Unit I

CHAPTER 1

Introduction to Data Science and Big Data

Syllabus Topics

Basics and need of Data Science and Big Data, Applications of Data Science, Data explosion, 5 V's of Big Data, Relationship between Data Science and Information Science, Business intelligence versus Data Science, Data Science Life Cycle, Data: Data Types, Data Collection. Need of Data wrangling, Methods: Data Cleaning, Data Integration, Data Reduction, Data Transformation, Data Discretization.

1.1 INTRODUCTION TO DATA SCIENCE AND BIG DATA

1.1.1 Introduction to Data Science

GQ. Explain data science in brief. (4 Marks)

Process

- **Data science** is a **process** in which it is examined that from where the information can be taken, what it signifies and how it can be converted into a useful resource in the creation of business and IT strategies.
- Mining huge quantity of structured and unstructured data to recognize patterns can help out an organization to reduce costs, raise efficiencies, identifies new market opportunities and enhances the organization's competitive benefit.
- The data science field manipulates the mathematics, statistics and computer science regulations, and includes methods like machine learning, cluster analysis, data mining and visualization.

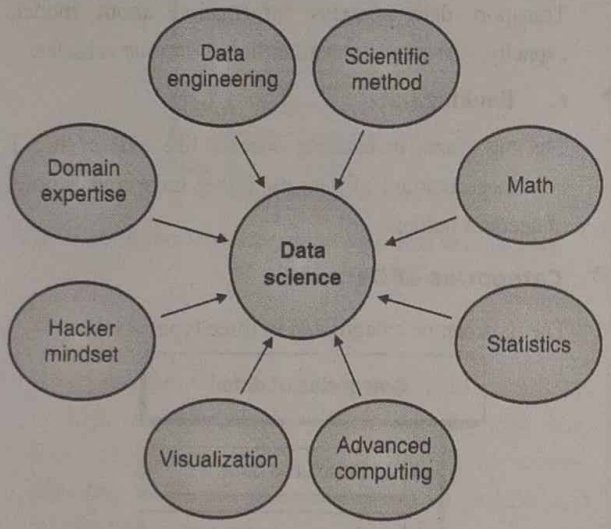


Fig. 1.1.1 : Data Science

Data scientists

- As we know that the amount of data generation can be increased by the typical modern businesses. Because of this the importance of data scientists can be increased.
- The task of data scientists is to convert the organization's raw data into the useful information.
- Data extraction is a method of retrieving particular data from unstructured or badly structured data sources for advance processing and analysis.
- Data scientists must acquire a mixture of analytic, machine learning, data mining and statistical skills, as well as familiarity with algorithms and coding.
- Another task for data scientists along with managing and understanding large amounts of data is to create data visualization models that facilitates demonstrating the business value of digital information.
- Data scientists must acquire an emotional intelligence in addition to education and experience in data analytics to make it effective.
- With the help of Smartphone's, Internet of Things (IoT) devices, social media, internet searches and behavior, the data scientists can illustrates the digital information very easily because they are studying them on regular basis.

❑ **Definition :** The **data mining** is the process of identifying the patterns to solve the problems by data scientists when such a large data sets are sorted with the help of data analysis.

Data science and machine learning

- Machine learning is often integrated in data science. Machine learning is an Artificial Intelligence (AI) tool that basically automates the data-processing piece of data science.
- Machine learning includes advanced algorithms that can be self learned and can process huge amounts of data within a fraction of time.
- After gathering and processing the structured data from the machine learning tools, data scientists takes data, transform it and summarize the data so it becomes useful for the company's decision-makers.
- **Example :** Examples of Machine learning applications in the data science field are image recognition and speech recognition, self-driving vehicles etc.

1.1.2 Introduction to Big Data

UQ. What is Big data ? Explain characteristics of big data. (SPPU – Q. 1(a), Aug. 18, May 19, 4 Marks)

GQ. Write a short note on Big Data? (4 Marks)

- Now a day the amount of data created by various advanced technologies like Social networking sites, E-commerce etc. is very large. It is really difficult to store such huge data by using the traditional data storage facilities.
- Until 2003, the size of data produced was 5 billion gigabytes. If this data is stored in the form of disks it may fill an entire football field. In 2011, the same amount of data was created in every two days and in 2013 it was created in every ten minutes. This is really tremendous rate.
- In this topic, we will discuss about big data on a fundamental level and define common concepts related to big data. We will also see in deep about some of the processes and technologies currently being used in this field.

☞ Big Data

❑ **Definition : Big data** means huge amount of data, it is a collection of large datasets that cannot be processed using traditional computing techniques. Big Data is complex and difficult to store, maintain or access in regular file system. Big Data becomes a complete subject, which involves different techniques, tools, and frameworks.

☞ Sources of big data

There are various sources of big data. Now a days in number of fields such huge data get created. Following are the some of fields.

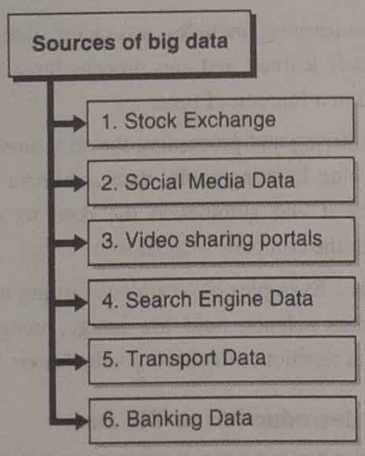


Fig. 1.1.2 : Sources of big data

1. Stock Exchange

The data in the share market regarding information about prices and status details of shares of thousands of companies is very huge.

2. Social Media Data

The data of social networking sites contains information about all the account holders, their posts, chat history, advertisements etc. On topmost sites like facebook and whatsapp, there are literally billions of users.

3. Video sharing portals

Video sharing portals like youtube, Vimeo etc. contains millions of videos each of which requires lots of memory to store.

4. Search Engine Data

The search engines like Google and Yahoo holds lot much of metadata regarding various sites.

5. Transport Data

Transport data contains information about model, capacity, distance and availability of various vehicles.

6. Banking Data

The big giants in banking domain like SBI or ICICI hold large amount of data regarding huge transactions of account holders.

☞ Categories of Data

The data can be categorized in three types :

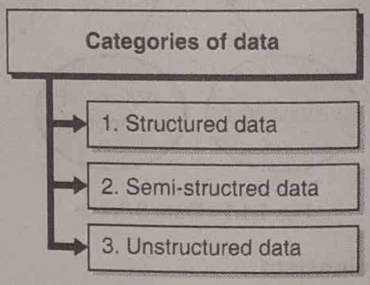


Fig. 1.1.3 : Categories of data

1. Structured Data

This type of data is stored in relations (tables) in Relational Database Management System.

2. Semi-structured Data

This type of data is neither raw data nor typed data in a conventional database system. A lot of data found on the web can be described as semi-structured data. This type of data does not have any standard formal model. This data is stored using various formats like XML and JSON.

3. Unstructured Data

This data do not have any pre-defined data model. The data of video, audio, Image, text, web logs, system logs etc. comes under this category.

☞ Important issues regarding data in traditional file

In general there are some important issues regarding data in traditional file storage system.

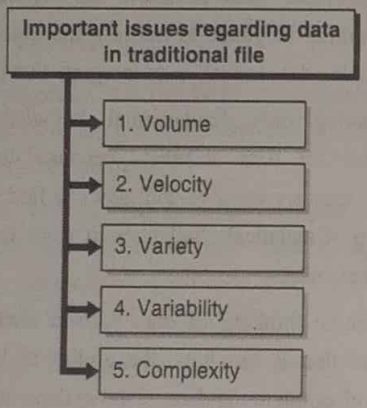


Fig. 1.1.4 : Important issues regarding data in regarding data in traditional file

1. Volume

Now a day the volume of data regarding different fields is high and potentially increasing day by day. Organizations collect data from a variety of sources, including business transactions, social media and information etc.

2. Velocity

The configuration of system with single processor, limited RAM and limited storage capacity cannot store and manage high volume of data.

3. Variety

The form of data from different sources is different.

4. Variability

The flow of data coming from sources like social media is inconsistent because of daily emerging new trends. It can show sudden increase in size of data which is difficult to manage.

5. Complexity

As the data is coming from various sources, it is difficult to link, match and transform such data across systems. It is necessary to connect and correlate relationships, hierarchies and multiple data linkages of the data.

All these issues are solved by the new advanced **Big Data Technology.**

1.2 DEFINING DATA SCIENCE AND BIG DATA

GQ. Define the term data science. **(2 Marks)**

GQ. Define Big Data. **(2 Marks)**

☞ Defining Data science

❑ **Definition : Data science** is a field of Big Data which searches for providing meaningful information from huge amounts of complex data.Data science is a system used for retrieving the **information** in different forms, either in structured or unstructured.

- Data Science unites different fields of work in statistics and computation in order to understand the data for the purpose of decision making.

☞ Defining Big Data

❑ **Definition : Big Data** is described as volumes of data available in changing level of complexity, produced at different velocities and changing level of ambiguity, that cannot be processed using conventional technologies, processing methods, algorithms, or any commercial off-the-shelf solutions.

- Data that can be defined as Big Data comes from variety of fields such as machine-generated data from sensor networks, nuclear plants, airplane engines, and consumer-driven data from social media.
- The producers of the Big Data that resides within organizations include legal, sales, marketing, procurement, finance, and human resources departments.

1.3 THE REQUISITE SKILL SET IN DATA SCIENCE

GQ. Explain requisite skill set in data science. **(4 Marks)**

Data science is a combination of skills consisting of three most important areas.

They are explained in brief in Fig. 1.3.1.

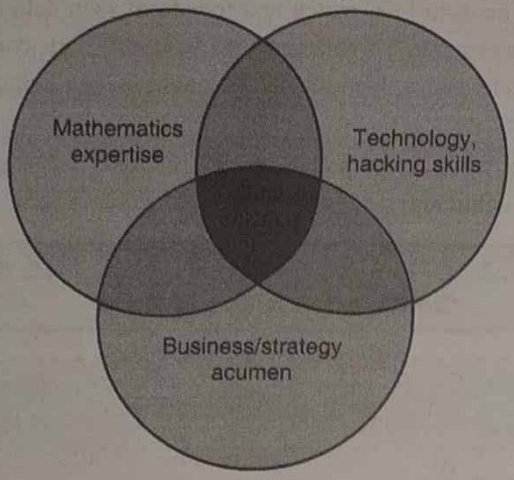


Fig. 1.3.1 : Requisite skill set in data science

1. Mathematics Expertise

- The most important thing required while constructing the data products and data mining insights is the capability to view the data via a quantitative way. There are texture, measurement, and relationship in data that can be illustrated mathematically.
- Solutions to numerous business problems occupies building analytic models grounded in the hard math, where being able to recognize the underlying mechanics of those models is key to success in building them.
- Also, a misunderstanding is that data science contains all about statistics. While statistics is important, it is not the only type of math utilized in the data science.
- There are two branches of statistics as given in Fig. 1.3.2.

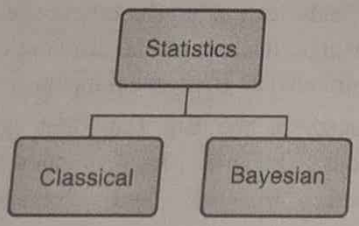


Fig. 1.3.2 : Branches of statistics

- Having knowledge of both classical and Bayesian statistics is helpful but when the majority of peoples refer to stats they are normally preferring to *classical statistics*.

2. Technology and Hacking

- Here the term "hacking" is related to the innovation not to the tempering any confidential data.
- We are going to refer hacking as a programmer's creativity and the cleverness to solve the problems that arise while building the things.
- The hacking is important because data scientists make use of *technology* in order to handle huge data sets and work with composite algorithms, and it needs tools far more difficult than Excel.
- Data scientists have to know the fundamentals of programming language to find out the quick solutions for complex data as well as to integrate that data.
- But having only fundamental knowledge is not sufficient for data scientists because data science hackers are very creative and they can find a way with the help of technical challenges to work their code in desired manner.
- Algorithmic thinking of data science hacker is very high, so that it can have the ability to break down confused problems and recompose them in ways that are solvable.

3. Strong Business Acumen

- For the data scientists, it is necessary to behave like a **tactical business consultant**. As the data scientists working are very close to the data so they can works like no one can do it.
- This will make a responsibility to transform observations to shared knowledge, and contribute to strategy on how to solve core business problems.
- This means a core ability of data science is using data to clearly inform a story. No data-puking – rather, present a unified description of problem and solution, with the help of data insights as supporting pillars, that lead to guidance.

1.4 EXAMPLES OF BIG DATA APPLICATIONS

GQ. List and explain the examples of big data. **(6 Marks)**

There are various big data applications as follows :

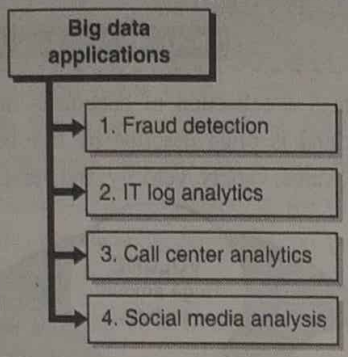


Fig. 1.4.1 : Big data applications

1. Fraud detection

- Fraud detection is a Big Data application example for businesses which has operations like any type of claims or transaction processing.
- Number of times the detection of fraud is concluded long after the fact. At this point the damage has been already done all that's left is to decrease the harm and revise policies to prevent it in future.
- The Big Data platforms can analyze claims and transactions of businesses. They identify large-scale patterns across many transactions or detect anomalous behaviour of some user. This helps to avoid the fraud.

2. IT log analytics

- An enormous quantity of logs and trace data is generated in IT solutions and IT departments. Many times such data go unexamined: organizations simply don't have the manpower or resource to go through all such information.
- Big data has the ability to quickly identify large-scale patterns to help in diagnosing and preventing problems. It helps the organization with a large IT department.

3. Call center analytics

- Now we turn to the customer-facing Big Data application examples, of which call center analytics are particularly powerful.
- Without a Big Data solution, much of the insight that a call center can provide will be ignored or exposed later.
- By making sense of time/quality resolution metrics, the Big Data solutions are able to identify recurring problems or customer and staff behaviour patterns. Big data can also capture and process call content itself.

4. Social media analysis

- With the help of Social media we can observe the real-time insights into how the market is responding to products and campaigns.
- With the help of these insights, it is possible for companies to adjust their pricing, promotion, and campaign placement to get optimal results.

1.5 DATA EXPLOSION

GQ. Explain Data explosion in detail. **(4 Marks)**

- **Definition :** The **data explosion** is nothing but the rapid growth of the data. One reason to this explosive growth of data is innovation.
- The Innovation has changed the way we engage in business, provide services, and the associated measurement of value and profitability.
- There are three basic trends available which becomes very essential to build up the data in the last few years. They are:

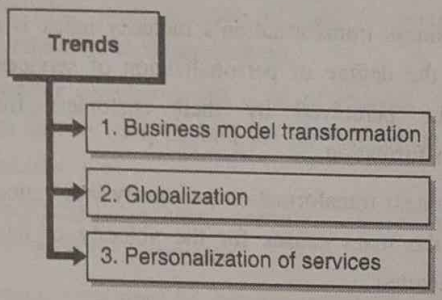


Fig. 1.5.1 : Basic Trends

1. Business model transformation

- The primary business models have been transformed through globalization and connectivity.
- Modern companies can be moved towards the service oriented technologies rather than product oriented.
- In the service oriented the value of the organization from customers point of view is measured by how

much the service is effective instead of how much product is useful.

- What this transformation commands means to every business is that you want to produce more data in terms of products and services to provide to each segment and channel of customers, and to hold lot of data from every customer, consisting of social media, surveys, forums, feedback from customers, call center, competitive market research, and many more.
- The amount of data created and stored by each organization today beats what the same organization produced prior to the business transformation.
- Now the data which is primary or having the higher priority are kept in center and the supporting data which is required but not available or accessible previously now can be available and also accessible with the help of multiple channels.
- Here the volume equation of data exploding to Big Data comes in the picture.

2. Globalization

- Globalization is a key trend that has radically changed the commerce of the world, starting from manufacturing to customer service.
- IT has also changed the variations and formats of data.

3. Personalization of services

- Business transformation's maturity index is measured by the degree of personalization of services and the value perceived by their customers from such transformation.
- Business transformation's maturity index model is one of the main causes for the velocity of data that is generated.

New sources of data

- As the technologies are growing the data can be generated from various sources such as social media, mobile devices, sensor media and many more which is not present before.
- The appearances of newer business models and the aggressive expansion of technology capabilities over the last decade or more has covered the way for incorporating all of the data across the enterprise into one holistic platform to create a significant and appropriate business decision support platform.

1.6 5 V'S OF BIG DATA

UQ. Explain 3V's of Big Data. **(SPPU - Q. 1(a), May 19, 5 Marks)**

- Big data is a collection of data from many different sources and is often describe by five characteristics: volume, value, variety, velocity, and veracity.

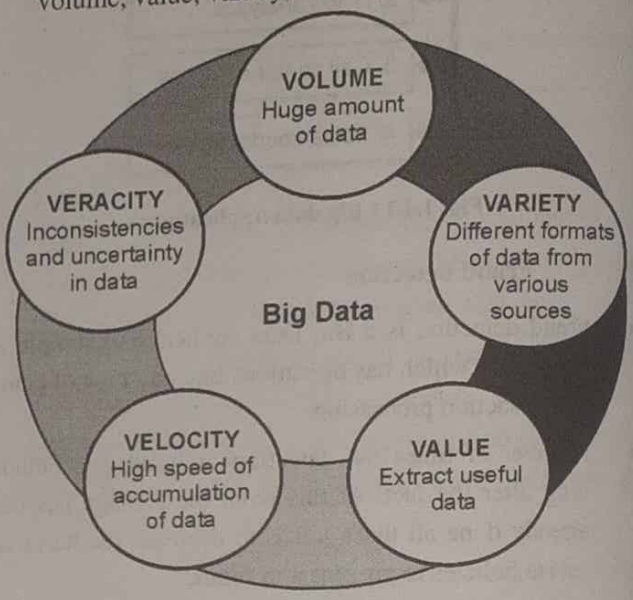


Fig. 1.5.2

- **Volume :** The size and amounts of big data that companies manage and analyze. The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, 'Volume' is one characteristic which needs to be considered while dealing with Big Data solutions.
- **Variety :** The diversity and range of different data types, including unstructured data, semi-structured data and raw data. Variety refers to heterogeneous sources and the nature of data, both structured and unstructured.
- Earlier days, most of the application was using database as a spreadsheet (Structured data). Now a day's data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc.(unstructured

data) are also being considered in the analysis applications.

- **Value :** The most important "V" from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and quantifiable business benefits. This refers to the value that big data can provide, and it relates directly to what organizations can do with that collected data. Being able to pull value from big data is a requirement, as the value of big data increases significantly depending on the insights that can be gained from them.
- **Velocity :** The speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time
- **Veracity :** The "truth" or accuracy of data and information assets, which often determines executive-level confidence.

1.7 RELATIONSHIP BETWEEN DATA SCIENCE AND INFORMATION SCIENCE

- Data Science is an interdisciplinary field which deals with all things data, providing ways to benefit from Big Data. The idea behind Data Science is to identify patterns, discover relationships, and to make sense of the raw data.
- It is a field that deals with the complex world of data while using a blend of tools and algorithms to extract useful information from the data.
- Every entity has some data associated with itself. This data helps us to identify an object or to categorize it.
- Data Science is a field that involves the use of statistical and scientific methods to draw useful insights from the data.
- "Information science is the science and practice dealing with the effective collection, storage, retrieval, and use of information.
- It is concerned with recordable information and knowledge, and the technologies and related services that facilitate their management and use.
- In Information science, Big Data is a term that refers to the vast amount of information about an entity either in the form of text, video, images or audio used for pattern recognition and decision making.
- The considerable differences between the two are listed below :

1.7.1 Data Science

- It involves the analysis of data and related pre-processing.
- Used for visualizing local and global trends in the data to generate useful insights.
- Applications include image processing, medical analysis, computer vision, etc
- Popular tools for Data Science are R, Python, etc.
- Data Science supporting software includes integrated development environments such as Spyder, R studio, Jupyter notebook, etc.
- Job opportunities are more as compared to Big Data. It is a multidisciplinary field that requires analysts, engineers as well as database professionals.
- The analysis involves a predictive and prescriptive approach to data.

1.7.2 Big Data

- It involves huge amounts of data for analyzing the patterns and prediction making.
- Used for visualizing and detailed analysis of complex data sets that are beyond the normal problem-solving methodologies.
- Applications include security, telecommunication services and E-commerce.
- Popular tools used for big data are the spark, Hadoop, etc.
- Technologies supporting big data include Apache Hadoop, Tableau, etc.
- Job opportunities are limited as it involves deep analysis of complex data. Hence, job opportunities get limited to the requirement of data analysts.
- The analysis used is mainly retrospective in nature.

1.7.3 Computer Science and Data Science

- Computer science can be referred to as the study of computers as well as computing concepts.
- It is basically the study of the processes which interact with data which is in the form of programs. It deals with the manipulation of the information by making the use of various algorithms.
- Thus computer science deals with the study of both hardware as well as software and other components like networking and the internet.
- The hardware part of computer science deals with the study of the basic design of computers and the working process of it. The software part of computer science deals with the study of programming concepts as well as languages. Computer science also deals with operating systems and compilers.
- Data science is basically a field in which information and knowledge are extracted from the data by using various scientific methods, algorithms, and processes. It can thus be defined as a combination of various mathematical tools, algorithms, statistics, and machine learning techniques which are thus used to find the hidden patterns and insights from the data which helps in the decision making process.
- Data science deals with both structured as well as unstructured data. It is related to both data mining and big data. Data science involves studying the historic trends and thus using its conclusions to redefine present trends and also predict future trends.

1.7.4 Difference Between Computer Science and Data Science

Sr. No	Computer Science	Data Science
1.	It is basically the study of the computational systems including both theory and applications.	It is a field that uses mathematics, statistics and various other tools to discover the hidden patterns in the data.
2.	It is mainly used for advancement and growth of technology.	It is mainly used for the management of data and data analysis.
3.	The advantage of using computer science is growth and development of technology.	The advantage of using data science is handling and maintenance of large volumes of data.
4.	It has been a branch of science from a really long time.	It has evolved recently as a developing branch of science.
5.	One after studying computer science becomes a computer science professional.	One after studying data science becomes a data scientist or a data analyst.
6.	Computer science is the super set of data science as it covers the entire technological field.	Data science is a subset of computer science which involves the study of data and its analysis.
7.	Its main benefit is technological advancement and improved speed and performance of technological devices.	Its main benefit is easy management of data and reduction of data redundancy.
8.	It is applied to nearly all the technical industries and companies.	It is basically applied to the industries and companies where data is of quite a lot importance.

1.8 BUSINESS INTELLIGENCE V/S DATA SCIENCE

UQ. Compare BI Vs. Data science.
(SPPU – Q. 2(a), Dec. 18, 6 Marks)

- The four business drivers which we have discussed in previous section need a variety of analytical techniques to address them properly.
- There are number of ways which helps to compare these groups of analytical techniques.
- One way for the evaluation of the type of analysis being carried out is to observe the time horizon and the type of analytical approaches being used.
- BI usually provides reports, dashboards, and queries on business questions for the current period or in the past.
- BI systems helps to simplify to answer questions regarding quarter-to-date revenue, progress toward quarterly targets, and know quantity of given product was sold in a prior quarter or year.
- These questions considered as closed-ended and explain current or past behavior, normally by the process of aggregating historical data and grouping it in some way.
- BI offers hindsight and little insight and usually answers questions regarding "when" and "where" events occurred.
- When compared with BI, it is found that Data Science like to use disaggregated data with a more forward-looking, exploratory technique, concentrating on analyzing the present and enabling informed decisions about the future.
- Instead of aggregating historical data to search for quantity of product sold in the previous quarter, it is possible for a team to employ Data Science techniques like time series analysis.
- Such techniques help to guess future product sales and revenue more precisely as compared to extending a simple trend line.
- Also, Data Science considered as more exploratory in nature and may like to refer scenario optimization for the purpose of dealing with more open-ended questions.
- This approach helps to get insight into current activity and foresight into upcoming events, while usually concentrating on questions regarding "how" and "why" events occur.
- Where BI problems needs highly structured data which has been organized in rows and columns for accurate reporting, Data Science projects mostly refer various kinds of data sources, including large or unconventional datasets.
- Based on an future goals of organization, it may prefer to board on a PI project if there is reporting, dashboards creation, or simple visualizations, or it may prefer to board on Data Science projects if it required to do a more sophisticated analysis with datasets which are in the form of disaggregated or distinct.

NOTES

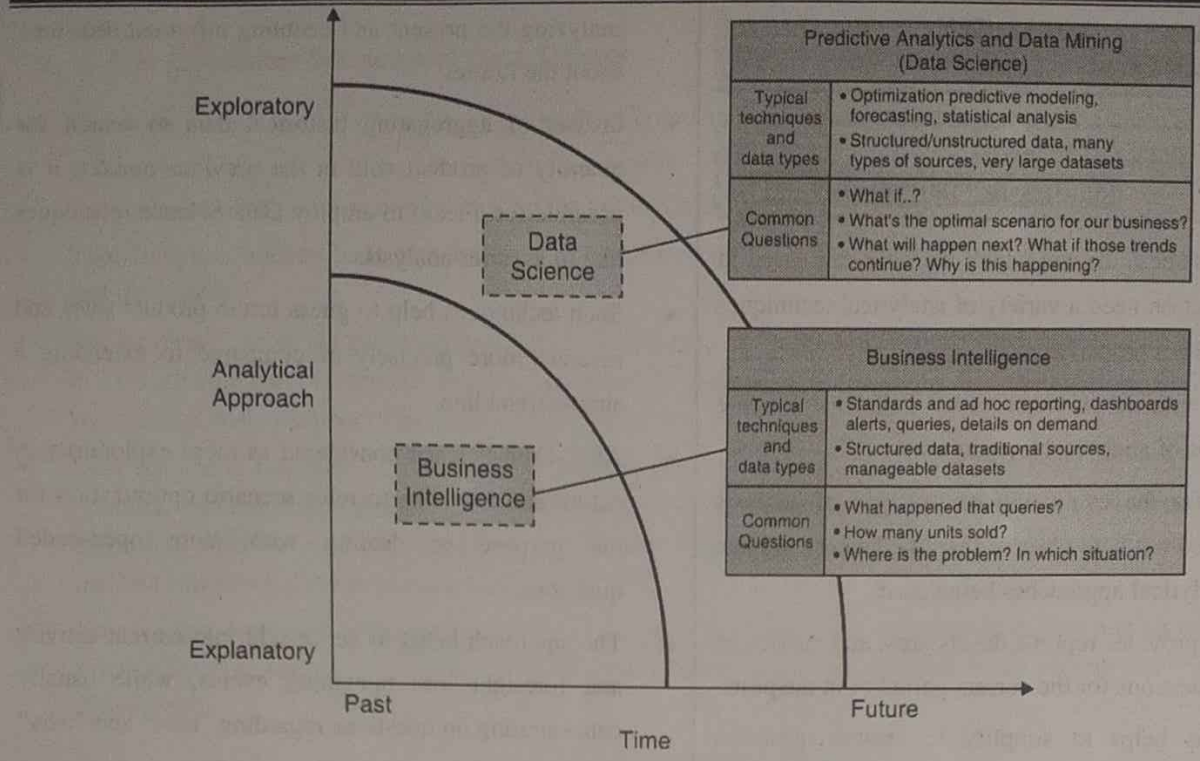


Fig. 1.8.1 : Comparing BI with Data Science

1.9 DATA SCIENCE LIFE CYCLE

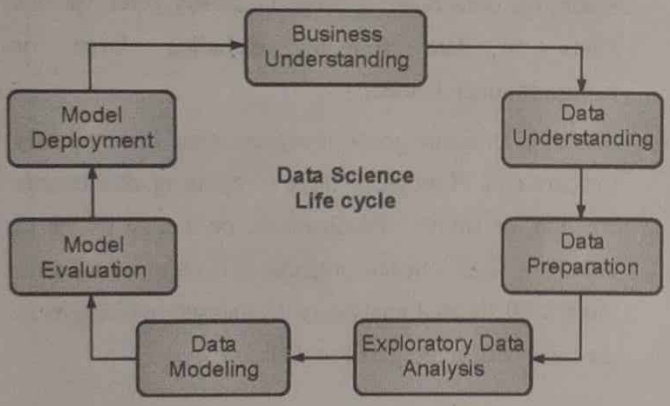


Fig. 1.9.1 : Data Science Lifecycle

1. **Business Understanding :** The complete cycle revolves around the enterprise goal. What will you resolve if you do no longer have a specific problem? It is extraordinarily essential to apprehend the commercial enterprise goal sincerely due to the fact that will be your ultimate aim of the analysis. After desirable perception only we can set the precise aim of evaluation that is in sync with the enterprise objective. You need to understand if the customer desires to minimize savings loss, or if they prefer to predict the rate of a commodity, etc.
2. **Data Understanding :** After enterprise understanding, the subsequent step is data understanding. This includes a series of all the reachable data. Here you need to intently work with the commercial enterprise group as they are certainly conscious of what information is present, what facts should be used for this commercial enterprise problem, and different information. This step includes describing the data, their structure, their relevance, their records type. Explore the information using graphical plots. Basically, extracting any data that you can get about the information through simply exploring the data.
3. **Preparation of Data :** Next comes the data preparation stage. This consists of steps like choosing the applicable data, integrating the data by means of merging the data sets, cleaning it, treating the lacking values through either eliminating them or imputing them, treating inaccurate data through eliminating them, additionally test for outliers the use of box plots and cope with them. Constructing new data, derive new elements from present ones. Format the data into the preferred structure, eliminate undesirable columns and

features. Data preparation is the most time-consuming but arguably the most essential step in the complete existence cycle. Your model will be as accurate as your data.

4. **Exploratory Data Analysis :** This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model. Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representations like scatter plots and warmth maps. Many data visualization strategies are considerably used to discover each and every characteristic individually and by means of combining them with different features.

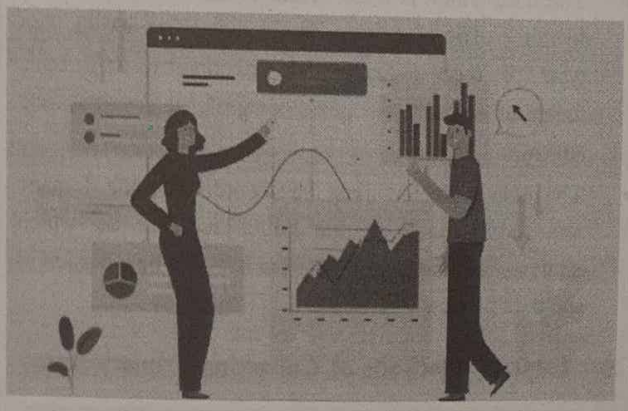
5. **Data Modeling :** Data modeling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output. This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem. After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them. We need to tune the hyperparameters of every model to obtain the preferred performance. We additionally need to make positive there is the right stability between overall performance and generalizability. We do no longer desire the model to study the data and operate poorly on new data.

6. **Model Evaluation :** Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics. We additionally need to make positive that the model conforms to reality. If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved. Any data science solution, a machine learning model, simply like a human, must evolve, must be capable to enhance itself with new data, adapt to a new evaluation metric. We can construct more than one model for a certain phenomenon, however, a lot of them may additionally be imperfect. The model assessment helps us select and construct an ideal model.

7. **Model Deployment :** The model after a rigorous assessment is at the end deployed in the preferred structure and channel. This is the last step in the data science life cycle. Each step in the data science life cycle defined above must be laboured upon carefully. If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste. For example, if data is no longer accumulated properly, you'll lose records and you will no longer be constructing an ideal model. If information is not cleaned properly, the model will no longer work. If the model is not evaluated properly, it will fail in the actual world. Right from Business perception to model deployment, every step has to be given appropriate attention, time, and effort.

1.10 DATA: DATA TYPES, DATA COLLECTION

- Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis. In the process of big data analysis, "Data collection" is the initial step before starting to analyze the patterns or useful information in data. The data which is to be analyzed must be collected from different valid sources.

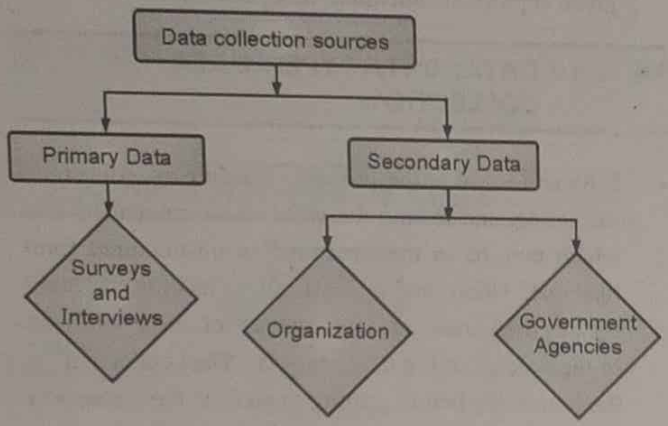


- The data which is collected is known as raw data which is not useful now but on cleaning the impure and utilizing that data for further analysis forms

information, the information obtained is known as "knowledge". Knowledge has many meanings like business knowledge or sales of enterprise products, disease treatment, etc. The main goal of data collection is to collect information-rich data.

- Data collection starts with asking some questions such as what type of data is to be collected and what is the source of collection. Most of the data collected are of two types known as "qualitative data" which is a group of non-numerical data such as words, sentences mostly focus on behavior and actions of the group and another one is "quantitative data" which is in numerical forms and can be calculated using different scientific tools and sampling data.

The actual data is then further divided mainly into two types known as:



A. Primary data

- The data which is Raw, original, and extracted directly from the official sources is known as primary data. This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys.
- The data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing.

1.10.1 Methods of Collecting Primary Data

1. Interview method

- The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee.
- Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

2. Survey method

- The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video.
- The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analyzing data. Examples are online surveys or surveys through social media polls.

3. Observation method

- The observation method is a method of data collection in which the researcher keenly observes the behavior and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats.
- In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behavior towards the products. The data obtained will be sent for processing.

4. Experimental method

The experimental method is the process of collecting data through performing experiments, research, and investigation. The most frequently used experiment methods are CRD, RBD, LSD, FD.

- **CRD** - Completely Randomized design is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.
- **RBD** - Randomized Block Design is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a

technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector.

- **LSD** - Latin Square Design is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is an arrangement of NxN squares with an equal amount of rows and columns which contain letters that occurs only once in a row. Hence the differences can be easily found with fewer errors in the experiment. Sudoku puzzle is an example of a Latin square design.
- **FD** - Factorial design is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived.

B. Secondary data

Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

1. Internal source

These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. The cost and time consumption is less in obtaining internal sources.

2. External source

The data which can't be found at internal organizations and can be gained through external third party resources is external source data. The cost and time consumption is more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labor bureau, syndicate services, and other non-governmental publications.

3. Other sources

- **Sensors data :** With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products.
- **Satellites data :** Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.
- **Web traffic :** Due to fast and cheap internet facilities many formats of data which is uploaded by users on different platforms can be predicted and collected with their permission for data analysis. The search engines also provide their data through keywords and queries searched mostly.

1.11 NEED OF DATA WRANGLING

Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data Wrangling is also known as Data Munging.

- Data wrangling helps data usability by transforming it to make it compatible with the end system as complex and intricate datasets can hinder data analysis and business processes. To make data usable for the end processes, data wrangling tools transform and organize data according to the target system's requirements.

Data Wrangling is a very important step. The below example will explain its importance as :

- Books selling Website want to show top-selling books of different domains, according to user preference. For example, a new user search for motivational books, then they want to show those motivational books which sell the most or having a high rating, etc.
- But on their website, there are plenty of raw data from different users. Here the concept of Data Munging or Data Wrangling is used. As we know Data is not Wrangled by System. This process is done by Data Scientists. So, the data Scientist will wrangle data in such a way that they will sort that motivational books that are sold more or have high ratings or user buy this book with these package of Books, etc. On the basis of that, the new user will make choice. This will explain the importance of Data wrangling.

1.12 DATA CLEANING

GQ. Write a short note on data cleaning. **(4 Marks)**

❑ **Definition :** Data cleaning is a process of finding the incorrect or corrupted data and removing it".

The data cleaning process is required because incorrect data can produce the wrong conclusions and bad analysis, particularly when considered the massive quantities of Big Data.

☞ **Example**

UK banking industry which is losing hundreds of millions of pounds due to Big Bad Data.

1.12.1 Data Issues

GQ. What are the data issues in data cleaning ? **(4 Marks)**

Following are the issues related to the data.

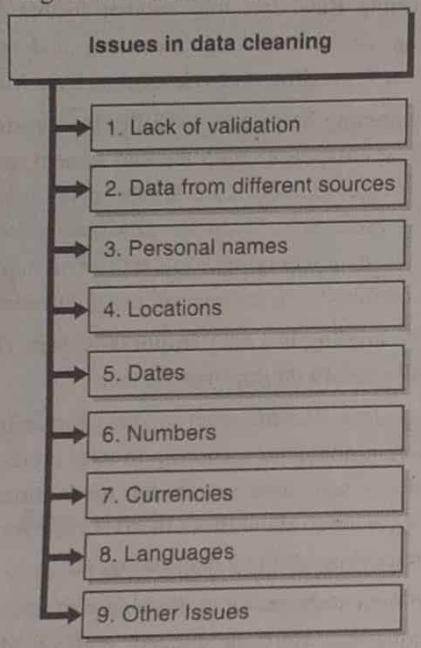


Fig. 1.12.1 : Issues in data cleaning

1. Lack of validation

- Whenever the data is inserted it is not always a case that data can be validated properly. It needs to be fixed after it is in the database.
- For example, phone numbers may be written with or without spaces, with letters instead of numbers, etc.

2. Data from different sources

- Data sources, maybe it is coming from inside or outside of the organization, may require cleansing when combining them together.
- Data which is present inside the organization could appear from different applications. These applications may provide data about the same entities, but not the same identities.
- Data which is coming from outside of organization could already appear with ambiguous or corrupt information.

3. Personal names

The same name could be written in several ways - full name, full name with comma, last name only, and more. For example: "John A. Smith", or "Smith, John A.", or "Smith", and so forth.

4. Locations

- Sometime the different names can be used for the same locations. For example, Bombay is the same city as Mumbai, different people may use either name while meaning the same thing.
- Even the US could be noted as USA or written out as United States of America or just United States.

5. Dates

Dates present a particularly big data migraine as there are plenty of formats. The same date could be written as 11 December 2017, 11/12/2017, 12.11.2017, 11-Dec-2017 and so on.

6. Numbers

- Dots and commas have a different role when it comes to writing numbers in German - the number 1,024.56 would be written by Germans as 1.024,56.
- Even Americans could write the same number without the comma as 1024.56 while Australians would write it with a space character as 1 024.56.

7. Currencies

Data related to finance may appear in different currencies. They have a different financial value and different symbols or codes that may need to be merge, e.g. $32, 32 USD, 1057.92 THB.

8. Languages

Data could also appear in different languages such as English, Chinese, or Arabic. It contains not only just free text, but also personal names, addresses, and so on.

9. Other Issues

Spelling mistakes, ambiguity in the upper or lower case also requires to be cleaned depends on the context of business.

1.12.2 Cleaning Methods

GQ. Explain various cleaning methods. **(4 Marks)**

Following are some methods are used for the cleaning the data :

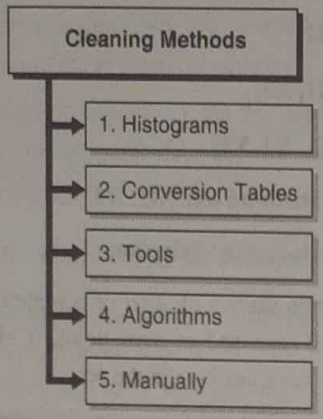


Fig. 1.12.2 : Cleaning Methods

1. Histograms

- The histograms method is used to find out which values are used less frequently and hence becomes invalid.
- These values can then be updated, but this is a problem with Hadoop because it does not provide update functionality.

2. Conversion Tables

- In a situation where the data issues are already known the conversion of tables can be preferable.
- The data should be sorted by the related key, lookups could be used to make the conversions, and at the last whatever results are produced can be stored for further use.

3. Tools

- A variety of vendors such as IBM, SAS, Oracle, and Lavastorm Analytics offers solutions for the data cleansing.
- There also free tools are available such as Open Refine, plye,reshape2 and so on ,but it is not clear that whether they can manage Big Data.

4. Algorithms

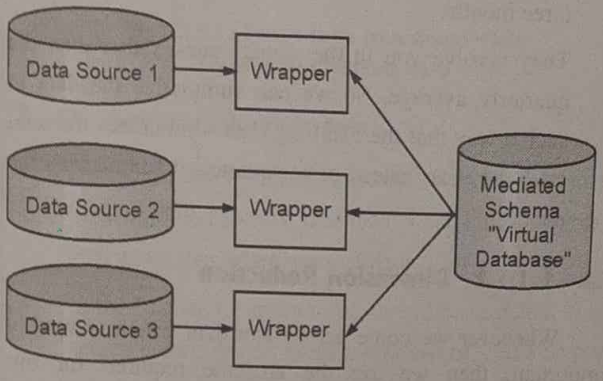
- To fix some of the data, spell checking or phonetic algorithms can be useful.
- This will create one problem that it can corrupt data by using the wrong suggestions therefore the manual work is necessary.

5. Manually

- In general most of the data is typically cleaned by hand.
- Even if there are multiple tools, histograms, and algorithms are available, the human interference is still required to understand and fix the data.

1.13 DATA INTEGRATION

- Data Integration is a data preprocessing technique that involves combining data from multiple heterogeneous data sources into a coherent data store and provide a unified view of the data. These sources may include multiple data cubes, databases, or flat files.
- The data integration approaches are formally defined as triple <G, S, M> where, G stand for the global schema, S stands for the heterogeneous source of schema, M stands for mapping between the queries of source and global schema.



- There are mainly 2 major approaches for data integration – one is the "tight coupling approach" and another is the "loose coupling approach".

1.13.1 Tight Coupling

- Here, a data warehouse is treated as an information retrieval component.
- In this coupling, data is combined from different sources into a single physical location through the process of ETL – Extraction, Transformation, and Loading.

1.13.2 Loose Coupling

- Here, an interface is provided that takes the query from the user, transforms it in a way the source database can understand, and then sends the query directly to the source databases to obtain the result.
- The data only remains in the actual source databases.

1.14 DATA REDUCTION

The method of data reduction may achieve a condensed description of the original data which is much smaller in quantity but keeps the quality of the original data.

Methods of data reductions are explained as following :

1.14.1 Data Cube Aggregation

- This technique is used to aggregate data in a simpler form. For example, imagine that information you gathered for your analysis for the years 2012 to 2014, that data includes the revenue of your company every three months.
- They involve you in the annual sales, rather than the quarterly average, So we can summarize the data in such a way that the resulting data summarizes the total sales per year instead of per quarter. It summarizes the data.

1.14.2 Dimension Reduction

Whenever we come across any data which is weakly important, then we use the attribute required for our analysis. It reduces data size as it eliminates outdated or redundant features.

a. Step-wise Forward Selection

- The selection begins with an empty set of attributes later on we decide best of the original attributes on the set based on their relevance to other attributes. We know it as a p-value in statistics.
- Suppose there are the following attributes in the data set in which few attributes are redundant.

Initial attribute Set: {X1, X2, X3, X4, X5, X6}

Initial reduced attribute set: { }

- **Step 1: {X1}**
- **Step 2: {X1, X2}**
- **Step 3: {X1, X2, X5}**

Final reduced attribute set: {X1, X2, X5}

b. Step-wise Backward Selection

- This selection starts with a set of complete attributes in the original data and at each point, it eliminates the worst remaining attribute in the set.
- Suppose there are the following attributes in the data set in which few attributes are redundant.

Initial attribute Set: {X1, X2, X3, X4, X5, X6}

Initial reduced attribute set: {X1, X2, X3, X4, X5, X6 }

- **Step 1: {X1, X2, X3, X4, X5}**
- **Step 2: {X1, X2, X3, X5}**
- **Step 3: {X1, X2, X5}**

Final reduced attribute set : {X1, X2, X5}

c. Combination of forwarding and Backward Selection

It allows us to remove the worst and select best attributes, saving time and making the process faster.

1.14.3 Data Compression

The data compression technique reduces the size of the files using different encoding mechanisms (Huffman Encoding & run-length Encoding). We can divide it into two types based on their compression techniques.

a. Lossless Compression

- Encoding techniques (Run Length Encoding) allows a simple and minimal data size reduction.
- Lossless data compression uses algorithms to restore the precise original data from the compressed data.

b. Lossy Compression

- Methods such as Discrete Wavelet transform technique, PCA (principal component analysis) are examples of this compression. For e.g., JPEG image format is a lossy compression, but we can find the meaning equivalent to the original the image.
- In lossy-data compression, the decompressed data may differ to the original data but are useful enough to retrieve information from them.

1.14.4 Numerosity Reduction

In this reduction technique the actual data is replaced with mathematical models or smaller representation of the data instead of actual data, it is important to only store the model parameter. Or non-parametric method such as clustering, histogram, sampling.

1.14.5 Discretization & Concept Hierarchy Operation

Techniques of data discretization are used to divide the attributes of the continuous nature into data with intervals. We replace many constant values of the attributes by labels of small intervals. This means that mining results are shown in a concise, and easily understandable way.

a. Top-down discretization

If you first consider one or a couple of points (so-called breakpoints or split points) to divide the whole set of attributes and repeat of this method up to the end, then the process is known as top-down discretization also known as splitting.

b. Bottom-up discretization

If you first consider all the constant values as split-points, some are discarded through a combination of the neighbourhood values in the interval, that process is called bottom-up discretization.

1.15 DATA TRANSFORMATION

GQ. Write a short note on data transformation. **(4 Marks)**

- There is no meaning to Big Data without structure. When effectively and quickly the unstructured data is converted to structured form, the Big data transforms to **smart data**.
- The unstructured data, when mixed with behavioral information as well as survey data, it provides rich insights. Smart data for business is used to understand the world perfectly.
- There are following important points to be considered while transforming the big data :

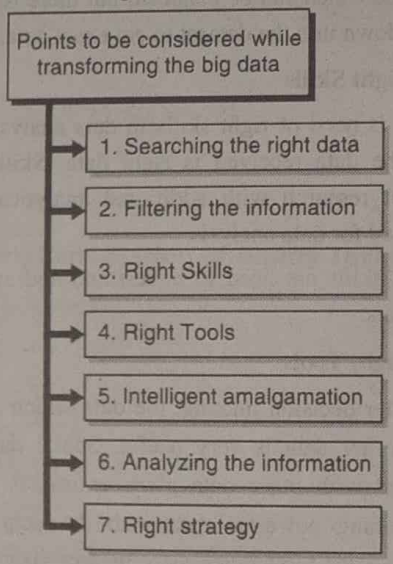


Fig. 1.15.1 : Points to be considered while transforming the big data

1. Searching the right data

- Nowadays there is huge amount of social data in the form of twitter tweets, facebook posts, blogs, forums as well as various other types of data.
- In the first step, certain topic should be searched in any search engine which gives a long list of URLs or posts which is not in defined order.
- Eventually, there is necessity to find the right data with the help of appropriate filters.

2. Filtering the information

- In the second stage, filters of different types can be added. The basic filter is Time frame. Consider we want information of result regarding previous years. Then the information can be filtered depending upon the media channels such as Facebook, Twitter, YouTube videos or other blog posts.
- Multinational companies search for posts origination as well as the language in which it is written. Some more filters can be added to transform disordered Big Data to manageable smart data.
- These multiple filters are used to retrieve fair idea regarding exactly when and where a particular topic is coming from. In this phase there are large number of insights which can be collected but there is need to dig deep down into the dataset to get exact data.

3. Right Skills

- There is need of right skills in data analysis to ensure that the data received is right data. Skills regarding market research with additional analytical skills are required for data analysis.
- These skills are used to understand and analyze huge data sets.

4. Right Tools

- In better decision making, the data which is combined with other data is very useful. Social data which is merged with other data gives enormous advantages. Combining powerful data with the help of various cutting edge tools gives more prefect view of behavior regarding customers.
- Better tools and advanced technology is available for analysts to handle large data sets. Transforming big data into smart data benefits in searching key insights regarding businesses and make decisions depending upon more relevant data.

5. Intelligent amalgamation

- In the process of converting big data to smart data, it is very important to understand customer behavior. The satisfaction of consumers can be achieved only when demand of consumers are understood otherwise it is difficult to make the big data as smart data.
- Intelligent amalgamation helps to combine customer behavior with data analytics so as to evade needless risks. With the help of intelligent amalgamation, it is possible to determine and target deep customer segments. It helps in getting new customers as well as retaining valued existing customers. It gives the experiences which is helpful in assuring big data for success.

6. Analyzing the information

- Filtering the information helps analysts to generate set of manageable data. With the help of marketing analytics, it is possible to measure and analyze the performance as we well as effectiveness of market and improve ROI (return on investment).
- With the help of advanced analytics, it is easy for professionals to observe the keywords associated with certain brands. They can also become witness for certain posts on twitter and facebook which have most effective impact on various brands. The effect may be positive or negative.

7. Right strategy

- The process regarding search, filter and analyze huge amount of big data gives rich source of actionable insights for various organizations. At various levels regarding filtering and analyzing, we can get key insights on multiple subjects, from various posts but as we go further, the data becomes more and more smart.
- These rich insights are used and incorporated by the professionals into business strategy and transform. As the data is refined, it can be used in different forms.
- Organization can use the smart data to build a healthy relationship with the customer. It is centralized, clean as well as context-based regarding customer insights as well as behavior. This data can also be used to promote trust between the customer and the organization which ultimately leads to improvement in customer experience.

1.16 DATA DISCRETIZATION OR BINNING

- **Discretization** is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes.
- Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values.
- For example, (1-10, 11-20) (age:- young, middle age, senior).
- Real-world data tend to be noisy. Noisy data is data with a large amount of additional meaningless information in it called noise. Data cleaning (or data cleansing) routines attempt to smooth out noise while identifying outliers in the data.
- There are three data smoothing techniques as follows –

1. **Binning :** Binning methods smooth a sorted data value by consulting its "neighborhood", that is, the values around it.
2. **Regression :** It conforms data values to a function. Linear regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other.
3. **Outlier analysis :** Outliers may be detected by clustering, for example, where similar values are organized into groups, or "clusters". Intuitively, values that fall outside of the set of clusters may be considered as outliers.

Data binning, bucketing is a data pre-processing method used to minimize the effects of small observation errors. The original data values are divided into small intervals known as bins and then they are replaced by a general value calculated for that bin. This has a smoothing effect on the input data and may also reduce the chances of overfitting in the case of small datasets There are 2 methods of dividing data into bins:

1. **Equal Frequency Binning :** bins have an equal frequency.
2. **Equal Width Binning :** bins have equal width with a range of each bin are defined as [min + w], [min + 2w] .... [min + nw] where w = (max – min) / (no of bins).

Equal frequency

Input : [5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]

Output

[5, 10, 11, 13]

[15, 35, 50, 55]

[72, 92, 204, 215]

Equal Width

Input : [5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]

Output

[5, 10, 11, 13, 15, 35, 50, 55, 72]

[92]

[204, 215]

Code : Implementation of Bining Technique

```
# equal frequency
defequifreq(arr1, m):
    a =len(arr1)
    n =int(a /m)
    fori inrange(0, m):
        arr =[]
    forj inrange(i *n, (i +1) *n):
        ifj >=a:
            break
        arr =arr +[arr1[j]]
    print(arr)
# equal width
defequiwidth(arr1, m):
    a =len(arr1)
    w =int((max(arr1) -min(arr1)) /m)
    min1 =min(arr1)
    arr =[]
    fori inrange(0, m +1):
     arr =arr +[min1 +w *i]
```

```
    arri=[]
     fori inrange(0, m):
        temp =[]
     forj inarr1:
        ifj >=arr[i] andj <=arr[i+1]:
            temp +=[j]
     arri +=[temp]
     print(arri)
# data to be binned
data =[5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]
# no of bins
m =3
print("equal frequency binning")
equifreq(data, m)

print("\n\nequal width binning")
equiwidth(data, 3)
```

Output

```
equal frequency binning
[5, 10, 11, 13]
[15, 35, 50, 55]
[72, 92, 204, 215]
equal width binning
[[5, 10, 11, 13, 15, 35, 50, 55, 72], [92], [204, 215]]
```

Chapter Ends...

❑❑❑

UNIT II

CHAPTER 2

Statistical Inference

Syllabus Topics

Need of statistics in Data Science and Big Data Analytics, Measures of Central Tendency: Mean, Median, Mode, Mid-range. Measures of Dispersion: Range, Variance, Mean Deviation, Standard Deviation. Bayes theorem, Basics and need of hypothesis and hypothesis testing, Pearson Correlation, Sample Hypothesis testing, Chi-Square Tests, t-test.

Unit II In Sem.

Someone has jokingly remarked

'If the head is in boiler and legs in fridge, then the temperature of the stomach is Statistics.'

☞ **Objectives**

By the end this chapter you should be able to :

- Distinguish between mean, median and mode;
- Calculate percentiles, deciles, quartiles etc.
- Distinguish between grouped data and ungrouped data.

2.1 INTRODUCTION TO STATISTICS

❒ **Definition (1) :** A Variate is any quantity or attribute whose value varies from one unit of investigation to another.

❒ **Definition (2) :**An observation is the value taken by a variate for a particular unit of investigation.

Variates differ in nature, and the methods of analysis of a variate depend on its nature. And we can distinguish between quantitative variates (like the birth-weight of the baby, etc.) and qualitative variates (such as the sex of the baby etc.)

❒ **Definition (3) :** A quantitative variate is a variate where values are numerical.

❒ **Definition (4) :** A qualitative variate or attribute is a variate whose values are not numerical.

Qualitative variates can also be divided into two types :

(i) They may be continuous, if they can take any value we specify in some range or

(ii) Discrete if their values change by steps or jumps.

❒ **Definition (5) :** A continuous variate is a variate which may take all values within a given range.

❒ **Definition (6) :**A discrete variate is a variate whose values change by steps.

The choice of which variates to record is important in any investigation. Once choice is made, the information can be summarized by the frequency distribution of the possible 'values'.

❒ **Definition (7) :** The frequency distribution of a (discrete) variate is the set of possible values of the variate, together with the associated frequencies.

❒ **Definition (8) :** The frequency distribution of a (continuous) variate is the set of class-intervals for the variate, together with the associated class-frequencies.

If we classify the whole population according to birth-weights, then instead of looking at the frequency of each variate, we first group the values into intervals, which is the sub-division of the total range of possible values of the variate.

In this example, the variate may be classified as 1-500, 500-1000, 4501-5000, 5001-5500 grams.

❒ **Definition (9) :** A class-interval is a sub-division of the total range of values which a (continuous) variate may take.

❒ **Definition (10) :** The class-frequency is the number of observations of the variate which fall in a given interval.

❒ **Definition (11) :** Cumulative frequency distribution is the sum of all observations which are less than the upper boundary of a given class interval : or this number is the sum of the frequencies upto and including that class to which the upper class boundary corresponds.

- For example, consider the heights of 50 students. We prepare the Table 2.1.1.

Table 2.1.1 : Cumulative frequency (more than) Table

Class (cm) interval	Frequency	Cumulative frequency more than
145-146	2	2
147-148	5	7
149-150	8	15
151-152	15	30
153-154	9	39
155-156	6	45
157-158	4	49
159-160	1	50
Total	**50**	

Table 2.1.2 : Cumulative frequency (less than) Table

Class (cm) interval	Frequency	Cumulative frequency more than
145-146	2	50
147-148	5	48
149-150	8	43
151-152	15	35
153-154	9	20
155-156	6	11
157-158	4	5
159-160	1	1
Total	**50**	

❒ **Definition (12)**

Points to note while constructing the Tables,

(1) Make the table self-explanatory provide a title, a brief description of a source of the data, State in what units the figures are expressed, label rows and columns where appropriate.

(2) Keep the table as simple as possible.

(3) Distinguish between zero values and missing observations.

(4) Make alternations clearly.

(5) Give the calculations of logical pattern on the sheet.

2.2 MEASURES OF CENTRAL TENDENCY

- One of the most important aspects of describing a distribution is the central value around which the observations are distributed.
- Any arithmetical measure which gives the centre or central value of a set of observations is known as a measure of central tendency or measure of location.

2.3 REVIEW OF BASIC RESULTS IN THE THEORY OF STATISTICS

2.3.1 Range and Mid-ranage

One way to measure the variability in a sample is simply to look at the highest and the lowest of the observations in a set, and calculate the difference between them.

❒ **Definition 1 :** The range of a set of observations is the difference in values between the largest and smallest observations in the set.

❒ **Definition 2 :** The mid-range is the average of the largest and smallest values in the data set.

e.g. for $X = \{1, 3, 5, 7, 9, 11, 13\}$

$$\text{mid range} = \frac{1 + 13}{2} = 7$$

Unit II In Sem.

2.3.2 Variance and Standard Deviation

Pursuing the idea of measuring how closely a set of observations cluster round their mean, we square each deviation $(x_i - \bar{x})$ instead of taking its absolute value.

The next measure of variability is the

Variance : It is the mean of the squared deviations.

❒ **Definition (I) : Variance**

(I) The variance of a set of observations $x_1, x_2, \ldots, x_n$ is the average of the squared deviations from their mean and equals $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

On simplification it is equal to

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{\sum x_i}{n}\right)^2$$

(II) For grouped data

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n} f_i (x_i - \bar{x})^2$$

❒ **Definition : Standard deviation**

(I) The standard deviation is the positive square root of the variance, and is equal to

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$ and denoted by σ

(II) For grouped data

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_i x_i^2 - \left(\frac{\sum f_i x_i}{n}\right)^2}$$

2.3.3 Arithmetic Mean

If f_1, f_2, f_n are frequencies of the variates $x_1, x_2, \ldots, x_n$ then,

$$M = \text{Arithmetic Mean (A.M.)} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Short-cut method of finding mean :

Let, $x' = \frac{x - x_0}{h}$; where x_0 is assumed mean and h is length of the class interval

Then $M = x_0 + hA$

Where, $A = \frac{\sum f \cdot x'}{\sum f}$

2.3.4 r^{th} Moments about Mean

Let, (x_i, f_i) be the given frequency distribution, then the r^{th} moment about mean M is given by,

$$\mu_r = \frac{\sum f_i (x_i - M)^2}{\sum f_i}$$

If $r = 1$, then $\mu_1 = 0$

$$\because \quad \mu_1 = \frac{\sum f_i (x_i - M)}{\sum f_i}$$

$$= \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i M}{\sum f_i} = \frac{\sum f_i x_i}{\sum f_i} - M\frac{\sum f_i}{\sum f_i}$$

$$= M - M \cdot 1 = 0$$

and for, $r = 2$

$$\mu_2 = \text{Variance} = \frac{\sum f_i (x_i - M)^2}{\sum f_i} = \sigma^2$$

= Square of standard deviation

Definition : r^{th} moment about origin are given by,

$$\mu_r' = \frac{\sum f_i x_i^r}{\sum f_i}$$

if, $r = 1$, $\mu_1' = \frac{\sum f_i x_i}{\sum f_i}$ = Mean M,

Again, $\mu_2' = \frac{\sum f_i x_i^2}{\sum f_i}$; $\mu_3' = \frac{\sum f_i x_i^3}{\sum f_i}$; $\mu_4' = \frac{\sum f_i x_i^4}{\sum f_i}$

2.3.5 Relation between Moments about Mean (μ_r) and Moments about Origin (μ_r')

(i) $\mu_2 = \mu_2' - (\mu_1')^2$

(ii) $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$

(iii) $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$

Also note that,

If $x' = \frac{x - x_0}{h}$ Then, $\mu_r' = h^r \frac{\sum f_i x_i'^r}{\sum f_i}$

2.3.6 Kari Pearson's Coefficients of Kurtosis

(i) β_1 = Measure of skewness $= \frac{\mu_3^2}{\mu_2^3}$

(ii) β_2 = Measure of flatness of single humped distribution

$$= \frac{\mu_4}{\mu_2^2}$$

Note : For normal distribution, $\beta_2 = 3$. If $\beta_2 > 3$, then distribution is peaked sharply than the normal cure and is known as lepto-kurtic.

If $\beta_2 < 3$, the distribution is flat compared to normal curve and is known as plato-kurtic.

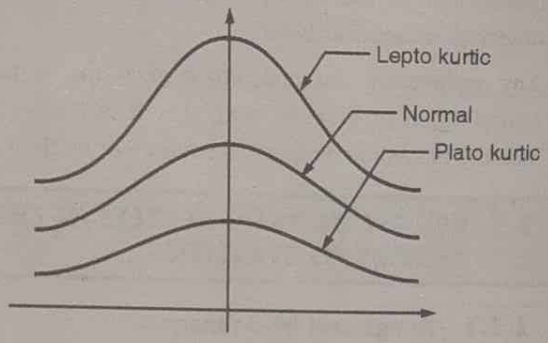


Fig. 2.3.1

2.3.7 The expected value of x : (mean value of x)

If X is a random variable then the expected value of X is denoted by E (X) and means the value, on average, that X takes.

Definition : If $X = x_i$, i = 1 to n, is a discrete random variable with frequencies f_i, i = 1 to n, then

$$E(X) = \sum_{i=1}^{n} x_i f(x_i)$$

Note : The expected value of X is also called as the Mean value of X and is also denoted by M. i.e. M = E (X)

Properties

(i) The r^{th} moment about origin is also written as,

$$\mu_r' = E(x^r) = \sum_{i=1}^{n} (x_i^r) f_i$$

Clearly, $E(x) = \mu_1' = \sum_{i=1}^{n} x_i f_i$

$$E(x^2) = \mu_2' = \sum_{i=1}^{n} x_i^2 f_i$$

$$E(x^3) = \mu_3' = \sum_{i=1}^{n} x_i^3 f_i \text{ and}$$

$$E(x^4) = \mu_4' = \sum_{i=1}^{n} x_i^4 f_i \text{ and so on.}$$

(ii) Moment about the mean $\bar{x}$ are defined as,

$$\mu_r = E\left[x_i - \bar{x}\right]^r = \sum_{i=1}^{n} (x_i - \bar{x}) f_i$$

and is called as r^{th} moment about mean $\bar{x}$.

Clearly, $\mu_1 = 0$

$$\mu_2 = E(x_i - \bar{x})^2 = \mu_2' - \mu_1'^2$$

$$\mu_3 = E(x_i - \bar{x})^3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

and $\mu_4 = E(x_i - \bar{x})^4$

$$= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

2.3.8 Covariance

- In probability theory and statistics, **covariance** is a measure of the joint variability of two **random variables**.
- If the greater values of one variable correspond with the greater values of the other variable, and the same holds for the lesser values (that is, the variables tend to show similar behaviour), the covariance is **positive**.
- In the opposite case, when the greater values of one variable correspond to the lesser values of the other, (that is, the variables tend to show opposite behaviour), the covariance is **negative**.
- The sign of the covariance shows the tendency in the linear relationship between the variables.

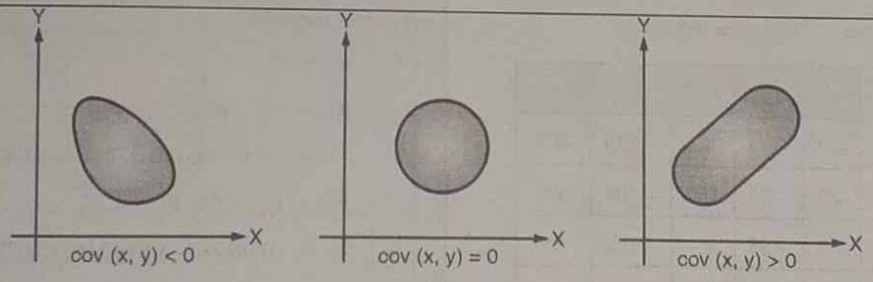


Fig. 2.3.2

Formulae of covariance

If X and Y are two random variables, then covariance between them is defined as :

$$\text{cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$$

$$= E\{XY - X E(Y) - Y E(X) + E(X) E(Y)\}$$

$$= E(XY) - E(X) E(Y) - E(Y) E(X) + E(X) \cdot E(Y)$$

$$\text{cov}(X, Y) = E(XY) - E(X) \cdot E(Y) \quad \ldots(i)$$

Unit II In Sem.

If X and Y are independent, then

$$E(XY) = E(X) \cdot E(Y)$$

and hence in this case,

$$\begin{aligned} \text{cov}(X, Y) &= E(X)E(Y) - E(X)E(Y) \\ &= 0 \end{aligned}$$

Remarks

(i) $$\begin{aligned} \text{cov}(aX, bY) &= E\{[aX - E(ax)][bY - E(bY)]\} \\ &= E\{[aX - aE(X)][bY - bE(Y)]\} \\ &= E\{a[X - E(X)]b[Y - E(Y)]\} \\ &= ab\,E\{[X - E(X)][Y - E(Y)]\} \\ &= ab\,\text{cov}(X, Y) \end{aligned}$$

(ii) $\text{cov}(X + a, Y + b) = \text{cov}(X, Y)$

(iii) $\text{cov}(aX + b, (Y + d)) = ac\,\text{cov}(X, Y)$

(iv) $\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z)$

(v) If X and Y are independent, then $\text{cov}(X, Y) = 0$ but the converse is not true.

2.3.9 Examples

Ex. 2.3.1 : For the following distribution, find :

(i) Arithmetic mean (ii) Standard derivation

(iii) First 4 moments about the mean (iv) β_1 and β_2.

x	2	2.5	3	3.5	4	4.5	5
f	5	38	65	92	70	40	10

Soln. :

Let, $x' = \dfrac{x - 3.5}{0.5}$

$x_0 = 3.5$; $h = 0.5$

x	f	x'	fx'	fx'^2	fx'^3	fx'^4
2	5	– 3	– 15	45	– 135	405
2.5	38	– 2	– 76	152	– 304	608
3	65	– 1	– 65	65	– 65	65
3.5	92	0	0	0	0	0
4	70	1	70	70	70	70
4.5	40	2	80	160	320	640
5	10	3	30	90	270	810
Total	**320**		**24**	**582**	**156**	**2598**

(i) Arithmetic mean : Using result of **A** in section 2.3 we have,

$$A = \frac{\sum fx'}{\sum r} = \frac{24}{320} = 0.075$$

and arithmetic mean $= x_0 + hA$

$$= 3.5 + (0.5)(0.075) = -3.538$$

(ii) Standard deviation : Using the result of **B** in section 2.3, we have

$$\begin{aligned} \sigma^2 &= h^2 \left\{ \frac{\sum fx'^2}{\sum f} - \left(\frac{\sum fx'}{\sum r} \right)^2 \right\} \\ &= (0.5)^2 \left\{ \frac{582}{320} - \left(\frac{24}{320} \right)^2 \right\} = 0.453 \\ \sigma &= 0.673 \end{aligned}$$

(iii) Moments about the mean M

When assumed mean is $A = x_0 = 3.5$ and using **C** in section 2.3, we have,

$$\mu_1' = h \frac{\sum fx'}{\sum f} = 0.5 \left(\frac{24}{320} \right) = 0.0375$$

$$\mu_2' = h^2 \frac{\sum fx'^2}{\sum f} = (0.5)^2 \left(\frac{582}{320} \right) = 0.4546$$

$$\mu_3' = h^3 \frac{\sum fx'^3}{\sum f} = (0.5)^3 \left(\frac{156}{320} \right) = 0.0609$$

$$\mu_4' = h^4 \frac{\sum fx'^4}{\sum f} = (0.5)^4 \left(\frac{2598}{320} \right) = 0.5074$$

Using result, D in section 2.3, we have for moments about the mean M.

$$\mu_1 = 0$$

$$\begin{aligned} \mu_2 &= \mu_2' - \mu_2'^2 \\ &= (0.4546) - (0.0375)^2 = 0.0453 \end{aligned}$$

$$\begin{aligned} \mu_3 &= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 \\ &= (0.0609) - 3(0.4546)(0.0375) + 2(0.0375)^2 \\ &= 0.0600 \end{aligned}$$

$$\begin{aligned} \mu_4 &= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^4 \\ &= (0.5074) - 4(0.0609)(0.0375) \\ &\quad + 6(0.0375)^2(0.4546) - 3(0.0375)^4 \\ &= 3.2385 \end{aligned}$$

(iv) By definition of β_1 and β_2 we get,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.0600)^2}{(0.0453)^3} = 32.22$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3.2385}{(0.0453)^2} = 1578.21$$

Since $\beta_1 > 3$, the distribution is lepto-curtic i.e. it is peaked up sharply than the normal distribution.

UEx. 2.3.2 (SPPU - Dec. 98)

From the following frequency distribution compute the standard deviation of 100 students.

Table P. 2.3.2

Mass in kg.	No. of students
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8

☑ **Soln. :**

We construct the table,

Let $\bar{x} = 67$ be assumed mean

and $h = 3$ = class width

Let $u = \frac{x - 67}{3}$

Table P. 2.3.2(a)

Class of masses	Midpoint of Classes x	Number of students f	$u = \frac{x-67}{3}$	fu	fu^2
60-62	61	5	– 2	– 10	20
63-65	64	18	– 1	– 18	18
66-68	67	42	0	0	0
69-71	70	27	1	27	27
72-74	73	8	2	16	32
Total ($\sum$)		100	0	15	97

We have $\sum f \cdot u = 15$, $\sum fu^2 = 97$

$h = 3$ and $N = \sum f = 100$

By definition, $\sigma = h\sqrt{\frac{1}{N}\sum fu^2 - \left(\frac{1}{N}\sum f_n\right)^2}$

$$= 3\sqrt{\frac{1}{100}(97) - \left(\frac{15}{100}\right)^2}$$

$\therefore \sigma = 2.92$

2.4 THE MEDIAN

Unit II In Sem.

- The median of a set of n observations $x_1, x_2, \ldots\ldots, x_n$ is the middle value when the observations are arranged in an array according to their order of magnitude.
- If n is odd, the middle value which is the $\left(\frac{n}{2}+1\right)^{th}$ in the ascending order of magnitude is unique and is the median.
- If n is even, there are two middle values and the average of three values is the median.
- For example, the median of the set 2, 3, 5, 6, 7 is 5 and that of the set – 3, – 1, 0, 1, 2, 3 is $\frac{0+1}{2} = 0.5$.
- The median is the value which divides the set of observations into two equal halves, such that 50% of the observations lie below the median and 50% above the median.

2.4.1 The Median of Grouped Data

In case of grouped data, median lies in a class interval. To find its value, we use the formula,

$$M = l_1 + \frac{l_2 - l_1}{f_1}(m - C)$$

M = Median

l_1 = The lower limit of the class in which median lies

l_2 = The upper limit of the class in which median lies

f_1 = The frequency of the class in which the median lies

m = Middle item and

C = Cumulative frequency of the group preceding the median group

Let us take an example.

Ex. 2.4.1 : Find the median of the following distribution :

Table P. 2.4.1

Class interval Rs.	Frequencies
1-3	6
3-5	53
5-7	85
7-9	56
9-11	21
11-13	16
13-15	4
15-17	4
Total	**245**

☑ **Soln. :**

Here N = Total frequency = 245

$\therefore \frac{N}{2} = 122.5$

We prepare the table with cumulative frequency.

Table P. 2.4.1(a)

Class-interval	Frequency	Cumulative frequency
1-3	6	6
3-5	53	59
5-7	85	144
7-9	56	200
9-11	21	221
11-13	16	237
13-15	4	241
15-17	4	245

Now, median = The value of $\frac{N}{2}$ i.e.

122.5 items, which lies in 5 – 7 group,

Applying the formula,

$$M = l_1 + \frac{l_2 - l_1}{f_1} (m - C) \text{ we get}$$

$$M = 5 + \frac{7-5}{85} (122.5 - 59) = 6.5$$

2.5 MODE

The word mode comes from the French word '*la* mode' (which means the fashion). The mode is the observation which occurs most frequently in a set.

2.5.1 Calculation of Mode

(1) The computation of mode from a set of observations is simple enough. It is the observation which occurs most frequently. For example the mode of 3, 2, 3, – 1, 0, – 3, 2, 5 is 3. The mode may not exist for certain sets such as set – 1, 0, 1, 5. Also note that there may be more than one mode.

(2) For grouped data, mode is determined as follows :

$$M = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) h$$

Where M is mode,

L_1 = Lower boundary of the modal class.

Δ_1 = Excess of the modal class frequency over the frequency of the class to its left and

Δ_2 = Excess of the modal class frequency over the frequency of the class to its right

h = Size of the class-intervals.

Ex. 2.5.1 : Calculate the mode for the given data :

Table P. 2.5.1

Class-interval (cm)	Frequency
145-146	2
147-148	5
149-150	8
151-152	15
153-154	9
155-156	6
157-158	4
159-160	1
Total	**50**

☑ **Soln. :** First we find the modal class i.e. the interval which corresponds to the maximum frequency. Here the modal class is 151-152. Class midpoint is 151.5. Its lower boundary is 150.5 and upper boundary is 152.5.

Again $h = 2, \Delta_1 = 15 - 8 = 7, \Delta_2 = 15 - 9 = 6.$

$$\therefore \quad M = 150.5 + \left(\frac{7}{7+6}\right) \times 2 = 150.577$$

2.6 GEOMETRIC AND HARMONIC MEAN

There are two other averages, the geometric mean and the harmonic mean which are sometimes used.

(1) The Geometric Mean (G.M.) of a set of observations is such that its logarithm is equal to the Arithmetic mean of the logarithms of the values of the observations. This is given by the same formula even if the observations occur with certain frequencies.

Consider the set of values 2, 3, 5, 6 which occur with frequencies 10, 16, 24, 10 respectively. If x is the Geometric Mean then,

$$\log x = \frac{10 \log 2 + 16 \log 3 + 24 \log 5 + 10 \log 6}{10 + 16 + 24 + 10}$$

$$= 0.5867$$

$$\therefore \quad x = 3.86$$

(2) The harmonic mean (H.M.) of a set of observations is such that its reciprocal is the arithmetic mean (A.M.) of the reciprocals of the values of the observations. Consider the above set of values. The harmonic mean y of the set of values is given by,

$$\frac{1}{y} = \frac{10 \times \frac{1}{2} + 16 \times \frac{1}{3} + 24 \times \frac{1}{5} + 10 \times \frac{1}{6}}{60}$$

$$= 0.28$$

$$\therefore \quad y = 3.57$$

Note :

1. The Geometric mean can be found only if the values assumed by the observations are positive.
2. It can be shown that A.M. ≥ G.M. ≥ H.M.

2.6.1 Other Measures of Location, Quartiles, Deciles and Percentiles

- We have seen that the median of a set of measurements is the value which divides the set into two equal halves, each containing 50% of the measurements.
- In the same way, some other measures of location can be considered. We define the three quartiles, Q_1, Q_2 and Q_3.
- They are such that when the measurements are arranged in increasing order, they divide the set of measurements into four equal parts, the first quartile Q_1 contains the 25% of the measurement, the second quartile Q_2 contains 50% of the measurements and the third quartile Q_3 contains 75% of the measurements.
- Actually the second quartile Q_2 is the median.
- Similarly we define the deciles. The first decile D_1 contains 10% of the measurements, the second decile D_2 contains 20% of the measurements and so on.
- The fifth decile is the median.
- In the same manner, we define percentiles. The 99 percentiles $P_1, \ldots P_{99}$ divide the set of measurements into 100 equal parts.
- The first percentile P_1 contains 1% of the measurements, the second percentile P_2 contains 2% of the measurements and so on, the 12^{th} percentile contains 12% of the measurements.
- The 50^{th} percentile is therefore the median. The method of finding out the quartiles, deciles and percentiles is basically the same as that of finding the median.
- The median divides the set of observations into two equal values, each containing 50% of the measurements, the 3^{rd} decile divides the set into two parts, the first part being 30% of the set and the other containing 70% of the observations.

Ex. 2.6.1 : The distribution of fortnightly wages of 280 employees of an undertaking is as given. Find the first quartile, the median and the third quartile, find D_4, P_{66}, P_{10} and P_{90}.

Table P. 2.6.1

Fortnightly wages (Rs.)	Frequency
Less than 200	12
200-400	16
400-600	38
600-800	78
800-1000	80
1000-1200	35
1200-1400	14
Above 1400	7
Total	**280**

Unit II In Sem.

☑ **Soln. :** First we prepare cumulative frequency table.

Table P. 2.6.1(a)

Wages (Rs.)	Frequency	Cumulative Frequency
Less than 200	12	12
200-400	16	28
400-600	38	66
600-800	78	144
800-1000	80	224
1000-1200	35	259
1200-1400	14	273
Above 1400	7	280

▶ **Step I :** To find quartiles

The observation for the first quartile Q_1 corresponds to $\frac{280}{4} = 70^{th}$ observations, which lies in the interval 600-800, with lower class boundary 600. This interval contains 78 observations and the interval preceding this contains 66 observations. Hence,

$$Q_1 = l_1 + \frac{l_2 - l_1}{f_1}(m - C)$$

where l_1 = Lower limit of the class in which Q_1 lies

l_2 = The upper limit of HM class in which Q_2 lies

f_1 = Positive frequency of the class

$m = \frac{N}{4}$

C = Cumulative frequency of the group preceding the Q_1 class.

$$\therefore Q_1 = 600 + \frac{200}{78}\left(\frac{280}{4} - 66\right)$$

$$= 610.25 \text{ Rs.}$$

The median which is the second quartile Q_2, is given by,

$$Q_2 = 600 + \frac{200}{78}\left(\frac{280}{2} - 66\right)$$

$$= 600 + 189.74 = 789.74 \text{ Rs.}$$

The third quartile Q_3 is given by,

$$Q_3 = 800 + \frac{200}{80}\left[280 \times \frac{3}{4} - 144\right]$$

$$= 965 \text{ Rs.}$$

▶ **Step II :** The observation for 4^{th} decile corresponds to the $\frac{280 \times 4}{10} = 112^{th}$ observation, which lies in the interval 600 – 800. Hence,

$$D_4 = 600 + \frac{200}{78}(112 - 66) = 717.95 \text{ Rs.}$$

▶ **Step III :** The observation for 66^{th} percentile corresponds to $280 \times \frac{66}{100} = 184^{th}$ observation which lies in the interval 800 – 1000. Thus the 66^{th} percentile P_{66} is given by,

$$P_{66} = 800 + \frac{200}{80}(184.8 - 144) = 902 \text{ Rs.}$$

In the same way,

$$P_{10} = 400 + 0 = 400$$

and

$$P_{90} = 1000 + \frac{200}{35}\left[\frac{280 \times 90}{100} - 220\right]$$

$$= \mathbf{1182.96 \text{ Rs.}}$$

2.7 SEMI-INTERQUARTILE RANGE

The range of set of numbers is the difference between the largest and the smallest items of the set. The range is a very crude measure. It does not tell us about the distribution of the values of the set relative to the average.

This is a more refined form of range. It is defined by,

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

And is called as semi-interquartile range.

Q_1 = First quartile ; Q_3 = Third quartile

2.8 THE MEAN DEVIATION

Consider a set of observations, $x_1, x_2, ..., x_n$. The mean deviation (or average deviation) is defined by,

$$\text{M.D.} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

where, $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

is the arithmetic mean and $|x_i - \bar{x}|$ is the absolute value of the deviation.

Ex. 2.8.1 : Find the mean deviation of the set of measurement 1, 3, 8.

Soln. : Here the arithmetic mean,

$$\bar{x} = \frac{1+3+8}{3} = 4$$

$$\therefore \quad \text{M.D.} = \frac{|1-4|+|3-4|+|8-4|}{3} = 2.67$$

2.8.1 Mean Deviation for Grouped Data

Let $x_1, x_2, \ldots, x_n$ occur with the corresponding frequencies $f_1, f_2, \ldots, f_n$, then

$$\text{M.D.} = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{\sum_{i=1}^{n} f_i}$$

where, $$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

Note that the above formula is also applicable in the case of a frequency distribution whose class intervals have mid-points $x_1, x_2, \ldots, x_n$ and the classes have frequencies $f_1, f_2, \ldots f_n$.

Ex. 2.8.2 : Calculate the mean deviation from the mean of the following distribution :

Table P. 2.8.2

Marks	Number of students
0-10	5
20-20	8
20-30	15
30-40	16
40-50	6
Total	**50**

Soln. : We first calculate mean then find mean deviation.

Table P. 2.8.2(a)

Mid value	$u = \frac{x-25}{10}$	f	fu	$x - \bar{x}$	$f\|x - \bar{x}\|$
5	– 2	5	– 10	– 22	110
15	– 1	8	– 8	– 12	96
25	0	15	0	– 2	30
35	1	16	16	8	128
45	2	6	12	18	108
Total		**50**	**10**		**472**

Here, Mean $= 25 + \frac{10}{50} \times 10 = 27$ marks and

Mean deviation $= \frac{\sum f_i |x_i - \bar{x}|}{\sum f_i} = \frac{472}{50} = 9.44$ marks

Ex. 2.8.3 : The mean annual salary paid to all employees of a company was Rs. 5000. The mean annual salaries paid to male and female employees were Rs. 5200 and Rs. 4200 respectively. Determine the percentage of males and females employed by the company.

Soln. : Let n_1 and n_2 represent percentage of males and females respectively.

Then $n_1 + n_2 = 100$ ...(1)

Now, mean annual salary of males $= \overline{x_1} = 5200$ Rs.

Mean annual salary of females $= \overline{x_2} = 4200$ Rs.

Mean annual salary of all employees $= \bar{x} = 5000$ Rs.

$$\text{Now,} \quad \bar{x} = \frac{n_1 \overline{x_1} + n_2 \overline{x_2}}{n_1 + n_2} \quad \text{(Note the formula)}$$

$$\therefore \quad 5000 = \frac{n_1\, 5200 + n_2\, 4200}{100}$$

$\therefore \quad 26\, n_1 + 21\, n_2 = 2500$ ...(2)

From Equations (1) and (2), $n_1 = 80$, $n_2 = 20$.

$\therefore$ Percentages are 80 and 20.

Ex. 2.8.4 : The first of the two samples has 100 items with mean 15 and standard deviation 3. If the whole group has 250 items with mean 15.6 and standard deviation $\sqrt{13.44}$, find the standard deviation of the second group.

Soln. : We have, $n_1 = 100$, $\overline{x_1} = 15$, $\sigma_1 = 3$

$$n = n_1 + n_2 = 250$$

$$\bar{x} = 15.6$$

$$\sigma = \sqrt{13.44}$$

Unit II In Sem.

$\therefore n_2 = 250 - 100 = 150$

We use the formula,

$$\bar{x} = \frac{n_1\bar{x_1} + n_2\bar{x_2}}{n_1 + n_2}$$

$$\therefore 15.6 = \frac{100(15) + 150\,\bar{x_2}}{250}$$

$$\therefore \bar{x_2} = 16$$

The variance of the combined group σ^2 is given by the formula,

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}$$

(Note the formula)

Where $d_1 = \bar{x_1} - \bar{x} = 15 - 15.6 = -0.6$

and $d_2 = \bar{x_2} - \bar{x} = 16 - 15.6 = 0.4$

$\because \sigma = \sqrt{13.44}$

$$\therefore 13.44 = \frac{100(9) + 150\,\sigma_2^2}{250} + \frac{100(-0.6)^2 + 250(0.4)^2}{250}$$

$$\therefore 250 \times 13.44 = 900 + 36 + 40 + 150\,\sigma_2^2$$

$$\therefore \sigma_2 = 4$$

2.9 PROBABILITY

- So far we have been looking at populations of observations, and at samples drawn from these populations. We have developed ways of summarising real-life data, so that we can make references from them.
- In order to devise mathematical ways of making inferences, we shall need to set up mathematical models of the real-life situations which underlie our data. These models make use of probability theory.
- Suppose that a supermarket employs 40 people, of whom 15 work full-time and the remaining 25 part-time. We can select one employee at random.
- Now the question is, will the one selected be a full-time or a part-time employee ? After selection, we shall know; before selection we do not know, but we do have the information that there are 15 who work full time and 25 part-time.

  Can we use this information to predict what our selection will be ?
- By making a random selection of one from the 40, the relative frequency of full times among the 40 is $\frac{15}{40}$ and of part-timers $\frac{25}{40}$. Since every one of the 40 is equally likely to be chosen, we say that the probability of choosing a full-timer is $\frac{15}{40}\left(=\frac{3}{8}\right)$, and the probability of choosing a part-times is $\frac{25}{40}\left(=\frac{5}{8}\right)$
- Thus, we have set the probability of selection of one type of worker, equal to the relative frequency of that type of worker in the whole population.
- We can also say that the probability of selecting a full-time employee is the ratio

$$= \frac{\text{Number of ways of selecting one full-time employee}}{\text{Total number of ways of selecting one employee}}$$

$$= \frac{15}{40} = \frac{3}{8}$$

Working definition of probability

- In n equally likely ways, r of these ways lead to a particular result, the probability of obtaining this result in one run of the process is $\frac{r}{n}$.
- Let us call the probability of selecting a full-timer be p. Since it is a ratio, it is clearly a ratio, a positive number between 0 and 1.
- Also, the probability of selecting a full-timer is p, then the probability of not doing so is (1 – p).

Terminology and notation

We define some terminologies used in probability theory.

❒ **Definition : Trial :** A trial is a process which, when repeated, generates a set of results or observations.

For example, tossing a coin, measuring journey time are trials.

❒ **Definition :** An outcome is the result of carrying-out a trial.

For example, noting H or T when tossing a coin is an outcome.

❒ **Definition :** An **event** is a set of which consists of one or more of the possible outcomes of a trial.

An event is thus a subset of all possible outcomes.

2.9.1 Inverse Probability

- One of the important applications of the conditional probability is the computation of unknown probabilities, on the basis of information supplied by the experiment or past records.
- For example, suppose an event has occurred through one of the various mutually disjoint events. Then the conditional probability that it has occurred due to a particular event is called its inverse or posterior probability.
- These probabilities are computed by Boye's rule.
- The revision of old (given) probabilities with the help of additional information supplied by the experiment is of extreme importance to business and management executives in arriving at valid decisions in the face of uncertainties.

2.9.2 Conditional Probability

UQ. Explain the following : (i) Conditional probability **(SPPU – Q. 4(b), Dec. 18, 3 Marks)**

It is defined as the probability of an event A given B. It is equal to the probability of B and A happening together divided by the probability of B.

For example : Assume two partially intersecting sets A and B as shown :

Set A represents one set of events and set B represents another. We wish to calculate the probability of A given B has already happened.

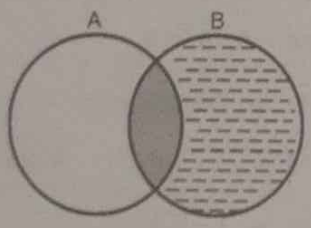


Fig. 2.9.1

Now since B has happened, the part which now matters for A is the part shaded which is actually $A \cap B$.

$\therefore$ Probability of A given B turns out to be

$$= \frac{\text{shaded area}}{\text{dotted area + shaded area}}$$

$$\therefore \quad P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \ldots(2.9.1)$$

Also, we can write the formula for event B given A has already occurred.

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \qquad \ldots(2.9.2)$$

From Equation (ii), we can write

$$P(A \cap B) = P(A) \cdot P(B/A) \qquad \ldots(2.9.3)$$

Now, the first Equation (i), can be written as, using Equation (iii),

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \qquad \ldots(2.9.4)$$

This is known as conditional probability.

Remark

Bayes' theorem is built on top of conditional probability and lies in the heart of Bayesian Inference.

2.10 ALGEBRA OF EVENTS : FOR EVENTS A,B,C

(i) $A \cup B = [\{e \in S \mid e \in A \text{ or } e \in B\}$

(ii) $A \cap B = \{e \in S \mid e \in A, \text{ and } e \in B\}$

(iii) $\bar{A}$ (A complement) $= \{e \in S \mid e \notin A\}$

(iv) $A - B = \{e \in S \mid e \in A \text{ but } e \notin B\}$

(v) $A \subset B \Rightarrow$ every $e \in A, e \in B$

$A \subset B \Rightarrow B \supset A.$

(vi) A = B if and only if A and B have same elements.

(vii) A and B disjoint (mutually exclusive) $\Rightarrow A \cap B = \phi$ (empty set).

(viii) $A \cup B = A + B$ if A and B are disjoint

(ix) $A \Delta B$ denotes those 'e' belonging to exactly one of A and B, i.e.

$$A\Delta B = \bar{A}B \cup A\bar{B} = \bar{A}B + A\bar{B} \quad \text{(disjoint events).}$$

(x) De-morgan's Laws :

$$\overline{A \cup B} = \bar{A} \cap \bar{B} \quad \text{and} \quad \overline{A \cap B} = \bar{A} \cup \bar{B}$$

Unit II In Sem.

(xi) Laws of distribution :

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

2.11 TABLE OF PROBABILITY TERMS

Sr. No.	Statement	Meaning
1.	At least one of the events A and B occur	$\omega \in A \cup B$
2.	Both the events A and B occur	$\omega \in A \cap B$
3.	Neither A nor B occurs	$\omega \in \bar{A} \cap \bar{B}$
4.	Event A occurs and B does not occur.	$w \in A \cap \bar{B}$
5.	Exactly one of the events A or B occurs	$w \in A \Delta B$
6.	If event A occurs, so does B	$A \subset B$
7.	Events A and B are mutually exclusive	$A \cap B = \phi$
8.	Complementary event of A	$\bar{A}$
9.	Sample space	Universal set S.

2.12 SOLVED EXAMPLE ON GIVEN EVENT IN TERMS OF STANDARD EVENTS

To express one of the required events in terms of another given events :

Ex. 2.12.1 : A , B and C are three arbitrary events. Find expression for the events noted below, in the context of A, B and C.

(i) Only A occurs (ii) Both A and B but not C, occur.

(iii) All three events occur. (iv) At least one occurs.

(v) At least two occur. (vi) One and no more occur.

(vii) Two and no more occur. (viii) None occurs.

☑ **Soln. :**

(i) $A \cap \bar{B} \cap \bar{C}$ (ii) $A \cap B \cap \bar{C}$

(iii) $A \cap B \cap C$ (iv) $A \cup B \cup C$

(v) $(A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C) \cup (A \cap B \cap C)$

(vi) $(A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap \bar{C}) \cup (\bar{A} \cap \bar{B} \cap C)$

(vii) $(A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap C) \cup (A \cap \bar{B} \cap C)$

(viii) $(\bar{A} \cap \bar{B} \cap \bar{C})$ or $\overline{(A \cup B \cup C)}$

2.13 AXIOMATIC DEFINITION OF PROBABILITY

GQ. State the three axioms of probability.

❒ **Definition :** Let S be the sample space. Let A be the event (or element) of S, then

(1) $P(A) \geq 0$ (Axiom of non-negativity)

(2) $P(S) = 1$ (Axiom of certainty)

(3) $P(A \cup B) = P(A) + P(B)$ (Axiom of Additivity)

where A and B are disjoint events.

2.14 THEOREMS ON PROBABILITY OF EVENTS

▶ **Theorem (1)**

(1) Probability of an impossible event ϕ is zero,

i.e. $P(\phi) = 0$.

Proof : $\because S \cup \phi = S$

$\therefore P(S \cup \phi) = P(S) = 1$ (Axiom 2)

$\therefore P(S) + P(\phi) = 1$ (Axiom 3)

$\therefore P(\phi) = 0$

(2) Complimentary events : The events A and $\bar{A}$ where $\bar{A}$ is the complement of A in S are called complimentary events.

▶ **Theorem (2) :** $P(\bar{A}) = 1 - P(A)$

Proof : We have $A \cup \bar{A} = S$

$\therefore P(A \cup \bar{A}) = P(S) = 1$

$\therefore P(A) + P(\bar{A}) = 1$, ($\because$ A and $\bar{A}$ are disjoint)

$\therefore P(\bar{A}) = 1 - P(A)$

☞ **Probability of union of Two event**

Let A and B be two events, then the probability of A or B denoted by $P(A \cup B)$ is given by

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

☞ **Corollary**

If A and B are mutually disjoint then,

$A \cap B = \phi$

$\therefore P(A \cap B) = P(\phi) = 0$

$\therefore P(A \cup B) = P(A) + P(B)$

2.15 SOLVED EXAMPLES ON AXIOMATIC PROBABILITY

Ex. 2.15.1 : A,B,C are bidding for a contract. A has exactly half the chance that B has; B in turn is $\frac{4}{5}$ as likely as C to win the contract. What is the probability for each to win the contract if the contract is to be given to one of them.

☑ **Soln. :**

▶ **Step (I) :** Since the events A,B and C are exclusive,

$\therefore P(A) + P(B) + P(C) = 1$ ...(i)

Now, $P(A) = \frac{1}{2} P(B)$ and ...(ii)

$P(B) = \frac{4}{5} P(C)$ ...(iii)

▶ **Step (II) :** Let p be the probability of C, from (i),

$$\frac{1}{2} \cdot \frac{4}{5} P(C) + \frac{4}{5} P(C) + P(C) = 1$$

$$\therefore \left(\frac{2}{5} + \frac{4}{5} + 1\right) P = 1$$

$$\therefore P = \frac{5}{11}$$

$\therefore P(A) = \frac{2}{11}$, $P(B) = \frac{4}{11}$, $P(C) = \frac{5}{11}$ ...Ans.

Ex. 2.15.2 : A ball is drawn at random from a box containing 12 red, 18 white, 10 blue and 15 orange balls. Find the probability that

(i) It is red or blue (ii) white, blue or orange

(iii) Neither white nor orange.

☑ **Soln. :**

▶ **Step (I) :** Let A,B,C and D be the events of drawing red, white, blue and orange balls respectively. Total number of balls = 12 + 18 + 10 + 15 = 55. Probability of drawing one red ball is

$$= P(A) = \frac{12C_1}{55C_1} = \frac{12}{55}, \quad \text{Similarly}$$

and $P(B) = \frac{18C_1}{55C_1} = \frac{18}{55}$, $P(C) = \frac{10C_1}{55C_1} = \frac{10}{55}$

and $P(D) = \frac{15C_1}{55C_1} = \frac{15}{55}$

▶ **Step (II) :** Probability of a red or blue ball

$$= P(A \cup C) = P(A) + P(C)$$

$\because$ A and C are mutually disjoint

$$= \frac{12}{55} + \frac{10}{55} = \frac{22}{55} = \frac{2}{5}$$

(i) Probability of white, blue or orange

$$= P(B \cup C \cup D) = P(B) + P(C) + P(D)$$

$\because$ B, C, D are mutually disjoint

$$= \frac{18}{55} + \frac{10}{55} + \frac{15}{55} = \frac{43}{55}$$

(ii) Probability of neither white nor orange

$$= 1 - P(A \cup B) = 1 - [P(A) + P(B)]$$

$$= 1 - \left[\frac{18}{55} + \frac{15}{55}\right] = 1 - \frac{33}{55} = \frac{22}{55} = \frac{2}{5} \quad \text{...Ans.}$$

Ex. 2.15.3 : If two dice are thrown, what is the probability that the sum is greater than 8.

☑ **Soln. :**

▶ **Step (I) :** Let S denote the sum on the two dice, then we want $P(S > 8)$.

The required event can happen in the following mutually exclusive ways :

(i) S = 9, (ii) S = 10, (iii) S = 11, (iv) S = 12.

$\therefore$ By addition theorem

$P(S > 8) = P(S = 9) + P(S = 10) + P(S = 11) + P(s + 12)$. ...(i)

▶ **Step (II) :** In a throw of two dice, the sample space contains $6^2 = 36$ points.

The number of favourable cases are as follows :

S = 9 : (3,6), (6,3), (4,5), (5,4) i.e. 4 sample points

Unit II In Sem.

$\therefore P(S=9) = \frac{4}{36}$

$S = 10 : (4,6), (6,4), (5,5)$, i.e. 3 points.

$\therefore P(s=10) = \frac{3}{36}$

$S = 11 : (5,6), (6,5)$, i.e. 2 sample points.

$\therefore P(S=11) = \frac{2}{36}$

$S = 12 : (6,6)$, i.e. 1 sample point

$\therefore P(S=12) = \frac{1}{36}$

$\therefore$ Required probability

$= \frac{4}{36}+\frac{3}{36}+\frac{2}{36}+\frac{1}{36} = \frac{10}{36} = \frac{5}{18}$ ...Ans.

Ex. 2.15.4 : A card is drawn from a pack of 52 cards. Find the probability of getting a king or a heart or a red card.

Soln. :

Step (I) : Let the events be :

A = the card drawn is a king

B = the card drawn is a heart

C = the card drawn is a red card.

Note that A, B, C are nor mutually exclusive $A \cap B$ = the card drawn is the king of hearts

$\therefore n(A \cap B) = 1$; $\therefore p(A \cap B) = \frac{1}{52}$

$B \cap C = B$: the card drawn is a heart ($\because B \subset C$)

$\therefore n(B \cap C) = 13$

$\therefore p(B \cap C) = \frac{13}{52}$ $C \cap A$: the card drawn is a red king

$n(C \cap A) = 2$

$\therefore p(C \cap A) = \frac{2}{52}$

and $A \cap B \cap C = A \cap B$: the card drawn is the king of hearts

$n(A \cap B \cap C) = 1$; $\therefore P(A \cap B \cap C) = \frac{1}{52}$

Step (II) : The required probability of getting a king or heart or a red card is given by

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C)$ ...(i)

$\because P(A) = \frac{4}{52}$, $P(B) = \frac{13}{52}$, $P(C) = \frac{26}{52}$ from Equation (i)

$P(A \cup B \cup C) = \frac{4}{52}+\frac{13}{52}+\frac{26}{52}-\frac{1}{52}-\frac{13}{52}-\frac{2}{52}+\frac{1}{52}$

$= \frac{7}{13}$ ...Ans.

UEx. 2.15.5 (SPPU – Dec. 20, 3 Marks)

If A and B are independent events with P(A) = 0.26, P(B) = 0.45, find

(a) $P(A \cap B)$ (b) $P(A \cap \bar{B})$, (c) $P(\bar{A} \cap \bar{B})$

Soln. :

(a) Since A and B are independent events

$\therefore P(A \cap B) = P(A) \cdot P(B) = (0.26)(0.45) = 0.117$ ...(i)

(b) We have

$P(A \cap \bar{B}) = P(A) - P(A \cap B)$

$= 0.26 - 0.117 = 0.143$ from (i)

(c) $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$

$= 1 - [P(A) + P(B) - P(A \cap B)]$

$= 1 - [0.26 + 0.45 - 0.117] = 0.407$

Ex. 2.15.6 : A fair dice is thrown thrice. Find the probability that the sum of the numbers obtained is 10.

Soln. :

Step (I) : When a dice is thrown thrice, total number of cases = $6 \times 6 \times 6 = 216$

$\therefore n = 216$

The number of favourable cases are as follows :

(1,4,5) in 3! Ways = 6 ways

(1,6,3) in 3! Ways = 6 ways

(2,5,3) in 3! Ways = 6 ways

(2,4,4) in $3C_2$ ways = $\frac{3!}{2!} = 3$ ways

(2,6,2) in $3C_2$ ways = 3 ways

(3,3,4) in $3C_2 = 3$ ways.

Step (II) : Total number of ways of getting the sum 10 is = 6 + 6 + 6 + 3 + 3 + 3 = 27 ways.

$\therefore$ Probability = $\frac{27}{216} = \frac{1}{8}$ ...Ans.

Ex. 2.15.7 : A letter of the English alphabet is chosen at random. Calculate the probability that the letter so chosen

(i) is a vowel, (ii) precedes m and is a vowel, (iii) follows m an is a vowel.

Soln. :

Step (I) : The sample space S of the experiment

Is S = {a,b,c,d, ...,y,z}, $\therefore n(S) = 26$.

Step (II)

(i) Let A be the event that the letter chosen is a vowel.

Thus, A = {a, e, i, o, u}, n (A) = 5

$\therefore P(A) = \frac{5}{26}$

(ii) Let B be the event that the letter precedes m and is a vowel, then B = {a,e,i}, n (B) = 3

$\therefore P(B) = \frac{3}{26}$

(iii) Let C be the event that the letter follows m and is a vowel,

Then C = {o, u}, n (C) = 2

$\therefore P(C) = \frac{2}{26} = \frac{1}{13}$ ...Ans.

UEx. 2.15.8 (SPPU – Dec. 19, 4 Marks)

If 5 of 20 tires in storage are defective and 5 of them are randomly chosen for inspection (that is, each tire has the same chance of being selected), what is the probability that the two of the defective tires will be included ?

Soln. :

Total number of tires = 20

5 tires can be selected in $20 C_5$ ways

$= \frac{20 \times 19 \times 18 \times 17 \times 16}{5 \times 4 \times 3 \times 2}$

$= 19 \times 3 \times 17 \times 16 = 15,504$

Let A be the event that 2 defective balls are chosen

$\therefore n(A) = 5C_2 = \frac{5.4}{2} = 10$

$\therefore P(A) = \frac{10}{15,504} = \frac{1}{1550.4} = 0.00065$

2.16 MULTIPLICATION THEOREM OF PROBABILITY

Theorem : *For two events A and B*

$P(A \cap B) = P(A) \cdot P(B/A)$, $P(A) > 0$

$= P(B) \cdot P(A/B)$, $P(B) > 0$

Proof : In the usual notations, we have

$P(A) = \frac{n(A)}{n(S)}$, $P(B) = \frac{n(B)}{n(S)}$, and

$P(A \cap B) = \frac{n(A \cap B)}{n(S)}$ ...(2.16.1)

Now, for the event A/B, the sample space is B and out of the n(B) sample points, $n(A \cap B)$ stands for the occurrence of the event A. Hence

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

Now, from Equation (2.16.1),

$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{n(B)}{n(S)} \cdot \frac{n(A \cap B)}{n(B)}$

$= P(B) \cdot P(A/B)$ ...(2.16.2)

Similarly, we can show that

$P(A \cap B) = P(A) \cdot P(B/A)$ ...(2.16.3)

2.16.1 Some Important Results

(1) If A and B are exclusive events, then P(A/B) = 0 and P(B/A) = 0

Proof : $\because$ A and B are exclusive event.

$\therefore A \cap B = \phi$

$\therefore P(A \cap B) = 0$

Now, $P(A \cap B) = P(A) \cdot P(B/A) = 0$

$\therefore P(B/A) = 0$ $\because P(A) > 0$

Similarly, $P(A \cap B) = P(B) \cdot P(A/B) = 0$

$\therefore P(A/B) = 0$

(2) If $A \subset B$ then $P(B/A) = 1$ and $P(A/B) = \frac{P(A)}{P(B)}$

Proof : $\because A \subset B$

$\therefore A \cap B = A$

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

Now, $P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A)}{P(A)} = 1$

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

(3) If A and B are any two events, then

$P(A/\bar{B}) = \frac{P(A) - P(A \cap B)}{1 - P(B)}$

Proof :

By definition,

$P(A/\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{P(A) - P(A \cap B)}{1 - P(B)}$

C/O [Theorem 1, Article 1.11]

(4) If A, B are any two events, then $P(\bar{B}/A) = 1 - P(B/A)$.

Proof

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

$$= 1 - \frac{P(A\cap B)}{P(A)} = 1 - P(B/A)$$

(5) For $P(B) > 0$, $P(A/B) \le P(A)$

Proof :

Now, $n(A\cap B) \le n(A)$ and $n(B) \le n(S)$

$$\therefore \frac{n(A\cap B)}{n(B)} \le \frac{n(A)}{n(S)}$$

$$\therefore P(A/B) \le P(A)$$

2.16.2 Solved Examples on Conditional Probability and Independent Events

Ex. 2.16.1 : For any three events A,B and C

$P(A\cup B/C) = P(A/C) + P(B/C) - P(A\cap B/C)$.

Soln. :

Recall that, $(A\cup B)\cap C = (A\cap C)\cup(B\cap C)$

and also, $P(A\cup B) = P(A) + P(B) - P(A\cap C)$

$\therefore P[(A\cup B)\cap C] = P[(A\cap C)\cup(B\cap C)]$

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

Dividing both sides by P(C), we get

$$\frac{P[(A\cup B)\cap C]}{P(C)} = \frac{P(A\cap C)}{P(C)} + \frac{P(B\cap C)}{P(C)} - \frac{P(A\cap B\cap C)}{P(C)}$$

$$\therefore P[(A\cup B)/C] = P(A/C) + P(B/C) - P(A\cap B/C)$$

...Ans.

Ex. 2.16.2 : For any three events A,B and C

$P(A\cap \bar{B}/C) + P(A\cap B/C) = P(A/C)$

Soln. :

$$L.H.S. = \frac{P(A\cap\bar{B}\cap C)}{P(C)} + \frac{P(A\cap B\cap C)}{P(C)}$$

$$= \frac{P(A\cap\bar{B}\cap C) + P(A\cap B\cap C)}{P(C)}$$

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

...Ans.

2.17 INDEPENDENCE

- In discussing conditional probability, we considered two events A and B which could be inter-related; one of them could give information about the other. It is perfectly possible to carry out two events in such a way that one event gives no information about the other.
- Suppose our trial is a game in which a fair die is thrown once and a fair coin is tossed twice.

  Let A : the score with the die is 4.

  B: both tosses of the coin give heads.
- Since A and B both refer to the same trial, they occur together. But the information on the die is of no help in predicting what may have happened with the coin.

  $\therefore P(B/A) = P(B)$ and similarly, $P(A/B) = P(A)$

  Since $P(A\cap B) = P(A)\cdot P(B/A) = P(A)\cdot P(B)$

Definition : Two events A and B are independent if and only if $P(A\cap B) = P(A)\cdot P(B)$

Theorem (1) : For any event A in S,

(i) A and null event ϕ are independent

(ii) A and S are independent.

Proof : (i) $P(A\cap\phi) = P(\phi) = 0 = P(A)\cdot P(\phi)$

$\therefore$ A and ϕ are independent.

(ii) $P(A\cap S) = P(A) = P(A)\cdot 1 = P(A)\cdot P(S)$

$\therefore$ A and S are independent

Theorem (2) : Multiplication theorem of probability for independent events :

If A and B are two events with $P(A) \neq 0$, $P(B) \neq 0$, then A and B are independent if $P(A\cap B) = P(A)\cdot P(B)$.

Proof : $\because$ A and B are independent,

$\therefore P(A/B) = P(A)$ and $P(B/A) = P(B)$

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

and $P(A\cap B) = P(B)\cdot P(A/B) = P(B)\cdot P(A)$

Theorem (3) : For any three events A,B,C defined on the sample space S such that $B \subset C$ and $P(A) > 0$, $P(B/A) \le P(C/A)$.

Proof : Now,

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

$$= \frac{P[(B\cap C\cap A)\cup(\bar{B}\cap C\cap A)]}{P(A)}$$

Note : $(B\cup D)\cup(\bar{B}\cap D) = D$

$$= \frac{P[(B\cap C\cap A)}{P(A)]} + \frac{P(\bar{B}\cap C\cap A)}{P(A)} = P(B\cap C/A) + P(\bar{B}\cap C/A)$$

$= P(B/A) + P(\bar{B}\cap C/A)$ ($\because$ BCC)

$\because P(\bar{B}\cap C/A) \ge 0$

$\therefore P(C/A) \ge P(B/A)$

Theorem (4) : If A and B are independent events, then

(i) A and $\bar{B}$, (ii) $\bar{A}$ and B,

(iii) $\bar{A}$ and $\bar{B}$ are also independent

Proofs

(i) A and B are independent

$$P(A\cap B) = P(A)\cdot P(B) \quad ...(2.17.1)$$

Now, $P(A\cap\bar{B}) = P(A) - P(A\cap B)$

$= P(A) - P(A)\cdot P(B)$

From Equation (2.17.1)

$= P(A)[1 - P(B)]$

$\therefore P(A\cap\bar{B}) = P(A)\cdot P(\bar{B})$

$\therefore$ A and $\bar{B}$ are independent

(ii) Again, $P(\bar{A}\cap B) = P(B) - P(A\cap B)$

$= P(B) - P(A)\cdot P(B)$ From Equation (2.17.1)

$= P(B)[1 - P(A)] = P(B)\cdot P(\bar{A}) = P(\bar{A})\cdot P(B)$

$\therefore \bar{A}$ and B are independent.

(iii) $P(\bar{A}\cap\bar{B}) = P(\overline{A\cup B}) = 1 - P(A\cup B)$

$= 1 - [P(A) + P(B) - P(A\cap B)]$

$= 1 - P(A) - P(B) + P(A)\cdot P(B)$

From Equation (2.14.2)

$= [1 - P(B)] - P(A)[1 - P(B)]$

$= 1 - P(B)][1 - P/A]$

$= P(\bar{B})\cdot P(\bar{A}) = P(\bar{A})\cdot P(\bar{B})$

$\therefore \bar{A}$ and $\bar{B}$ are also independent

2.18 PAIR WISE INDEPENDENT EVENTS

The events $A_1, A_2, \ldots, A_n$ are said to be pair wise independent if and only if $P(A_i \cap A_j) = P(A_i)\cdot P(A_j)$, $i \neq j = 1, 2, \ldots, n$.

In particular, three events A_1, A_2, A_3 are pair wise independent if and only if

$P(A_1\cap A_2) = P(A_1)\cdot P(A_2)$

$P(A_1\cap A_3) = P(A_1)\cdot P(A_3)$

and $P(A_2\cap A_3) = P(A_2)\cdot P(A_3)$

Unit II In Sem.

2.18.1 Mutually Independent Events

The n events $A_1, A_2, \ldots, A_n$ in a sample space S are said to be mutually independent if

$P(A_1\cap A_2\cap \ldots\ldots\cap A_n) = P(A_1)\cdot P(A_2)\ldots P(A_n)$

2.18.2 Independence and Exclusiveness

Independence and exclusiveness of two events A and B are totally different concepts.

(i) For **independence**, the condition is

$$P(A\cap B) = P(A)\cdot P(B)$$

(ii) For **exclusiveness**, the condition is

$$P(A\cap B) = 0$$

(1) If A and B are independent then they cannot be exclusive,

$\because$ If A and B are independent, then

$P(A\cap B) = P(A)\cdot P(B)$, but since $P(A) \neq 0$, $P(B) \neq 0$, $\therefore P(A\cap B) \neq 0$

$\therefore$ A and B are not exclusive

(2) If A and B are exclusive, then they cannot be independent.

If A and B exclusive, then

$P(A\cap B) = 0$; but $\because P(A) \neq 0, P(B) \neq 0$

We cannot have $P(A\cap B) = P(A)\cdot P(B)$

$\therefore$ A and B are not independent.

Ex. 2.18.1 : An urn A contains 8 cards bearing numbers 1 to 8. Another urn B contains 9 cards bearing numbers 1 to 9. An urn is selected at random and a card is drawn. If the card bears an odd number one more card is drawn from the same

urn. If the card bears an even number the second card is drawn from the other urn. Find the probability that both cards shown (i) even numbers, (ii) odd numbers.

☑ **Soln.** : We observe that there are two possibilities : Either first urn A is selected or urn B is selected.

▶ **Step (I)** : Urn A is selected

∵ there are two urns, probability of selecting an urn A is $\frac{1}{2}$ and getting an even number is $\frac{1}{2}$. Now a ticket is drawn from the other urn B.

Probability of getting an even number from B is $\frac{4}{9}$.

∴ Probability of getting two even numbers.

Since the events are independent, we take product.

$$= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{4}{9} = \frac{1}{9}.$$

▶ **Step (II)** : Urn B is selected

Probability of selecting urn B is $\frac{1}{2}$ and getting an even number is $\frac{4}{9}$. Now a ticket is drawn from the other urn A.

Probability of getting an even number from A is $\frac{1}{2}$.

∴ Probability of getting two even numbers

$$= \frac{1}{2} \cdot \frac{4}{9} \cdot \frac{1}{2} = \frac{1}{9}$$

∴ Required probability $= \frac{1}{9} + \frac{1}{9} = \frac{2}{9}$

▶ **Step (III)** : Urn A is selected

Probability of selecting urn A is $\frac{1}{2}$ and getting an odd number is $\frac{1}{2}$. Now another ticket is drawn from the same urn A. Probability of getting an odd number again is $\frac{3}{7}$.

∴ Probability of getting two odd numbers

$$= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{7} = \frac{3}{28}$$

▶ **Step (IV)** : Urn B is selected

Probability of selecting urn B is $\frac{1}{2}$ and getting an odd number is $\frac{5}{9}$. Now another ticket is drawn from the same urn B. Probability of getting an odd number is $\frac{4}{8} = \frac{1}{2}$.

∴ Probability of getting two odd numbers. Since the events are independent, we take product.

$$= \frac{1}{2} \cdot \frac{5}{9} \cdot \frac{1}{2} = \frac{5}{36}$$

∴ Required probability $= \frac{3}{28} + \frac{5}{36} = \frac{167}{252}$

Since the events are mutually exclusive, we add the probabilities.

Ex. 2.18.2 : There are two bags. The first contains 5 red and 7 white balls and the second contains 3 red and 12 white balls. One ball is taken out at random from the first bag and is put in the second bag. Now a ball is drawn from the second bag. What is the probability that this last ball is red ?

☑ **Soln.** :

We have to consider two possibilities :

(1) The ball transferred from the first bag is **white** or

(2) the ball is red.

▶ **Step (I)** : When the first ball transferred from first bag to the second is white, then probability of this event is $\frac{7}{12}$.

Now in the second bag there are 13 white and 3 red balls.

Probability of drawing red ball $= \frac{3}{16}$

∵ the events are independent,

Probability of combined event $= \frac{7}{12} \times \frac{3}{16} = \frac{7}{64}$.

▶ **Step (II)** : When the red ball is transferred from first bag to second, then, the probability of drawing red ball is $\frac{5}{12}$

Now in the second bag, there are 4 red balls and 12 white balls.

∴ Probability of drawing a red ball from second bag

$$= \frac{4}{16} = \frac{1}{4}$$

∴ Probability (total) $= \frac{5}{12} \times \frac{1}{4} = \frac{5}{48}$

∴ Required probability $= \frac{7}{64} + \frac{5}{48} = \frac{41}{192}$ ...Ans.

Ex. 2.18.3 : A student takes his examination in four subjects A,B,C and D. He estimates his chances of passing in A as $\frac{4}{5}$, in B as $\frac{3}{4}$, in C as $\frac{5}{6}$ and D as $\frac{2}{3}$. To quality he must pass in A and in at least 2 other subjects. What is the probability that he will qualify ?

☑ **Soln.** : The following are the possibilities in order that he qualifies.

A_1 : He passes in A, and also in B, in C, in D

A_2 : He passes in A, and in B and in C and fails in D

A_3 : He passes in A, in B and in D and fails in C

A_4 : He passes in A, in C and in D but not in B.

Now we find the probabilities of these events :

$$P(A_1) = \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{5}{6} \cdot \frac{2}{3} = \frac{1}{3}$$

$$P(A_2) = \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{5}{6} \left(1 - \frac{2}{3}\right) = \frac{1}{6}$$

$$P(A_3) = \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} \left(1 - \frac{5}{6}\right) = \frac{1}{15}$$

$$P(A_4) = \frac{4}{5} \cdot \frac{5}{6} \cdot \frac{2}{3} \left(1 - \frac{3}{4}\right) = \frac{1}{9}$$

Required probability is the sum of all these probabilities.

In any of the options he can be qualified.

∴ Required probability $= \frac{1}{3} + \frac{1}{6} + \frac{1}{15} + \frac{1}{9} = \frac{61}{90}$ ...Ans.

Ex. 2.18.4 : A lot contains 1 percent defective items. What should be the number of items in a random sample so that the probability of finding at least one defective in it is at least 0.95?

☑ **Soln.** :

Let, p = Probability of a defective item $= \frac{1}{100}$

$= 0.01$

q = probability of a non-defective item

$= 1 - 0.01 = 0.99$

In n trials, the probability of non-defective items $= \left(\frac{99}{100}\right)\left(\frac{99}{100}\right) \ldots$ n times $= \left(\frac{99}{100}\right)^n$

∴ Probability of at least one defective in n trials

$$= 1 - \left(\frac{99}{100}\right)^n$$

We want this probability as greater than 0.95

$$\therefore 1 - \left(\frac{99}{100}\right)^n > \frac{95}{100} \quad ; \quad \therefore 1 - \frac{95}{100} > \left(\frac{99}{100}\right)^n$$

$$\therefore \frac{5}{100} > \left(\frac{99}{100}\right)^n$$

Taking logs, $\log(0.05) > n \log(0.99)$

∴ **n > 296** ...Ans.

Ex. 2.18.5 : An urn contains two black balls and 3 white balls. Two balls are drawn at random from the urn. Find the probability that both the balls are white.

☑ **Soln.** : There are totally 5 balls, 2 balls out of 5 balls can be drawn in $5C_2$ ways.

And 2 white balls out of 3 white balls can be drawn in $3C_2$ ways.

∴ Probability of drawing 2 white balls

$$P = \frac{{}^3C_2}{{}^5C_2} = \left(\frac{3 \cdot 2}{2}\right)\left(\frac{2}{5 \cdot 4}\right)$$

$$\therefore P = \frac{3}{10}$$ ...Ans.

Unit II In Sem.

Ex. 2.18.6 : 5 cards are drawn at random from a pack of 52 cards. If all cards are red, what is the probability that all of them are hearts.

☑ **Soln.** : Let A be the event of drawing 5 red cards of heart, let B be the event of drawing red cards.

Then, $P(A/B) = \frac{P(A \cap B)}{P(B)}$ ...(1)

$P(A \cap B) = {}^{13}C_5$

= Probability of 5 heart cards.

and $P(B) = {}^{26}C_5$ (∵ there are 26 red cards)

$$\therefore P(A/B) = \frac{{}^{13}C_5}{{}^{26}C_5} = \frac{13 \cdot 12 \cdot 11 \cdot 10 \cdot 9}{26 \cdot 25 \cdot 24 \cdot 23 \cdot 22}$$

$$= \frac{9}{460} = 0.02$$ ...Ans.

2.19 PARTITION OF SAMPLE SPACE

Let S be sample space and let $A_1, A_2, \ldots, A_n$ be the mutually disjoint subsets of S; i.e.

(i) $A_i \cap A_j = \phi$, $i \neq j = 1$ to n

(ii) $S = A_1 \cup A_2 \cup \ldots \cup A_n$

(iii) $A_1, A_2, \ldots A_n$ are events.

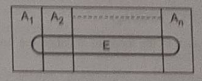


Fig. 2.19.1

urn. If the card bears an even number the second card is drawn from the other urn. Find the probability that both cards shown (i) even numbers, (ii) odd numbers.

☑ **Soln. :** We observe that there are two possibilities : Either first urn A is selected or urn B is selected.

▶ **Step (I) :** Urn A is selected

∵ there are two urns, probability of selecting an urn A is $\frac{1}{2}$ and getting an even number is $\frac{1}{2}$. Now a ticket is drawn from the other urn B.

Probability of getting an even number from B is $\frac{4}{9}$.

∴ Probability of getting two even numbers.

Since the events are independent, we take product.

$$= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{4}{9} = \frac{1}{9}.$$

▶ **Step (II) :** Urn B is selected

Probability of selecting urn B is $\frac{1}{2}$ and getting an even number is $\frac{4}{9}$. Now a ticket is drawn from the other urn A.

Probability of getting an even number from A is $\frac{1}{2}$.

∴ Probability of getting two even numbers

$$= \frac{1}{2} \cdot \frac{4}{9} \cdot \frac{1}{2} = \frac{1}{9}$$

∴ Required probability $= \frac{1}{9} + \frac{1}{9} = \frac{2}{9}$

▶ **Step (III) :** Urn A is selected

Probability of selecting urn A is $\frac{1}{2}$ and getting an odd number is $\frac{1}{2}$. Now another ticket is drawn from the same urn A. Probability of getting an odd number again is $\frac{3}{7}$.

∴ Probability of getting two odd numbers

$$= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{7} = \frac{3}{28}$$

▶ **Step (IV) :** Urn B is selected

Probability of selecting urn B is $\frac{1}{2}$ and getting an odd number is $\frac{5}{9}$. Now another ticket is drawn from the same urn B. Probability of getting an odd number is $\frac{4}{8} = \frac{1}{2}$.

∴ Probability of getting two odd numbers. Since the events are independent, we take product.

$$= \frac{1}{2} \cdot \frac{5}{9} \cdot \frac{1}{2} = \frac{5}{36}$$

∴ Required probability $= \frac{3}{28} + \frac{5}{36} = \frac{167}{252}$

Since the events are mutually exclusive, we add the probabilities.

Ex. 2.18.2 : There are two bags. The first contains 5 red and 7 white balls and the second contains 3 red and 12 white balls. One ball is taken out at random from the first bag and is put in the second bag. Now a ball is drawn from the second bag. What is the probability that this last ball is red ?

☑ **Soln. :**

We have to consider two possibilities :

(1) The ball transferred from the first bag is **white** or

(2) the ball is red.

▶ **Step (I) :** When the first ball transferred from first bag to the second is white, then probability of this event is $\frac{7}{12}$.

Now in the second bag there are 13 white and 3 red balls.

Probability of drawing red ball $= \frac{3}{16}$

∵ the events are independent,

Probability of combined event $= \frac{7}{12} \times \frac{3}{16} = \frac{7}{64}$.

▶ **Step (II) :** When the red ball is transferred from first bag to second, then, the probability of drawing red ball is $\frac{5}{12}$

Now in the second bag, there are 4 red balls and 12 white balls.

∴ Probability of drawing a red ball from second bag

$$= \frac{4}{16} = \frac{1}{4}$$

∴ Probability (total) $= \frac{5}{12} \times \frac{1}{4} = \frac{5}{48}$

∴ Required probability $= \frac{7}{64} + \frac{5}{48} = \frac{41}{192}$ **...Ans.**

Ex. 2.18.3 : A student takes his examination in four subjects A,B,C and D. He estimates his chances of passing in A as $\frac{4}{5}$, in B as $\frac{3}{4}$, in C as $\frac{5}{6}$ and D as $\frac{2}{3}$. To quality he must pass in A and in at least 2 other subjects. What is the probability that he will qualify ?

☑ **Soln. :** The following are the possibilities in order that he qualities.

A_1 : He passes in A, and also in B, in C, in D

A_2 : He passes in A, and in B and in C and fails in D

A_3 : He passes in A, in B and in D and fails in C

A_4 : He passes in A, in C and in D but not in B.

Now we find the probabilities of these events :

$$P(A_1) = \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{5}{6} \cdot \frac{2}{3} = \frac{1}{3}$$

$$P(A_2) = \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{5}{6}\left(1 - \frac{2}{3}\right) = \frac{1}{6}$$

$$P(A_3) = \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{2}{3}\left(1 - \frac{5}{6}\right) = \frac{1}{15}$$

$$P(A_4) = \frac{4}{5} \cdot \frac{5}{6} \cdot \frac{2}{3}\left(1 - \frac{3}{4}\right) = \frac{1}{9}$$

Required probability is the sum of all these probabilities.

In any of the options he can be qualified.

∴ Required probability $= \frac{1}{3} + \frac{1}{6} + \frac{1}{15} + \frac{1}{9} = \frac{61}{90}$ **...Ans.**

Ex. 2.18.4 : A lot contains 1 percent defective items. What should be the number of items in a random sample so that the probability of finding at least one defective in it is at least 0.95?

☑ **Soln. :**

Let, p = Probability of a defective item $= \frac{1}{100}$

$= 0.01$

a = probability of a non-defective item

$= 1 - 0.01 = 0.99$

In n trials, the probability of non-defective items $= \left(\frac{99}{100}\right)\left(\frac{99}{100}\right)$.....n times $= \left(\frac{99}{100}\right)^n$

∴ Probability of at least one defective in n trials

$$= 1 - \left(\frac{99}{100}\right)^n$$

We want this probability as greater than 0.95

$$\therefore 1 - \left(\frac{99}{100}\right)^n > \frac{95}{100} \quad \therefore \quad 1 - \frac{95}{100} > \left(\frac{99}{100}\right)^n$$

$$\therefore \frac{5}{100} > \left(\frac{99}{100}\right)^n$$

Taking logs, $\log(0.05) > n \log(0.99)$

∴ **n > 296** **...Ans.**

Ex. 2.18.5 : An urn contains two black balls and 3 white balls. Two balls are drawn at random from the urn. Find the probability that both the balls are white.

☑ **Soln. :** There are totally 5 balls, 2 balls out of 5 balls can be drawn in $5C_2$ ways.

And 2 white balls out of 3 white balls can be drawn in $3C_2$ ways.

∴ Probability of drawing 2 white balls

$$P = \frac{{}^3C_2}{{}^5C_2} = \left(\frac{3.2}{2}\right)\left(\frac{2}{5.4}\right)$$

$$\therefore \mathbf{P = \frac{3}{10}} \quad \textbf{...Ans.}$$

Ex. 2.18.6 : 5 cards are drawn at random from a pack of 52 cards. If all cards are red, what is the probability that all of them are hearts.

☑ **Soln. :** Let A be the event of drawing 5 red cards of heart, let B be the event of drawing red cards.

Then, $P(A/B) = \frac{P(A \cap B)}{P(B)}$ ...(i)

$P(A \cap B) = {}^{13}C_5$

= Probability of 5 heart cards.

and $P(B) = {}^{26}C_5$ (∵ there are 26 red cards)

$$\therefore P(A/B) = \frac{{}^{13}C_5}{{}^{26}C_5} = \frac{13 \cdot 12 \cdot 11 \cdot 10 \cdot 9}{26 \cdot 25 \cdot 24 \cdot 23 \cdot 22}$$

$$= \frac{9}{460} = \mathbf{0.02} \quad \textbf{...Ans.}$$

2.19 PARTITION OF SAMPLE SPACE

Let S be sample space and let $A_1, A_2, ..., A_n$ be the mutually disjoint subsets of S; i.e.

(i) $A_i \cap A_j = \phi$, $i \neq j = 1$ to n

(ii) $S = A_1 \cup A_2 \cup ... \cup A_n$.

(iii) $A_1, A_2, ... A_n$ are events.

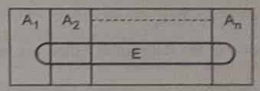


Fig. 2.19.1

Unit II In Sem.

Note : $A_1, A_2, \ldots A_n$ are mutually exclusive and exhaustive events, i.e. when experiment is performed one and only one of the events A_i must occur.

For example, in tossing a die, the events $A_1 = \{3, 4\}$, $A_2 = \{1, 2, 6\}$, $A_3 = \{5\}$ represents a portion of sample space S.

But $B_1 = \{1, 2, 5\}$, $B_2 = \{3, 4, 5, 6\}$ do not represent a portion (∵ Not mutually disjoint). And $C_1 = \{3\}$, $C_2 = \{1, 2, 5, 6\}$ do not represent a portion.

$(\because S \neq C_1 \cup C_2)$

Formation of partition for any event E of S :

Let E be any event in S.

Then $E = E \cap S$

$\because \quad S = A_1 \cup A_2 \cup \ldots \cup A_n$

$\therefore \quad E = E \cap (A_1 \cup A_2 \cup \ldots \cup A_n)$

$= (E \cap A_1) \cup (E \cap A_2) \cup \ldots \cup (E \cap A_n)$

Note that $E \cap A_i$, $i = 1$ to n, are disjoint and they form partition of E.

2.20 LAW OF TOTAL PROBABILITY

GQ. Explain total probability theorem.

Let E be the sample space and $A_1, A_2, \ldots, A_n$, any be events and $\because (E \cap A_i)$ for all $i = 1$ to n are disjoint.

$\therefore$ We have,

$$P(E) = P(E \cap A_1) + P(E \cap A_2) + \ldots + P(E \cap A_n)$$

Using the multiplication theorem for conditional probability i.e.

$$P(E \cap A_i) = P(A_i \cap E) = P(A_i) \cdot P(E / A_i)$$

We have the Theorem.

Theorem

$$P(E) = P(A_1) \cdot P(E / A_1) + P(A_2) \cdot P(E/A_2) + \ldots + P(A_n) P(E/A_n)$$

The equation in the theorem is called the **law of total probability**.

We again observe that the sets $A_1, A_2, \ldots A_n$ are pair wise disjoint and their union is all of S, i.e. the A's form partition of S.

2.20.1 Solved Examples on Total Probability

Here we use the law of total probability using the concept of conditional probability and partition of sample space.

Ex. 2.20.1 : A factory uses three machines A, B, C to produce certain items. Suppose

(1) Machine A produces 40 percent of the items of which 4 percent are defective.

(2) Machine B produces 30 percent of the items of which 4 percent are defective.

(3) Machine C produces 30 percent of the items of which 5 percent are defective.

Find the probability P that a randomly selected item is defective.

Soln. :

Let D be the event that an item is defective. Then by law of total probability.

Since the events are independent, we take product.

Let $P(A)$ = Probability of items of produced by machine A

$$= \frac{40}{100} = 0.4$$

$\therefore$ 4% items produced A are defective.

$$\therefore \quad P\left(\frac{D}{A}\right) = \frac{4}{100} = 0.04$$

Similarly, $P(B) = \frac{30}{100} = 0.3$ and

$$P\left(\frac{D}{B}\right) = \frac{4}{100} = 0.04$$

and $P(C) = \frac{30}{100} = 0.3$ and

$$P\left(\frac{D}{C}\right) = \frac{5}{100} = 0.05$$

$$P(D) = P(A) \cdot P(D/A) + P(B) \cdot P(D/B) + P(C) \cdot P(D/C)$$
$$= (0.40)(0.04) + (0.30)(0.04) + (0.30)(0.05)$$
$$= \mathbf{0.0430 \text{ i.e. } 4.3\%} \quad \ldots\textbf{Ans.}$$

Ex. 2.20.2 : In an experiment a coin is tossed. If it shows head, then a die is tossed and the result is recorded. If the coin shows a tail, two dice are thrown and their sum is recorded. Find the probability that the number recorded will be 5.

Soln. : In two ways the number 5 will be recorded.

(1) A coin shows head and a single die is thrown and it shows 5.

Now, $P(H) = \frac{1}{2}$ and $P(5) = \frac{1}{6}$

$\therefore$ Probability of getting (5) $= \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$ ...(i)

(∵ Events are independent.)

(2) The coin shows tail and a pair of dice is thrown which shows (1, 4), (2, 3), (3, 2), (4, 1)

$\therefore \quad P(T) = \frac{1}{2}$ and $P(5) = \frac{4}{36}$

$\therefore$ Probability of getting $5 = \frac{1}{2} \times \frac{4}{36} = \frac{1}{18}$ ...(ii)

$\therefore$ By the theorem of total probability,

$$P(\text{getting } 5) = \frac{1}{12} + \frac{1}{18} = \frac{5}{36} \quad \ldots\textbf{Ans.}$$

Ex. 2.20.3 : In a box there are four tags numbered 1 and six tags numbered 2. There are two urns U_1 and U_2 containing 3 red and 7 black balls and 8 red and 2 black balls respectively. One tag is drawn from the box and one ball is drawn from the urn whose number is found on the tag drawn. Find the probability that a red ball is drawn.

Soln. : Let E be the event of drawing a red ball.

Let A_1 be the event of drawing a tag bearing number 1.

And A_2 of drawing a tag bearing no. 2.

$$\therefore \quad P(A_1) = \frac{4}{10} \quad \text{and} \quad P(A_2) = \frac{6}{10}$$

(∵ There are totally 10 tags)

∵ u_1 contains 3 red balls, and u_2 contains 8 red balls

$$P(E / A_1) = \frac{3}{10} \quad \text{and} \quad P(E / A_2) = \frac{8}{10}$$

$\therefore$ By the theorem of total probability

$$P(E) = P(A_1) \cdot P(E/A_1) + P(A_2) \cdot P(E/A_2)$$
$$= \frac{4}{10} \cdot \frac{3}{10} + \frac{6}{10} \cdot \frac{8}{10}$$
$$= \frac{60}{100} = \mathbf{0.6} \quad \ldots\textbf{Ans.}$$

Ex. 2.20.4 : Four roads lead away from a jail. A prisoner trying to escape from the jail selects a road at random. If road A is selected, the probability of escaping is $\frac{1}{8}$ and for road B, it is $\frac{1}{6}$, for road C, it is $\frac{1}{4}$ and for road D it is $\frac{9}{10}$. What is the probability that a prisoner will succeed in escaping from the jail ?

Soln. : Let E be the event of success in escaping.

∵ There are four roads,

Probability of selecting a road is $\frac{1}{4}$

$\therefore \quad P(A) =$ Probability of selecting road $A = \frac{1}{4}$

$P(B) = \frac{1}{4}$, $P(C) = \frac{1}{4}$, $P(D) = \frac{1}{4}$

Now, probability of escaping from road A is $\frac{1}{8}$.

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

Similarly, $P(E/B) = \frac{1}{6}$ and $P(E/C) = \frac{1}{4}$ and

$P(E/D) = \frac{9}{10}$

By Law of total probability,

$$P(E) = P(A) \cdot P(E/A) + P(B) \cdot P(E/B) + P(C) \cdot P(E/C) + P(D) \cdot P(E/D)$$
$$= \frac{1}{4} \cdot \frac{1}{8} + \frac{1}{4} \cdot \frac{1}{6} + \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{9}{10}$$
$$= \frac{1}{4}\left[\frac{1}{8} + \frac{1}{6} + \frac{1}{4} + \frac{9}{10}\right]$$
$$= \frac{1}{4} \cdot \frac{163}{120} = \frac{\mathbf{163}}{\mathbf{480}} \quad \ldots\textbf{Ans.}$$

2.21 BAYES' THEOREM

UQ. Explain Bayes 'theorem. **(SPPU - Q. 3(a), Dec. 18, 4 Marks)**

UQ. Explain : Posterior probability. **(SPPU - Q. 4(b), Dec. 18, 3 Marks)**

(i) In conditional probability, we saw that revising probability when new information is given, is an important phase of probability analysis. Given this new information, we update the **prior probability** values by calculating revised probabilities.

(ii) These are called as **posterior** probabilities Bayes' theorem provides a means for making these probability calculations.

(iii) In short, if the result is known Bayes' theorem enables us to find the probability of cause.

Unit II In Sem.

(iv) For this reason Bayes' theorem is also known as the formula for 'Probability of causes'.

We exhibit the steps in this probability revision process in the following diagram :

Prior Probabilities	→ New Information	→ Application of Bayes' Theorem	→ Posterior probabilities

Bayes' Theorem

Statement :

Let $A_1, A_2, \ldots, A_n$ be events and let them represent a partition of sample space S. Let B be any other event defined on S. If $P(A_i) \neq 0$; $i = 1, 2, \ldots, n$ and $P(B) \neq 0$, then

$$P(A_i / B) = \frac{P(A_i) \cdot P(B/A_i)}{\sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)} = \frac{P(A_i) \cdot P(B/A_i)}{P(B)}$$

Proof :

Note : $S = A_1 \cup A_2 \cup \ldots \cup A_n$ and $B \subset S$

$$\therefore B = B \cap S$$
$$= B \cap (A_1 \cup A_2 \cup \ldots \cup A_n)$$
$$= (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n)$$

(By distributive law)

$\because$ $(B \cap A_i)$ are mutually disjoint for $i = 1$ to n;

So by addition theorem of probability

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_n)$$
$$= \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(A_i) \cdot P(B/A_i) \quad \ldots(i)$$

By multiplication theorem of probability.

(Using conditional probability theorem)

Also, we have,

$$P(B \cap A_i) = P(B) \cdot P(A_i / B)$$

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

$$= \frac{P(A_i) \cdot P(B/A_i)}{\sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)}$$

Remarks

(i) The probabilities $P(A_1), P(A_2), \ldots P(A_n)$ are termed as the **a priori probabilities'** because they exist before we gain any information from the expt. itself.

(ii) The probabilities $P(B / A_i)$, $i = 1$ to n are called 'likelihoods' because they indicate how likely the event A is to occur, given each and every **'a prior'** probability.

(iii) The probabilities $P(A_i / B)$, $i = 1$ to n, are called **'posterior** probabilities' because they are determined after the results of the experiment are known.

2.21.1 Advantages of Bayesian Analysis

In the past decade, there have been enormous advances in the use of Bayesian methodology for analysis of epidemologic data, and there are now many practical advantages to the Bayesian approach.

(i) Bayesian models can easily accommodated unobserved variables such as an individual's true disease status in the presence of diagnostic error.

(ii) It provides inferences that are conditional on the data and are exact. Small sample inferences proceeds in the same manner as if one had a large sample Bayesian analysis and also it can estimate any fluctuation of parameters directly.

(iii) It obeys the likelihood principle. If two distinct sampling designs yield proportional likelihood function, then all inferences for the same likelihood function are identical from these two designs.

Classical inference does not in general obey the likelihood principle.

(iv) It provides a convenient setting for a wide range of models, such as hierarchical models and missing data problems.

2.21.2 Disadvantages of Bayesian Analysis

There are also some **disadvantages** of using Bayesian analysis.

(i) There is no correct way to choose a prior. Bayesian inferences require skills to translate subjective prior beliefs into a mathematically formulated prior. If one does not proceed with caution, one can generate misleading results.

(ii) It often comes with a high computational cost, especially in models with a large number of parameters.

(iii) It can produce posterior distributions that are heavily influenced by the priors. From a practical point of view, it might sometimes be difficult to convince a subject matter experts who do not agree with the validity of the chosen prior.

2.21.3 Applications in probabilistic inference

- The scope of applications of Baye's theorem can be widened considerably if the calculus of probability can be applied not just to events as subsets of a sample space but also to more general statements about the world.
- Scientific theories and hypothesis are much deeper statements, which have great explanatory and predictive power, and which are not so much true or false as gaining or lacking in evidence.
- One way to assess the extent to which some evidence E supports a hypothesis. H in terms of conditional probability P (H/E).
- The relative frequency interpretation of probability does not apply in this situation, so a subjective interpretation is adopted.
- The quantity P (H/E) is regarded as **degree of belief** in hypothesis H on the basis of evidence E. In order to apply the theory as objective as possible, the rules of probability are strictly applied, and an inference mechanism based on Bayes' theorem is employed.
- Suppose that there are two competing hypothesis H_1 and H_2. Let X represent all background information and evidence relative to the two hypothesis, the probabilities $P(H_1/X)$ and $P(H_1 / X \cap E)$ are called the prior and posterior probabilities of H_1, where E is a piece of evidence.
- Similarly, there are prior and posterior probabilities of H_2. Applying Bayes' theorem to both H_1 and H_2 and cancelling the denominator, we have :

$$P(H_1 / X \cap E) = \frac{P(H_1 \cap X \cap E)}{P(X \cap E)} \quad \ldots(2.21.1)$$

$$P(H_2 / X \cap E) = \frac{P(H_2 \cap X \cap E)}{P(X \cap E)} \quad \ldots(2.21.2)$$

$$\frac{P(H_1 / X \cap E)}{P(H_2 / X \cap E)} = \frac{P(H_1 \cap X \cap E)}{P(H_2 \cap X \cap E)} = \frac{P(E \cap H_1 \cap X)}{P(E \cap H_2 \cap X)}$$

$$= \frac{P(E / H_1 \cap X) \cdot P(H_1 / X)}{P(E / H_2 \cap X) \cdot P(H_2 / X)}$$

The left-hand side and the second factor on the right-hand side are called the **posterior odds** and **prior odds** respectively, favouring H_1 and H_2.

The first factor on the right-hand side is called the **likelihood ratio** and it measures how much more likely it is that the evidence event E would occur if the hypothesis H_1 were true than if H_2 were true. The new evidence E 'updates' the odds, and the process can be repeated till the likelihood ratios can be calculated.

Unit II In Sem.

2.21.4 Solved Examples on Finding the Cause When the Probability of the Event is Given

Here knowing the probability of the result, we shall be finding the probability of the cause. We shall finding conditional probability and use Baye's theorem.

Ex. 2.21.1 : There are in a bag three true coins and one false coin with head on both sides. A coin is chosen at random and tossed four times. If head occurs all the four times, what is the probability that the false coin was chosen and used.

Soln. :

Let A_1 = Event of selecting true coin.

$\therefore$ $P(A_1) = \frac{3}{4}$ ($\because$ 3 true coins)

And let A_2 = Event of selecting false coin

$\therefore$ $P(A_2) = \frac{1}{4}$

Let B be the event of getting head.

Now, $P(B/A_1)$ = Probability of getting heads with true coin

$$= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$$

($\because$ independent events)

And $P(B/A_2)$ = Probability of getting head with false coin.

$$= 1 \cdot 1 \cdot 1 \cdot 1$$

$= 1$ (∵ False coin has head on both sides)

∴ Probability of choosing false coin

$$= \frac{P(A_2) \cdot P(B/A_2)}{P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2)}$$

$$= \frac{\left(\frac{1}{4}\right) \cdot 1}{\left(\frac{3}{4}\right)\left(\frac{1}{16}\right) + \left(\frac{1}{4}\right) \cdot 1} = \frac{16}{19} \quad \text{...Ans.}$$

Ex. 2.21.2 : A bag contains five balls, the colours of which are not known. Two balls were drawn from the bag and they were found to be white. What is the probability that all balls are white ?

☑ **Soln. :** The two balls drawn are white, so there may be also 3 white, 4 white, 5 white balls.

Let, A_1 = Event that there are 2 white balls
A_2 = Event with 3 white balls
A_3 = Event with 4 white balls
A_4 = Event with 5 white balls

We assume equal probabilities for these events.

$$\therefore P(A_1) = P(A_2) = P(A_3) = P(A_4) = \frac{1}{4} = P(A_5)$$

For convenience, we regard $A_1 = A_5$

Let B be the event of white balls,

∴ $P(B/A_2)$ = Probabilities of 2 white balls drawn

$$= \frac{{}^2C_2}{{}^5C_2} = \frac{2.1}{5.4} = \frac{1}{10}$$

$P(B/A_3)$ = Probability of 3 white balls when 2 balls drawn

$$= \frac{{}^3C_2}{{}^5C_2} = \frac{3.2}{5.4} = \frac{3}{10}$$

$P(B/A_4)$ = Probability of 4 white balls when 2 balls drawn

$$= \frac{{}^4C_2}{{}^5C_2} = \frac{4 \times 3}{5 \times 4} = \frac{3}{5}$$

$$P(B/A_5) = \frac{{}^5C_2}{{}^5C_2} = 1$$

∴ By Bayes' theorem,

Probability that all are white

$$= \frac{P(A_5) \cdot P(B/A_5)}{P(A_2) \cdot P(B/A_2) + P(A_3) \cdot P(B/A_3) + P(A_4) \cdot P(B/A_4) + P(A_5) \cdot P(B/A_5)}$$

$$= \frac{\frac{1}{4} \cdot 1}{\frac{1}{4} \cdot \frac{1}{10} + \frac{1}{4} \cdot \frac{3}{10} + \frac{1}{4} \cdot \frac{3}{5} + \frac{1}{4} \cdot 1}$$

$$= \frac{1}{\frac{1}{10} + \frac{3}{10} + \frac{6}{10} + \frac{10}{10}} = \frac{10}{20} = \frac{1}{2} \quad \text{...Ans.}$$

Ex. 2.21.3 : A certain test for a particular cancer is known to be 95% accurate. A person submits to the test and the result is positive. Suppose that a person comes from a population of 1,00,000 where 2000 people suffer from that disease. What can we conclude about the probability that the person under test has that particular disease ?

☑ **Soln. :**

Let A_1 be the event that a person has cancer.

∴ $P(A_1)$ = Probability that person has cancer

$$= \frac{2000}{100000} = \frac{2}{100} = 0.02$$

Let A_2 be the event that person has no cancer

∴ $P(A_2) = 1 - 0.02 = 0.98$

Let B be the event with the particular disease

∴ $P(B/A_1) = \frac{95}{100} = 0.95$

and $P(B/A_2)$ = Probability that the test is positive when a person has no cancer

$= 1 - 0.95 = 0.05$

∴ Required probability

$$= \frac{P(A_1) \cdot P(B/A_1)}{P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2)}$$

$$= \frac{\left(\frac{2}{100}\right)\left(\frac{95}{100}\right)}{\left(\frac{2}{100}\frac{95}{100}\right) + \left(\frac{95}{100} \cdot \frac{5}{100}\right)} = \frac{190}{680} = 0.279 \quad \text{...Ans.}$$

Ex. 2.21.4 : In a communication system a zero is transmitted with probability 0.4 and a one is transmitted with probability 0.6. Due to noise in the channel a zero can be received as one with probability 0.1 and as a zero with probability 0.9. Similarly, one can be received as zero with probability 0.1 and as a one with probability 0.9.

(i) If a one is observed, what is the probability that a zero was transmitted ?

(ii) If a one is observed what is the probability that one was transmitted ?

☑ **Soln. :**

Let A_0 be an event that '0' is transmitted

∴ $P(A_0) = 0.4$

Let A_1 = a '1' be transmitted

∴ $P(A_1) = 1 - 0.4 = 0.6$

Similarly,

$P(B_0/A_0)$ = a 0 is received when 0 is transmitted
$= 0.9$

and $P(B_1/A_0)$ = a 1 is received when 0 is transmitted
$= 0.1$

Again $P(B_1/A_1)$ = a 1 is received when 1 is transmitted
$= 0.9$

and $P(B_0/A_1)$ = a 0 is received when 1 is transmitted
$= 0.1$

▶ **Step (I) :** Probability when 1 is received.

P (1 is received) = P (1 is received when 1 is transmitted) + P (1 is received when 0 is transmitted)
(∵ events are mutually exclusive)

By total probability

$$\therefore P(B_1) = P(B_1/A_1) \cdot P(A_1) + P(B_1/A_0) \cdot P(A_0)$$
$$= (0.9)(0.6) + (0.1)(0.4) = 0.58$$

▶ **Step(II) :** Probability 0 is received

$$P(B_0) = P(B_0/A_0) \cdot P(A_0) + P(B_0/A_1) \cdot P(A_1)$$
$$= (0.9) \cdot (0.4) + (0.1)(0.6) = 0.42$$

▶ **Step (III) :** P (0 was transmitted given that 1 was received)

$$= P(A_0/B_1) = \frac{P(B_1/A_0) \cdot P(A_0)}{P(B_1)}$$

$$= \frac{0.1 \times 0.4}{0.58} = 0.07$$

▶ **Step (IV) :** P(1 was transmitted given that 0 was received)

$$\therefore P(A_1/B_1) = \frac{P(B_1/A_1) \cdot P(A_1)}{P(B_1)}$$

$$= \frac{0.9 \times 0.6}{0.58} = 0.93 \quad \text{...Ans.}$$

Ex. 2.21.5 : In a factory four machines A_1, A_2, A_3 and A_4 produce 10%, 20%, 30% and 40% of the items respectively. The percentage of defective items produced by them is 5%, 4%, 3% and 2% respectively. An item is selected at random is found to be defective. What is the probability that it was produced by machine A_2 ?

☑ **Soln. :** We are given,

$$P(A_1) = \frac{10}{100} = 0.1 ; \quad P(A_2) = \frac{20}{100} = 0.2$$

$$P(A_3) = \frac{30}{100} = 0.3 ; \quad P(A_4) = \frac{40}{100} = 0.4$$

Let B be the event of defective item,

∴ $P(B/A_1) = 0.05$; $P(B/A_2) = 0.04$

$P(B/A_3) = 0.03$; $P(B/A_4) = 0.02$

Now, the probability of defective item by A_2 is

$$= \frac{P(A_2) \cdot P(B/A_2)}{P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2) + P(A_3) \cdot P(B/A_3) + P(A_4) \cdot P(B/A_4)}$$

$$= \frac{0.2 \times 0.04}{(0.1) \times (0.05) + (0.2)(0.04) + (0.3)(0.03) + (0.4)(0.02)}$$

$$= \frac{0.008}{0.03} = 0.27 \quad \text{...Ans.}$$

Ex. 2.21.6 : For a certain communication channel the probability that '0' is received as 'O' is 0.8 while the probability that '1' is received as 1 0.95. If the probability of transmitting '0' is 0.45, find

(i) a '1' is received.
(ii) '1' was transmitted given that '1' was received.
(iii) Probability that an error has occurred.

☑ **Soln. :**

Let A_0 be the event that a '0' is transmitted.

and A_1 be the event that '1' is transmitted.

∴ $P(A_0) = 0.45$, $P(A_1) = 1 - P(A_0) = 0.55$

Now, Let B be the event that the number is received,

∴ $P(B_0/A_0)$ = a '0' is received when '0' is transmitted
$= 0.8$

and $P(B_1/A_0)$ = a '1' is received when a '0' was transmitted
$= 1 - 0.8 = 0.2$

Also, $P(B_1/A_1)$ = a '1' is received when a '1' was transmitted
$= 0.95$

Unit II In Sem.

and $P(B_0/A_1)$ = a '0' is received when a '1' was transmitted

$= 1 - 0.95 = 0.05$

With this information, we calculate the required probabilities : By Bayes' theorem.

(i) P (1 is received) = P (1 is received when 1 is transmitted) + P (1 is received when 0 is transmitted)

$P(B_1) = P(B_1/A_1)\,P(A_1) + P(B_1/A_0)\cdot P(A_0)$

$= (0.95)(0.55) + (0.2)(0.45) = 0.6125$

(ii) P (1 was transmitted given that 1 was received)

$$\therefore\ P(A_1/B_1) = \frac{P(B_1/A_1)\cdot P(A_1)}{P(B_1)}$$

$$= \frac{0.95 \times 0.55}{0.6125} = 0.853$$

(iii) P (error) = P (0 was received when 1 is transmitted given that 1 was transmitted) + P (1 was received when 0 was transmitted given that 0 was transmitted).

$\therefore$ P (error) $= P(B_0/A_1)\cdot P(A_1) + P(B_1/A_0)\,P(A_0)$

$= (0.05)(0.55) + (0.2)(0.45)$

$= \mathbf{0.1175}$ **...Ans.**

Ex. 2.21.7 : A newly constructed flyover is likely to collapse. The chance that the design is faulty is 0.5. The chance that the flyover will collapse if the design is faulty is 0.95 otherwise it is 0.30. The flyover collapsed. What is the probability that it collapsed because of faulty design ?

☑ **Soln. :**

Let A_1 be the event that the design is faulty.

$\therefore\ P(A_1) = 0.5$

Let A_2 be the event that the design is not faulty.

$\therefore\ P(A_2) = 1 - P(A_1) = 0.5$

Let B be the event that the flyover will collapse.

Now, $P(B/A_1)$ = Flyover will collapse because of faulty design

$= 0.95$

and $P(B/A_2)$ = Probability that flyover will collapse otherwise

$= 0.30$

Now, by Bayes' theorem,

P (Flyover collapses because of faulty design)

$$= \frac{P(A_1)\cdot P(B/A_1)}{P(A_1)\,P(B/A_1) + P(A_2)\cdot P(B/A_2)}$$

$$= \frac{(0.5)(0.95)}{(0.5)(0.95) + (0.5)(0.30)}$$

$$= \frac{0.95}{0.95 + 0.30} = \frac{0.95}{1.25} = \mathbf{0.76}$$ **...Ans.**

Ex. 2.21.8 : A factory produces a certain type of outputs by three types of machines. The respective daily production figures are :

Machine I : 3000 units; Machine II : 2500 units,

Machine III : 4500 units

Past experience shows that 1 percent of the output produced by machine I is defective. The corresponding fraction of defectives for the other two machines are 1.2 percent and 2 percent respectively. An item is drawn at random from the day's production run and is found to be defective. What is the probability that it comes from the output of :

(i) Machine I (ii) Machine II (iii) Machine III ?

☑ **Soln. :** Let A_1, A_2, A_3 be the events that the output is produced by machines I, II and III respectively and let B denote the event that the output is defective.

Then we have,

$$P(A_1) = \frac{3000}{10000} = 0.30$$

$$P(A_2) = \frac{2500}{10000} = 0.25$$

$$P(A_3) = \frac{4500}{10000} = 0.45$$

Also we are given that

$P(B/A_1) = 1\% = 0.01$

$P(B/A_2) = 1.2\% = 0.012$

$P(B/A_3) = 2\% = 0.02$

The probability that an item selected at random from day's production is defective is given by,

$P(B) = P(A_1)\cdot P(B/A_1) + P(A_2)\cdot P(B/A_2)$

$+ P(A_3)\cdot P(B/A_3)$

$= 0.30 \times 0.01 + 0.25 \times (0.012) + 0.45\,(0.02)$

$= 0.015$

By Bayes' theorem, the required probabilities are given by,

(i) $P(A_1/B) = \dfrac{P(A_1)\cdot P(B/A_1)}{P(B)} = \dfrac{0.003}{0.015} = \dfrac{1}{5}$

(ii) $P(A_2/B) = \dfrac{P(A_2)\cdot P(B/A_2)}{P(B)} = \dfrac{0.003}{0.015} = \dfrac{1}{5}$

(iii) $P(A_3/B) = \dfrac{P(A_3)\cdot P(B/A_3)}{P(B)} = \dfrac{0.009}{0.015} = \dfrac{3}{5}$

Probabilities in (i), (ii), (iii) are known as **posterior** probabilities of events A_1, A_2, A_3 respectively.

Ex. 2.21.9 : There are two bags A and B. A contains n white and 2 black balls and B contains 2 white and n black balls. One of the two is selected at random and two balls are drawn from it without replacement. It both the balls drawn are white and the probability that the bag A was used to draw the balls is $\frac{6}{7}$, find the value of n.

☑ **Soln. :** Let A_1 denote the event that bag A is selected and A_2 denote the event that bag B is selected. Let E be the event that two balls drawn are white. We have

$$P(A_1) = P(A_2) = \frac{1}{2}$$

Now, $P(E/A_1) = \dfrac{{}^{n}C_2}{{}^{n+2}C_2} = \dfrac{n(n-1)}{(n+2)(n+1)}$

and $P(E/A_2) = \dfrac{{}^{2}C_2}{{}^{n+2}C_2} = \dfrac{1}{\frac{(n+2)(n+1)}{2}}$

$$= \frac{2}{(n+2)(n+1)}$$

Now, by Bayes' theorem, the probability that the two white balls drawn are from the bag A is given by,

$$P(A_1/E) = \frac{P(A_1)\,P(E/A_1)}{P(A_1)\,P(E/A_1) + P(A_2)\,P(E/A_2)} = \frac{6}{7}$$

$$\therefore\ \frac{\frac{1}{2}\cdot\frac{n(n-1)}{(n+2)(n+1)}}{\frac{1}{2}\cdot\frac{n(n-1)}{(n+2)(n+1)} + \frac{1}{2}\cdot\frac{2}{(n+2)(n+1)}} = \frac{6}{7}$$

$$\therefore\ \frac{n(n-1)}{n(n-1)+2} = \frac{6}{7}$$

$\therefore\ 7n(n-1) = 6n(n-1) + 12$

$\therefore\ n^2 - n - 12 = 0 \quad \therefore\ n = 4, -3$

$\because$ n cannot be negative.

$\therefore\ n = 4.$ **...Ans.**

Ex. 2.21.10 : A letter is known to have come from TATANAGAR or from CALCUTTA. On the envelope just two consecutive letters TA are visible. What is the probability that the letter came from CALCUTTA ?

☑ **Soln. :** Let A_1 and A_2 be the events that the letter came from TATANAGAR and CALCUTTA respectively. Let B denote the event that two consecutive visible letters on the envelope are TA.

$\therefore$ We have

$$P(A_1) = P(A_2) = \frac{1}{2} \text{ and}$$

$$P(B/A_1) = \frac{2}{8} \quad \text{and} \quad P(B/A_2) = \frac{1}{7}$$

Using Bayes' theorem, we get,

$$P(A_2/B) = \frac{P(A_2)\cdot P(B/A_2)}{P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2)}$$

$$= \frac{\frac{1}{2}\cdot\frac{1}{7}}{\frac{1}{2}\cdot\frac{2}{8} + \frac{1}{2}\cdot\frac{1}{7}} = \frac{4}{11}$$ **...Ans.**

Ex. 2.21.11 : A and B are two weak students of Statistics and their chances of solving a problem in Statistics correctly are $\frac{1}{6}$ and $\frac{1}{8}$ respectively. If the probability of their making a common error is $\frac{1}{525}$ and they obtain the same answer, find the probability that their answer is correct.

☑ **Soln. :** We define the events :

A_1 : Both A and B solve the problem correctly.

A_2 : Exactly one of them solves the problem correctly.

A_3 : Neither of them solves the problem correctly.

Let B : They get the same answer.

From the given data :

$P(A_1) = \frac{1}{6}\cdot\frac{1}{8} = \frac{1}{48}$: $P(B/A_1) = 1$

$P(A_2) = \frac{1}{6}\cdot\frac{7}{8} + \frac{5}{6}\cdot\frac{1}{8} = \frac{12}{48}$: $P(B/A_2) = 0$

and $P(A_3) = \frac{5}{6}\cdot\frac{7}{8} = \frac{35}{48}$

$\therefore\ P(B/A_3) = \frac{1}{525}$ (given)

By Bayes' theorem, the required probability is given by,

$$P(A_1/B) = \frac{P(A_1)\cdot P(B/A_1)}{\sum_{r=1}^{3} P(A_i)\,P(B/A_i)}$$

Unit II In Sem.

$$= \frac{\frac{1}{48} \times 1}{\frac{1}{48} \times 1 + \frac{12}{48} \times 0 + \frac{35}{48} \times \frac{1}{525}}$$

$$= \frac{1}{1 + 35 \cdot \frac{1}{525}} = \frac{15}{16} \quad \text{...Ans.}$$

Ex. 2.21.12 : A speaks truth 4 out of 5 times. A die is tossed. He reports that there is a six. What is the chance that actually there was six ?

☑ **Soln. :** We define the following events :

A_1 : A speaks truth ; A_2 : A tells a lie ; B : A reports a six

$\therefore$ $P(A_1) = \frac{4}{5}$, $P(A_2) = \frac{1}{5}$; $P(B/A_1) = \frac{1}{6}$, $P(B/A_2) = \frac{5}{6}$

By Bayes' theorem, the required probability that there was six is given by,

$$P(A_1/B) = \frac{P(A_1) \cdot P(B/A_1)}{P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2)}$$

$$= \frac{\frac{4}{5} \cdot \frac{1}{6}}{\frac{4}{5} \cdot \frac{1}{6} + \frac{1}{5} \cdot \frac{5}{6}} = \frac{4}{9} \quad \text{...Ans.}$$

Ex. 2.21.13 : In a certain college 25% of boys and 10% of girls are studying mathematics. The girls constitute 60% of the student body.

(a) What is the probability that mathematics is being studied ?

(b) If a student is selected at random and is found to be studying mathematics, find the probability that the student is a girl ?

(c) A boy ?

☑ **Soln. :**

(a) We are given,

$$P(\text{Boy}) = P(B) = \frac{40}{100} = \frac{2}{5}$$

$$P(\text{Girl}) = P(G) = \frac{60}{100} = \frac{3}{5}$$

Probability that maths is studied given that the student is a boy $= P(M/B) = \frac{25}{100} = \frac{1}{4}$

Similarly,

$$P(M/G) = \frac{10}{100} = \frac{1}{10}$$

By the theorem as total probability, probability that maths is studied.

$$= P(M) = P(G) \cdot P(M/G) + P(B)\, P(M/B)$$

$$= \frac{3}{5} \cdot \frac{1}{10} + \frac{2}{5} \cdot \frac{1}{4} = \frac{4}{25}$$

(b) By Bayes' theorem :

Probability that a maths student is girl

$$= P(G/M) = \frac{P(G) \cdot P(M/G)}{P(M)} = \frac{\frac{3}{5} \cdot \frac{1}{10}}{\frac{4}{25}} = \frac{3}{8}$$

(c) Probability that a maths student is a boy :

$$= P(B/M) = \frac{P(B) \cdot P(M/B)}{P(M)} = \frac{\frac{2}{5} \cdot \frac{1}{4}}{\frac{1}{25}} = \frac{5}{8} \quad \text{...Ans.}$$

Ex. 2.21.14 : The contents of urns I, II, III are as follows :

1 white, 2 black and 3 red bulls

2 white, 1 black and 1 red balls and

4 white, 5 black and 3 red balls.

One urn is chosen at random and two balls drawn from it. They happen to be white and red. What is the probability that they come from urns I, II and III ?

☑ **Soln. :** Let A_1, A_2, A_3 be the events that the urns I, II, III are chosen, respectively. Let B be the event that the two balls taken from the selected urn are white and red.

Now, $P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}$

Now, $P(B/A_1) = \frac{1 \times 3}{{}^6C_2} = \frac{1}{5}$

$P(B/A_2) = \frac{2 \times 1}{{}^4C_2} = \frac{1}{3}$

and $P(B/A_3) = \frac{4 \times 3}{{}^{12}C_2} = \frac{2}{11}$

Now, by Bayes' theorem,

$$P(A_2/B) = \frac{P(A_2) \cdot P(B/A_2)}{P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2) + P(A_3) \cdot P(B/A_3)}$$

$$= \frac{\frac{1}{3} \cdot \frac{1}{3}}{\frac{1}{3} \cdot \frac{1}{5} + \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{11}} = \frac{55}{118}$$

Similarly, $P(A_3/B) = \frac{\frac{1}{3} \times \frac{2}{11}}{\frac{1}{3} \cdot \frac{1}{5} + \frac{1}{5} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{11}} = \frac{30}{118}$

and $P(A_1/B) = 1 - \frac{55}{118} - \frac{30}{118} = \frac{33}{118}$ ...Ans.

2.22 D'ALEMBERT'S PARADOX

D'Alembert considered the tossing of two coins. He argued that there are three possible cases namely, (i) both heads, (ii) both tails, (iii) one head and one tail.

He concluded that the probability of getting one head one tail is $\frac{1}{3}$. As a matter of fact the actual probability of getting one head and one tail is $\frac{1}{4}$. This is known as D'Alembert's paradox which has arisen due to the difficulty of deciding equally likely alternatives.

2.23 SAMPLING DISTRIBUTIONS

- A group of pupils in a school plan to investigate how long it takes to travel between home and school. There are 2000 pupils in their school, and they realise that they do not have the time to collect and analyse such a large amount of data.
- The argue that information from some of the pupils should give them what they want provided these pupils are chosen properly. So they decide to collect data from only a part of the complete school population.

We call this part a sample-of the school population.

❒ **Definition :** A sample is any subset of a population.

An investigation of this type is said to be a **survey** of a population.

❒ **Definition :** It the above example, the information is collected by sampling; such an investigation is called is **sample survey**.

❒ **Definition :** If a survey plans to collect information from every member of a population, it is called a **census** of that population.

- The sample chosen should be reflection of the whole population, it should reproduce characteristics of the population. In our problem, the mean journey time is the characteristic for the population of school children.

2.23.1 Random sampling

There are many sampling schemes that may be called random. We shall only define a **simple** random sample, which is very straight forward. Other, more complex random sampling schemes are of particular use in certain special types of problem.

☞ **Definition of a simple random sample**

❒ **Definition :** A (simple) **random sample** is a sample which is chosen so that every member of the population is equally likely to be a member of the sample, independently of which other members of the population are chosen.

☞ **Some useful terms**

For practical reasons the investigator often has to settle for obtaining information about a population which has similar properties to the population.

It is convenient to distinguish these two populations by giving them separate names.

(i) Definition : The target population

The target population is the population about which we want information.

(ii) Definition of study population

The study population is the population about which we can obtain information.

(iii) Definition of a sample unit

A sample unit is a potential member of the sample.

2.24 TESTING OF HYPOTHESIS

UQ. Explain Hypothesis testing with example. (SPPU – Q. 3(b), Aug. 18, 4 Marks)

UQ. Explain hypothetical testing in detail with example. (SPPU – Q. 3(b), Oct. 19, 5 Marks)

- Inference based on deciding about the characteristics of the population on the basis of sample study is called the inductive inference.

Unit II In Sem.

- Such decisions involve risk of taking wrong decisions. For example, a pharmaceutical concern may be interested to find if a new drug is really effective for the particular ailment, say, in reducing blood pressure, or inducing sleep.
- It is here that the modern theory of probability pays a very vital role in decisions making and the branch of statistics which helps us in arriving at the criterion for such decisions is known as **testing of hypothesis**.
- The theory of testing of hypothesis employs statistical techniques to arrive at decisions in certain situations where there is an element of uncertainty on the basis of sample whose size is fixed in advance.

2.24.1 Statistical Hypothesis

- A statistical hypothesis is some assumption or statement, which may or may not be true, about a population, or about the probability distribution which characteristics the given population.
- We are supposed to test it on the basis of the evidence from a random sample.
- If the hypothesis completely specifies the population, then it is known as simple hypothesis, otherwise it is known as composite hypothesis.

2.24.2 Test of Hypothesis

- A test of a statistical hypothesis is a two-action decision-after observing a random sample from the given population. The two actions are the acceptance or rejection of the hypothesis under consideration.
- The truth or falsity of a statistical hypothesis is based on the information contained in the sample. It may be consistent or inconsistent with the hypothesis and accordingly the hypothesis may be accepted or rejected.
- The acceptance of a statistical hypothesis is due to insufficient evidence provided by the sample to reject it and does not necessarily imply that it is true.

2.24.3 Tests of Significance

From the knowledge of the sampling distribution of a statistic, it is possible to find the probability that a sample statistic would differ from a given hypothetical value of the parameter or from another sample value, by more than a certain amount and then to answer the question of significance, between two independent statistics. It is known as **test of significance**.

Thus we can say that :

(i) The difference between a statistic and the corresponding population, or

(ii) The difference between two independent statistics is not **significant** if it depends on fluctuations of sampling, otherwise it is said to be significant.

2.24.4 Null Hypothesis

UQ. Explain Null Hypothesis.
(SPPU – Q. 2(b), May 19, 3 Marks)

- For applying any test of significance, we set up a hypothesis-**'a definite statement about the population parameter (s)'**.
- In the words of Prof. A. R. Fisher : "Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true."

Setting up a null hypothesis

As the name suggests, it is always taken as a hypothesis of no difference.

To set the null hypothesis

(i) Express the claim or hypothesis to be tested in the symbolic form.

(ii) Identify the null hypothesis and the alternate hypothesis as :

- Take the expression involving equality sign as the null hypothesis (H_0) and the other as the alternative hypothesis (H_1).
- Thus, depending on the wording of the original claim, the original claim can be regarded as H_0 (if it contains equality sign) and sometimes it can be regarded as H_1 (if it does not contain the equality sign).

2.24.5 Alternate Hypothesis

UQ. Explain Alternative Hypothesis.
(SPPU – Q. 2(b), May 19, 3 Marks)

- Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis. It is usually denoted by H_1.
- Alternative hypothesis H_1 is stated in respect of any null hypothesis H_0, because the acceptance or rejection of H_0 is meaningful only if it is being tested against a rival hypothesis.

2.24.6 Types of errors

UQ. Explain the following :
(i) Type 1 and 2 errors
(SPPU – Q. 4(a), Aug. 18, Q. 3(b), May 19 4 Marks)

- The decision to accept or reject the null hypothesis H_0 is made on the basis of the information supplied by the observed sample observations.
- The four possible situations that arise in any test procedure are as follows :

		Decision from sample	
		Reject H_0	**Accept H_0**
True state	H_0 true	Wrong (Type I error)	Correct
	H_0 False (H_1 true)	Correct	Wrong (Type II error)

- From the above table, it is clear that we may commit two types of errors.

1. Type (I) error, definition

The error of rejecting H_0 (accepting H_1) when H_0 is true is called Type I error.

2. Type (II) error definition

- The error of accepting H_0 when H_0 is false (H_1 is true) is called type II error.
- The probabilities of Type I and Type II errors are denoted by α and β respectively.

Thus, α = Probability of Type I error
= Probability of rejecting H_0 when H_0 is true.
and β = Probability of Type II error
= Probability of accepting H_0 when H_0 is False.

Difference of Means

- Now suppose that we have not just a single sample but two samples from different populations and that we wish to compare the separate means.
- We also assume that the variances of the two populations are equal but unknown (the most common situation).

2.24.6(A) Comparison between Type I and Type II Errors

UQ. Compare Type - I and Type - II errors.
(SPPU – Q. 4(a), Oct. 19, 5 Marks)

- We make type I error by rejecting a true null hypothesis. And,
- We make Type II error by accepting a wrong null hypothesis.

If we write :

P [rejecting H_0 when it is true]
= P [type I error] = α

and P [accept H_0 when it is wrong]
= p [type II error] = β;

then α and β are also called as **sizes** of Type I and Type II errors respectively.

- In the terminology of Industrial quality control, the type I error amounts to rejecting a good lot and type II error amounts to accepting a bad lot. Hence
  α = P [rejecting a good lot] and
  β = P [accepting a bad lot]
- The sizes of type I and type II errors are also known as producer's risk and consumer's risk respectively.
- Practically it is not possible to minimise both the errors simultaneously.
- An attempt to decrease α results in an increase in β and vice-versa.
- And it is more risky to accept a wrong hypothesis than to reject a correct one; i.e. consequences of type II error are likely to be more serious than the consequences of type I error.
- So for a given sample, a compromise is made by minimising more serious errors after fixing up the less serious error.

- Thus we fix α, the size of type I error and then try to obtain a criterion which minimises β, the size of type II error.
- We have

  β = P [type II error]

  = P [accepting H_0 when H_0 is false or H_1 is true]

  Since,

  P [accept H_0 when H_0 is wrong]

  + P [accept H_0 when H_0 is true] = 1

  $\therefore$ P [accept H_0 when H_0 is true]

  = 1 – P [Accept H_0 when H_0 is wrong]

  = 1 – β

2.24.7 Power of Test

$1 - \beta$ = P [Accept H_0 when H_0 is true]

is called the power of test.

- Naturally, when H_0 is true, it ought to be accepted. Hence, minimizing β amounts to maximizing $(1 - \beta)$, which is called the **power of the test**.
- Hence, the usual practice in testing hypothesis is to fix α, the size of type I error and then try to obtain a **criterion** which minimizes β, the size of the type II error or maximises $(1 - \beta)$, the power of the test.

2.24.8 Level of Significance

- The maximum size of type I error, which we prepare to risk is known as **level of significance**. It is denoted by α and is

  P [rejecting H_0 when H_0 is true] = α
- Commonly used levels of significance in practice are 5% (0.05) and 1% (0.01).
- If we adopt 5% level of significance, it implies that we are 95% confident that our decision to reject H_0 is correct.
- Level of significance is always fixed in advance before collecting the sample information.

2.24.9 Critical Region

- Suppose we take several samples of the same size from a given population and compute some statistic t, (say x, p etc.), for each of these samples.
- Let $t_1, t_2, \ldots, t_k$ be the values of the statistic for these samples. Each of these values may be used to test some null hypothesis H_0.
- These sample statistics $t_1, t_2, \ldots, t_k$ (comprising the sample space), may be divided into two mutually disjoint groups, one leading to rejection of H_0 and other leading to acceptance of H_0.
- The statistics which lead to the rejection of H_0 give us a region called critical region (C) or Rejection Region (R), while those which lead to the acceptance of H_0 give us a region called Acceptance Region (A).
- Thus, if the statistic $t \in C$, H_0 is rejected and if $t \in A$, H_0 is accepted.
- The sizes of type (I) and type (II) errors in terms of the critical region are defined as

  α = P [Rejecting H_0 when H_0 is true]

  = P [Rejecting H_0 / H_0]

  = P [$t \in C / H_0$]

  and

  β = P [Accepting H_0 when H_0 is wrong]

  = P [accepting H_0 when H_1 is true]

  = P [Accepting H_0 / H_1]

  = P [$t \in A / H_1$]
- Where C is the critical (rejection) region, A is acceptance region and $C \cap A = \phi$, $C \cup A = S$ (sample space)

2.24.10 Examples

Ex. 2.24.1 : In order to test whether a coin is perfect, it is tossed 5 times. The null hypothesis of perfection is rejected if and only if more than 4 heads are obtained. Obtain the

(i) Critical region,

(ii) Probability of type-I error, and

(iii) Probability of type II error, when the corresponding probability of getting a head is 0.2.

☑ **Soln. :**

Let X be the number of heads obtained in 5 tosses of a coin.

H_0 : The coin is perfect i.e. unbiased,

Let H_0 : $P = \frac{1}{2}$

We use binomial distribution.

under H_0 : $X \sim B\left(n = 5, P = \frac{1}{2}\right)$

$\therefore P(X = x \mid H_0) = {}^nC_x p^x q^{n-x}$

$\because n = 5$

$\therefore P(X = x \mid H_0) = {}^5C_x \left(\frac{1}{2}\right)^5 = \frac{1}{32}{}^5C_x$; $x = 0, 1, 2, 3, 4, 5$

(i) Critical region or region of rejection

Reject H_0 if more than 4 heads are obtained

Critical region = $\{x > 4\} = \{x = 5\}$

(ii) Probability of type I error (α) is

$$\alpha = P[\text{Reject } H_0 \mid H_0] = P[X = 5 \mid H_0] = \frac{1}{32} \times {}^5C_5 = \frac{1}{32} = 0.03125$$

(iii) The probability of type II error (β) is

$$\beta = P[\text{Accept } H_0 \mid H_1] = 1 - P[\text{Reject } H_0 \mid H_1] = 1 - P[X = 5 \mid P = 0.2] = 1 - [{}^5C_x p^x q^{5-x}]\ x = 5, P = 0.2 = 1 - {}^5C_5 (0.2)^5 \cdot 1 = 0.99968$$

Ex. 2.24.2 : In order to test whether a coin is perfect, it is fussed 5 times. The null hypothesis of perfectness of the coin is accepted if an only if atmost 3 heads are obtained. Then the power of the test corresponding to the alternative hypothesis that probability of head is 0.4 is

(i) $\frac{272}{3125}$ (ii) $\frac{2853}{3125}$ (iii) $\frac{56}{3125}$ (iv) none of these.

☑ **Soln. :**

Let X be the number of heads in n = 5 tosses of a coin and let P = probability of a head in a random toss of the coin.

Null Hypothesis : $H_0 : P = \frac{1}{2}$

Alternate hypothesis : $H_1 : P = 0.4$,

Critical region : $X > 3$

Power of the test for testing H_0 against H_1 is given by :

$$1 - \beta = P[\text{reject } H_0 \text{ when } H_1 \text{ is true}] = P[\text{reject } H_0 \mid H_1] = P(X > 3 \mid P = 0.4) = \sum_{r=4}^{5} {}^5C_r (0.4)^r (0.6)^{5-r}$$

[$\because$ binomial distance X – B (n = 5, P = 0.4, under H_1)]

$$= {}^5C_4 (0.4)^4 (0.6)^1 + {}^5C_5 (0.4)^5 (0.6)^0 = 5 \cdot \left(\frac{4}{10}\right)^4 \cdot \left(\frac{6}{10}\right) + \left(\frac{4}{10}\right)^5 = \left(\frac{2}{5}\right)^4 \left[3 + \frac{2}{5}\right] = \frac{16 \times 17}{625 \times 5} = \frac{272}{3125}$$

$\therefore$ (i) is correct answer.

2.25 CHI-SQUARE TEST OF GOODNESS OF FIT

Before going into the details of the chi-square test, we study some terms used in this connection:

2.25.1 Contingency Table

Let A and B be the attributes of the given data. Let the data be classified into S classes $A_1, A_2, \ldots A_s$ according to attribute A and into t classes $B_1, B_2, \ldots B_t$ according to the attribute B. Let O_{ij} be the observed frequency of the cell belonging to the classes A_i (i = 1, 2, ..., s) and B_j (j = 1, 2, ..., t). The data can be set into a s × t contingency table of s rows and t columns as follows :

Classes	B_1	B_2	...	B_j	...	B_t	Total
A_i	O_{1i}	O_{12}	...	O_{1j}	...	O_{1t}	A_1
A_2	O_{21}	O_{22}	...	O_{2j}	...	O_{2t}	A_2
:		:					
A_i							
A_s	O_{s1}	O_{s2}	...	O_{5j}	...	O_{st}	A_s
Total	B_1	B_2	...	B_j	...	B_t	N

2.25.2 Degrees of Freedom

The term degrees of Freedom refers to the number of "independent constraints" in a set of data. We explain this concept with a few examples :

(1) If the data are given in a contingency table, then the degree of freedom is calculated by the formula $\gamma = (c - 1)(r - 1)$.

Where γ stands for degree of freedom, c for number of columns and r for the number of rows. Thus in a 2 × 2

Unit II In Sem.

table, degrees of freedom are $(2-1)\cdot(2-1)=1$ and so on.

(2) If the data are not given in the form of contingency table but given in the shape of a series of individual observations then the degrees of freedom are calculated in a different way.

Consider the following distributions :

Number of heads	Actual frequencies
0	2
1	10
2	38
3	106
4	188
5	257
6	226
7	128
8	59
9	7
10	3
Total	**1024**

Here, if we write down the expected frequencies we have freedom to write any ten figures we choose but the eleventh figure must be equal to 1024 minus the total of the ten figures we have written, because the total of the expected frequencies must be equal to the total of the actual frequencies. Thus, there are ten degrees of freedom in the above question. In such cases the degrees of freedom are equal to $(n-1)$ ***where n is number of frequencies.***

2.26 CHI-SQUARE TEST

The square of a standard normal variable is called a chi-square (pronounced as Sky, without S) variate with 1 degree of freedom (d.f.)

Thus if X is a random variable following normal distribution with mean μ and standard deviation σ, then $\left(\frac{X-\mu}{\sigma}\right)$ is a standard normal variate.

$\therefore \left(\frac{X-\mu}{\sigma}\right)^2$ is a chi-square (χ^2) variate with 1 d.f.

If $X_1, X_2, \ldots X_r$ are r independent random variables following normal distribution with means $\mu_1, \mu_2, \ldots, \mu_r$ and standard deviations $\sigma_1, \sigma_2, \ldots \sigma_r$ respectively then the variate

$$\chi^2 = \left(\frac{X_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{X_2-\mu_2}{\sigma_2}\right)^2 + \ldots + \left(\frac{X_r-\mu_r}{\sigma_r}\right)^2$$

$$= \sum_{i=1}^{r}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2,$$

Which is the sum of the squares of r independent standard normal variates, follows Chi-square distribution with r d.f.

2.26.1 Probability Density Function (p.d.f.) of Chi-square Distribution

If χ^2 is a random variable following Chi-square distribution with γ d.f. then its probability function is given by,

$$P(\chi^2) = \frac{1}{2^{\gamma/2}\lceil\gamma/2}\, e^{-\chi^2/2}\,(\chi^2)^{(\gamma/2-1)}, \quad 0 < \chi^2 < \infty$$

Where $\lceil\gamma/2$ is Gamma function

2.26.2 Remark

(1) The probability function $P(\chi^2)$ depends on degrees of freedom. As γ changes, $P(\chi^2)$ changes.

(2) Constants of χ^2 distribution with γ d.f.

Mean $= \gamma$; Mode $= \gamma - 2$

Variance $= 2\gamma$

$\mu_1 = \gamma_1$, $\mu_2 = 2\gamma$, $\mu_3 = 8\gamma$, $\mu_4 = 48\gamma + 12\gamma^2$.

(3) Pearson;s Co-efficient of skewness :

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{s.d.}} = \frac{\gamma-(\gamma-2)}{\sqrt{2\gamma}} = \sqrt{\frac{2}{\gamma}}$$

(i) Since coeff. of skewness > 0 for $\gamma \geq 1$; χ^2 - distribution is positively skewed.

(ii) Since skewness is inversely proportional to the square-root of d.f., it tends to symmetry as d.f. increases.

Thus for large d.f, χ^2-distribution tends to normal distribution.

(4) For large γ, the standard variate

$$Z = \frac{\chi^2 - E(\chi^2)}{\text{s.d.}(\chi^2)} = \frac{\chi^2-\gamma}{\sqrt{2\gamma}}$$

is a standard normal variate.

(5) **Additive Property :**

If $\chi_1^2, \chi_2^2, \ldots \chi_k^2$ are independent χ^2- variates with $n_1, n_2, \ldots, n_k$ d.f.s., then the sum

$\sum_{i=1}^{k} \chi_i^2 = \chi_1^2 + \chi_2^2 + \ldots + \chi_k^2$ is a χ^2 - variate with $(n_1 + n_2 + \ldots + n_k)$ d.f.

2.26.3 Applications of χ^2 - Distribution

Some of the applications of χ^2-distribution are :

(i) Chi- square test of goodness of fit

(ii) χ^2-test for independence of attributes

2.26.4 Chi-Square Test of Goodness of Fit

χ^2-test of goodness of fit is used to test if the deviation between observation (experiment) and theory may be attributed to chance or if it is really due to the inadequacy of the theory to fit the observed data.

karl Pearson proved that the statistic

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i-E_i)^2}{E_i}\right]$$

$$= \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} + \ldots + \frac{(O_n-E_n)^2}{E_n}$$

Follows χ^2- distribution with $\gamma = n-1$ d.f., where $O_1, O_2, \ldots, O_n$ are observed frequencies and $E_1, E_2, \ldots E_n$ are corresponding expected frequencies under some theory of hypothesis.

2.26.5 Steps to Compute χ^2 and Drawing Conclusion

(i) Compute the expected frequencies $E_1, E_2, \ldots, E_n$ corresponding to observed frequencies $O_1, O_2, \ldots, O_n$ under same theory.

(ii) Compute the deviation $(O_i - E_i)$ and square them to obtain $(O_i - E_i)^2$.

(iii) Divide $(O_i - E_i)^2$ by E_i and add the values to compute $\chi^2 = \sum\left[\frac{(O_i-E_i)^2}{E_i}\right]$

(iv) Under the hypothesis H_0 : the theory fits the data well, the above statistic follows χ^2-distribution with $\gamma = (n-1)$ d.f.

(v) Look up tabulated (critical) values of χ^2 at 5% or 1% level of significance and draw the conclusion.

2.26.6 Conditions for the Validity of Chi-Square Test

The χ^2-test can be used only if the following conditions are satisfied :

(i) The total frequency, N, should be large, say greater than 50.

(ii) The sample observations should be independent; i.e., no individual item should be included twice or more in the sample.

(iii) The constraints should not involve square or higher powers of the frequencies.

(iv) No theoretical frequency should be small; i.e. it should be larger than 10 but not less than 5.

(v) The data should be in original units.

2.26.7 Examples

Ex. 2.26.1 : The number of scooter accidents per month in a certain town were as follows : 12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that accidently conditions were the same during this 10 month period?

Soln. :

Null Hypothesis :

H_0 : The given frequencies (i.e. number of accidents per month) are consistent with the belief that the accident conditions were same during the 10-month period.

Now, total number of accidents =

$12 + 8 + 20 + 2 + 14 + 10 + 15 + 6 + 9 + 4 = 100$,

under the null hypothesis,

Expected number of accidents for each of the 10 months. $= \frac{100}{10} = 10$. ($\because$ these accidents are uniformly distributed)

Now, d.f. $= 10 - 1 = 9$

$\therefore$ Tabulated $\chi^2_{0.05}$ for 9 d.f. $= 16.919$ ...(i)

We prepare table for computation of χ^2.

Month	Observed no. of accidents (O)	Expected no. of accidents (E)	(O – E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
1.	12	10	2	4	0.4
2.	8	10	–2	4	0.4
3.	20	10	10	100	10.0
4.	2	10	–8	64	6.4
5.	14	10	4	16	1.6
6.	10	10	0	0	0
7.	15	10	5	25	2.5
8.	6	10	–4	16	1.6
9.	9	10	–1	1	0.1
10.	4	10	–6	36	3.6
Total	100	100	0	–	26.6

$$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 26.6 \quad \text{...(ii)}$$

Since calculated value of $\chi^2 = 26.6$ is greater than tabulated value from (i) = 16.919, it is significant and hence the null hypothesis is rejected at 5% level of significance.

Hence, we conclude that the accident conditions are certainly not uniform over the 10-month period.

Ex. 2.26.2 : The theory predicts the proportion of beans, in the four groups A, B, C and D should be 9 : 3 : 3 : 1. In an experiment among 1,600 beans, the numbers in the four groups were 882, 313, 287 and 118. does the experimental result support the theory ? (The table value of χ^2 for 3 d.f. at 5% level of significance is 7.81).

☑ Soln. :

Null Hypothesis : H_0 : There is no significant difference between the experimental values and the theory; i.e. the theory supports the experiment.

The proportion of beans in four groups A, B, C and D should be 9 : 3 : 3 : 1. Hence the theoretical (expected) frequencies are as shown.

Category	Expected frequency (E)
A	$\frac{9}{16} \times 1600 = 900$
B	$\frac{3}{16} \times 1600 = 300$
C	$\frac{3}{16} \times 1600 = 300$
D	$\frac{1}{16} \times 1600 = 100$

Computation of χ^2

Category	Observed frequency (O)	Expected frequency (E)	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
A	882	900	–18	324	0.360
B	313	300	+13	169	0.563
C	287	300	–13	169	0.563
D	118	100	18	324	3.240
Total	1600	1600	0	986	4.726

$$\therefore \quad \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 4.726$$

Now, d.f. = 4 – 1 = 3 and tabulated χ^2 for 3 d.f.

$$\chi^2_{0.05} = 7.81$$

Conclusion : Since calculated value of χ^2 is less than tabulated value, it is not significant. Hence we accept the null Hypothesis at 5% level of significance. Thus, the experimental results support the theory.

Ex. 2.26.3 : A die is rolled 100 times with the following distribution.

Number	1	2	3	4	5	6
Observed frequency	17	14	20	17	17	15

At 0.01 level of significance, determine whether the die is true (or uniform).

☑ Soln. :

We have number of categories = 6

N = Total Frequency = 17 + 14 + 20 + 17 + 17 + 15 = 100

Null Hypothesis : H_0 : The die is true (uniform)

Under H_0, the probability of obtaining each of the six faces 1,2,...,6 is same, i.e. $P = \frac{1}{6}$

$\therefore$ Expected frequency for each face = N. P

$$= 100 \cdot \frac{1}{6} = 16.67$$

Computation of χ^2

Number	Observed frequency (O)	Expected frequency (E)	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
1.	17	16.67	0.33	0.1089	0.0065

Number	Observed frequency (O)	Expected frequency (E)	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
2.	14	16.67	– 2.67	7.1289	0.4276
3.	20	16.67	3.33	11.0889	0.6652
4.	17	16.67	0.33	0.1089	0.0065
5.	17	16.67	0.33	0.1089	0.0065
6.	15	16.67	– 1.67	2.7889	0.1673
total	-	-	0	-	1.2796

$$\therefore \quad \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 1.2796 \qquad ...(i)$$

The degrees of freedom (d.f.) = 6 – 1 = 5

The critical or tabulated value of Chi-square for $\gamma = 5$ and at 1% level of significance is : $\chi^2_5(0.01) = 15.086$ ...(ii)

Since calculated valued of χ^2 is less than critical value, it is not significant.

Hence H_0 may be accepted at 1% level of significance; i.e. the die may be regarded as true or uniform.

Ex. 2.26.4 : Records taken of the number of male and female births in 800 families having four children are given in the Table P. 2.26.4

Table P. 2.26.4

No. of births		Frequency
Male	Female	
0	4	32
1	3	178
2	2	290
3	1	236
4	0	64

Test whether the data are consistent with the hypothesis that the binomial law holds and the chance of a male birth is equal to that of a female birth.

☑ Soln. :

Null Hypothesis : H_0 : Data are consistent with the binomial law of equal probability for male and female births, so that $p = q = \frac{1}{2}$; where p is the probability of a male birth. We have n = 4, N = 800

According to binomial probability law, the frequency of 'r' male births is given by

$$f(r) = N\,p(r) = N \times {}^nC_r p^r q^{n-r}$$
$$= 800 \times 4C_r\, p^4 q^{4-r}$$
$$= 800 \times 4C_r \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{4-4}$$
$$= 800 \times 4C_r \left(\frac{1}{2}\right)^4 \quad r = 0, 1, 2, 3, 4.$$
$$= 50 \times 4C_r \qquad ...(i)$$

Substituting r = 0,1,2,3,4 successively in (i) we get the theoretical frequencies.

Table P. 2.26.4(a)

	No. of male births (γ)	Expected Frequency F(γ)
Now, we prepare a table for **testing Goodness of fit**	0	$50 \times 4C_0 = 50 \times 1 = 50$
	1	$50 \times 4C_1 = 50 \times 4 = 200$
	2	$50 \times 4C_2 = 50 \times 6 = 300$
	3	$50 \times 4C_3 = 50 \times 4 = 200$
	4	$50 \times 4C_4 = 50 \times 1 = 50$
	Total	800

No. of male births	Observed frequency (O)	Expected frequency (E)	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	32	50	– 18	324	6.48
1	178	200	– 22	484	2.42
2	290	300	– 10	100	0.33
3	236	200	36	1296	6.48
4	64	50	14	196	3.92
Total	800	800	0	-	19.63

$$\therefore \quad \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 19.63$$

Here, we have 5 frequencies.

$\therefore$ d.f. is = 5 – 1 = 4

Now, tabulated value (critical value) of χ^2 for 4 d.f. at 5% level of significance is 9.488.

Since calculate value of $\chi^2 = 19.63$ is greater than the tabulated value, it is significant.

Hence we reject the null hypothesis at 5% level of significance.

Unit II In Sem.

Conclusion : The hypothesis of equal male and female births is wrong.

Hence the binomial distribution with $P = q = \frac{1}{2}$, is not a good fit to the given data.

2.26.8 Levels of Significance

We have seen that the values of χ^2 are compared with table values. Usually we find the value of χ^2 and 5 % level of significance for the given degrees of freedom from the table. If the calculated value of χ^2 is greater than the tabled value it means the difference is significant, and if it is less than the tabled value, the difference can be ignored.

2.26.9 Method of solving the problem

(i) Assume the correctness of hypothesis or make a hypothesis of independence of arithmetic and on this assumption find the expected frequencies.

(ii) Calculate the value of χ^2 using the formula,

$$\sum\left[\frac{(f - f_1)^2}{f_1}\right]$$

(iii) Find the number of 'Degrees of freedom'.

(iv) Find the table value of χ^2 at 0.05 level of significance with the above degrees of freedom.

(v) Compare the tabled value and calculated value of χ^2 and derive the conclusion.

Examples

Ex. 2.26.5 : The following table shows the number of people interviewed by age-groups and the number in each age group estimated to have peptic ulcers.

Age group	15-20	20-25	25-35	35-45	45-55	55-65	65-75	Total
Nos. interviewed	199	300	1128	1375	1089	625	155	4871
P. U. cases	1	8	38	96	105	56	12	316

Do these figures justify the hypothesis that peptic ulcer is equally popular in all age groups ?

Soln. :

Our hypothesis is : Peptic ulcer (p.u.) is equally popular in all age groups.

Total no. Interviewed = 4871

p.u. cases = 316

$\therefore$ In each age group $\frac{316}{4871} \times 100 = 6.5$; i.e. 6.5 % of the people should suffer from p.u. on this basis we evaluate expected frequencies as follows :

Age group	15-20	20-25	25-35	35-45	45-55	55-65	65-75
Observed cases	1	8	38	96	105	56	12
Expected cases	13	19.5	73	89	71	40.5	10

Now, we evaluate χ^2. By definition,

$$\chi^2 = \sum\left[\frac{(f - f_1)^2}{f_1}\right]$$

$$\therefore \chi^2 = \frac{(1-13)^2}{13} + \frac{(8-19.5)^2}{19.5} + \frac{(38-73)^2}{73} + \frac{(96-89)^2}{89} + \frac{(105-71)^2}{71} + \frac{(56-40.5)^2}{40.5} + \frac{(12-10)^2}{10}$$

$$= 57.6 \quad \ldots(1)$$

Now, here the data are given in the form of series of individual observations, hence γ = degrees of freedom = 7 – 1 = 6. Now, for six degrees of freedom the table value of χ^2 at 5 % level of significance is 12.59 ...(2)

Here, the calculated value is much higher than the table value and hence the hypothesis is not justified.

Ex. 2.26.6 : Two hundred digits were chosen at random from a set of tables. The frequencies of the digits were as follows :

Digit	Frequency
0	18
1	19
2	23
3	21
4	16
5	25
6	22
7	20
8	21
9	15

Use χ^2 test to assess the correctness of hypothesis that the digits were distributed in equal number in the tables from which they were chosen.

Soln. : We take the hypothesis that the digits are distributed in equal numbers in the table as correct. With this assumption, the expected frequencies of the digits will be,

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	20	20	20	20	20	20	20	20	20	20

We substitute the observed and the expected frequencies in the formula,

$$\chi^2 = \sum\left[\frac{(f - f_1)^2}{f_1}\right], \text{ and we get,}$$

$$\chi^2 = \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(23-20)^2}{20} + \frac{(21-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(25-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(21-20)^2}{20} + \frac{(15-20)^2}{20}$$

$$= \frac{1}{20}[4+1+9+1+16+25+4+0+1+25]$$

$$= 4.3 \quad ...(1)$$

The degrees of freedom = 10 – 1 = 9 [or we can use the formula, degrees of freedom]

$$= (c-1)(r-1) = (10-1)(2-1) = 9$$

Now, the values of χ^2 at 5% level significance with 9 degrees of freedom is 16.919 ...(2)

Form Equation (1) and Equation (2), we see that the calculated value of χ^2 is much less than the figure. Hence the hypothesis seems reasonable.

2.26.10 Student's 't' distribution

Definition : Let $x_1, x_2, ...x_n$ be a random sample of size n from a **normal** population with mean μ and variance σ^2, (S. P. σ), then student's 't' statistic is defined as

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \quad ...(2.26.1)$$

Where $\bar{x} = \frac{\sum x}{n}$ and $S^2 = \frac{1}{(n-1)}\sum(x - \bar{x})^2$, is an unbiased estimate of the population variance σ^2.

't' defined in (i) follows student's t-distribution with '$\gamma = (n-1)$' degrees of freedom.

Probability density function (p. d. f.) of student's t distribution :

with $\gamma = (n-1)$ degrees of freedom is given by

$$p(t) = \frac{1}{\sqrt{\gamma}\,\beta\left(\frac{1}{2}, \frac{\gamma}{2}\right)}\left(1 + \frac{t^2}{\gamma}\right)^{-\frac{(\gamma+1)}{2}} ; -\infty < t < \infty$$

Remark

We write $t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

In general statistic 't' following student's t-distribution with γ d.f. is written as $t \sim t_\gamma$.

2.26.11 Properties of t-distribution

Some of the important properties of t-distribution are :

(1) Moments : Since p. d. f. R(t) involves only even powers of t, hence all bold order moments about origin are zero; i.e.

$$\mu'_{2r+1} = 0 \quad ...(2.26.2)$$

In particular, if r = 0, then

$$\mu'_1 = 0 \quad \text{i.e. mean} = \mu'_1 = 0 \quad ...(2.26.3)$$

$\therefore \mu'_{2r+1}$ (about origin) = 0

i.e., μ'_{2r+1} (about mean) = 0

i.e. central moments $\mu_{2r+1} = 0, \gamma = 0, 1, 2,$

Thus all order moments about mean are zero,

It can be shown that

$$\mu_{2r} = \frac{\gamma(2r-1)}{(\gamma - 2r)}\mu_{2r-2} \quad ...(2.26.4)$$

Is the recurrence relation for the moments of t-distribution

In particular, μ_2 = variance $= \frac{\gamma(2-1)}{(\gamma-2)}$,

$$= \frac{\gamma}{\gamma - 2}, (\gamma > 2) \quad [\because \mu_2 \text{ is always +ve}, \therefore \gamma > 2]$$

And $$\mu_4 = \frac{\gamma(4-1)}{(\gamma-4)}\mu_2 = \frac{3\gamma}{(\gamma-4)}\frac{\gamma}{(\gamma-2)}$$

$$\therefore \mu_4 = \frac{3\gamma^2}{(\gamma-2)(\gamma-4)} \quad ...(2.26.5)$$

$(\gamma > 4, \because \mu_4$ is always positive)

Unit II In Sem.

Now, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$ and ...(2.26.6)

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\gamma^2}{(\gamma-2)(\gamma-4)} \cdot \frac{(\gamma-2)^2}{\gamma^2}$$

$$\therefore \beta_2 = \frac{3(\gamma-2)}{(\gamma-4)} \quad ...(2.26.7)$$

Now, as $\gamma \to \infty$, $\beta_1 = 0$ and

$$\beta_2 = \lim_{\gamma \to \infty} \frac{3(\gamma-2)}{(\gamma-4)}$$

$$= 3 \lim_{\gamma \to \infty} \frac{\left(1-\frac{2}{\gamma}\right)}{\left(1-\frac{4}{\gamma}\right)}$$

$$= 3 \frac{(1-0)}{(1-0)} = 3$$

$$\therefore \beta_2 = 3 \quad ...(2.26.8)$$

Hence for large number of degrees of freedom, t-distribution approaches the Normal distribution. Hence, for $\gamma > 30$, we can directly apply normal test.

(2) Curve (as shown in the Fig. 2.26.1) is unimodas with Mean = Median = Mode = 0.

(3) As $|t|$ increase, $p|t|$ decreases and tends to zero as $|t| \to \infty$. Hence the curve is asymptotic to the t-axis, at each end.

(4) A very interesting property of the sampling distribution of t is that it does not depend on the population parameters and depends only on $\gamma = (n-1)$, i.e., on the sample size.

Comparison between normal curve and t-curve

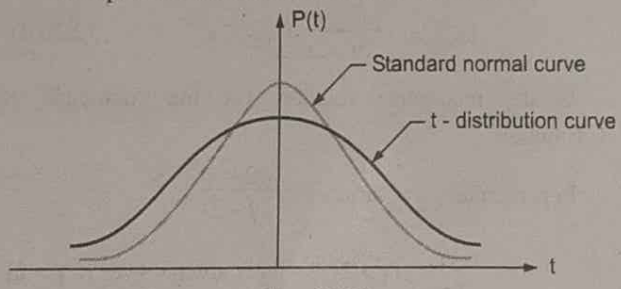


Fig. 2.26.1

(5) The probability curve is symmetric about t = 0.

(6) It does not depend on the population parameters and depends only on $\nu = n-1$; i.e. on the sample size.

(7) The shape of t - distribution is dependent on the sample size n.

(8) As sample size n increases, the distribution becomes approximately normal.

(9) The standard deviation is greater than 1.

(10) We use 'S' as an estimate of σ. So variability is introduced in t-distribution.

Hence the area in the tails of t-distribution is a little greater than the area in the tails of the standard normal distribution.

(11) As sample size n increases, the values of S get closer to σ.

2.26.12 Applications of t-distribution

(i) The t-distribution has a number of applications in statistics, among them we shall discuss the following :

(ii) t-test for the significance of single mean, population variance being unknown.

Assumptions for student's t-test

The sample variance (S^2) is given by

$$S^2 = \frac{1}{n}\sum (x-\bar{x})^2$$

and $$S^2 = \frac{1}{(n-1)}\sum (x-\bar{x})^2$$

$$\therefore nS^2 = (n-1)S^2$$

$$\therefore \frac{S^2}{n} = \frac{s^2}{(n-1)}$$

In numerical problem, in general we are given the sample standard deviation. Hence for numerical problems, the test-statistic 't' is given by

$$t = \frac{\bar{x}-\mu_0}{\sqrt{S^2/n}} = \frac{\bar{x}-\mu_0}{\sqrt{s^2/(n-1)}} \sim t_{n-1}$$

Ex. 2.26.7 : A machine is designed to produce insulating washers for electrical devices of average thickness 0.025 cm. A random sample of 10 washers was found to have an average thickness of 0.024 cm. with a standard deviation of 0.002 cm. Test the significance of deviation. Value of t for 9 degrees of freedom at 5% level is 2.262.

Soln. :

We have n = 10, $\bar{x}$ = 0.024 cm.

s = 0.002 cm (sample standard deviation)

Null Hypothesis

$H_0 : \mu = 0.025$ cm; i.e. there is no significant deviation between sample mean $\bar{x} = 0.024$ and population mean $\mu = 0.025$.

Alternative Hypothesis

$H_1 : \mu \neq 0.025$ cm

Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{10-1} = t_9$$

$$\therefore t = \frac{0.024 - 0.025}{0.002/\sqrt{9}} = \frac{-0.001 \times 3}{0.002}$$

$$= -1.5$$

Now, tabulated valve for 9 d. f. is = 2.262 since $|t| < 2.262$, it is not significant, at 5% level of significance.

Hence the deviation $(\bar{x} - \mu)$ is not significant.

Ex. 2.26.8 : Godrej soaps manufacturing company was distributing a particular brand of soap through a large number of retail shops. Before a heavy advertisement campaign, the mean sales per week per shop was 140 dozens. After the campaign, a sample of 26 shops was taken and the mean sales was found to be 147 dozens with standard deviation 16. Can you consider the advertisement effective?

Soln. :

We have $n = 26$, $\bar{x} = 147$,

$s = 16$ dozens

Null Hypothesis : $H_0 : \mu = 140$ dozens; i.e. the deviation between $\bar{x}$ and μ is just due to fluctuations of sampling. In other words, advertisement is not effective.

Alternate Hypothesis, $H_1 : \mu > 140$ and

Under H_0, the test statistic is

$$t = \frac{147 - 140}{16/\sqrt{25}} = \frac{7 \times 5}{16} = 2.19$$

$$\left[\because t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{n-1} = t_{25}\right]$$

Tabulated value of t for 25 d. f. At 5% level of significance for single tail test is 1.798, i.e. $t_{25}(0.05) = 1.708$.

Since calculated value of t is greater than the tabulated value, it is significant and we reject H_0 at 5% level of significance.

Hence, the increase in sales cannot be attributed to fluctuations of sampling and we conclude that advertisement is certainly effective in increasing the sales.

2.27 HYPOTHESIS

Unit II In Sem.

2.27.1 Simple Hypothesis

It is a statistical hypothesis which completely specifies an exact parameter. Null hypothesis is always a simple hypothesis stated as an equality specifying an exact value of the parameter.

Examples of Null Hypothesis (N.H.)

(i) N.H. $= H_0 : \mu = \mu_0$;

i.e. population mean equals to a specified constant μ_0.

(ii) N.H. $= H_0 : \mu_1 - \mu_2 = \delta$

i.e. the difference between the sample means equals to a constant δ.

2.27.2 Composite Hypothesis

It is stated in terms of several possible values i.e., by an inequality.

Examples of Alternate Hypothesis

(i) A.H. : $H_1 : \mu > \mu_0$

(ii) A.H. : $H_1 : \mu < \mu_0$

(iii) A.H. : $H_1 : \mu \neq \mu_0$

2.27.3 One Tailed Test (O.T.T) and Two Tailed Test (T.T.T)

UQ. Explain (i) Left-tail test (ii) Right-tail test **(SPPU – Dec. 18, 4 Marks)**

(I) Right One Tailed Test (R.O.T.T)

(a) When the alternative hypothesis (A.H.) : H_1 is of the greater than type i.e. $H_1 : \mu > \mu_0$ or $H_1 : \sigma_1^2 > \sigma_2^2$ etc then the entire critical region of area α lies on the right side tail of the probability density curve as shown in the shaded Fig. 2.27.1.

(b) In such case, the test of hypothesis (T.O.H.) is known as Right One-Tailed Test (R.O.T.T)

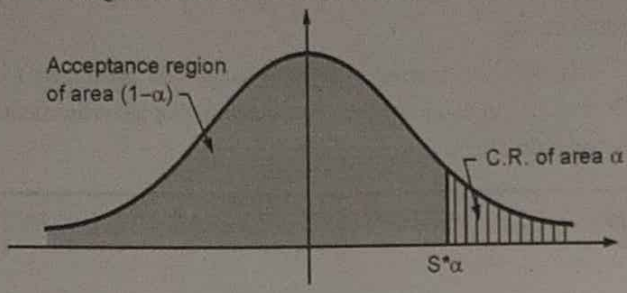


Fig. 2.27.1 : One sided right-tail test

(II) Left One Tailed Test (L.O.T.T)

(a) When the alternative hypothesis (A.H.) : is of the less than type i.e. $H_1 : \mu_1 < \mu_0$ or $H_1 : \sigma_1^2 < \sigma_2^2$ etc. then the entire critical region of area α lies on the left side tail of the curve shown in the Fig. 2.27.2.

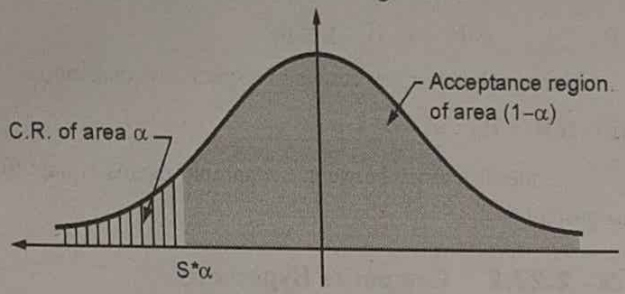


Fig. 2.27.2 : Left one tailed test

☞ Remark

(a) Generally a one-tailed test of hypothesis is used when one talks of type I error.

(b) A hypothesis test is also called as **one-sided test** and is designed to identify a difference from a hypothesized value in only one direction.

(c) It is also called **directional test,** because it includes the directional prediction in the statement of hypothesis and the location of the critical region.

(d) The critical region for a one-sided test is the set of values less than the critical value of the test or the set of values greater than the critical value of the test.

(e) A one-tailed test is one where H_1 is directional and includes < or >.

(f) A one-tailed test looks for an increase or decrease in the parameter.

(g) If we reject the **null hypothesis** at the 5% level of significance, we say "there is significant evidence to reject the hypothesis at 5% level".

(h) We can perform the test at any level, usually 1%, or 5%. For example, performing the test at a 1% level of significance means that there is a 1% chance of wrongly rejecting H_0.

(III) Two Tailed Test (T.T.T)

(a) If alternative hypothesis (A.H.) is of the not equals type i.e. $H_1 : \mu_1 \neq \mu_2$ or $H_1 : \sigma_1 \neq \sigma_2$ etc., then C.R. lies on both sides of the right and left tails of the curve such that C.R. of area $\frac{\alpha}{2}$ lies on the right tail and C.R. of area $\frac{\alpha}{2}$ lies on the left tail as shown in the Fig. 2.27.3

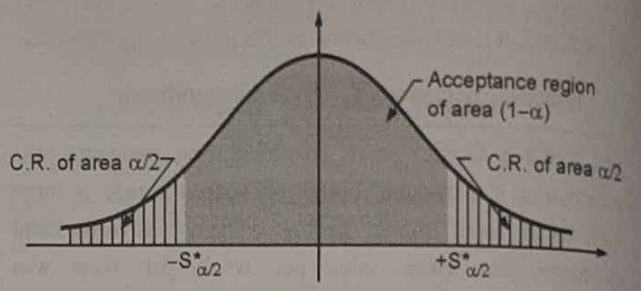


Fig. 2.27.3 : Two tailed test

(b) A two-tailed test (T.T.T.) is one where H_1 has no direction.

(c) It is a statistical hypothesis test in which the values for which we can reject the hypothesis, H_0 are located in both tails of the probability distribution.

(d) The critical region for a two-tailed test is the set of values less than a first critical value of the test and the set of values greater than a second critical value of the test.

2.27.4 Steps for Test of Hypothesis

The test of hypothesis or test of significance or rule of decision consists of the following steps :

(1) Formulate Null hypothesis N.H. : H_0.

(2) Formulate alternative hypothesis A.H. : H_1

(3) Choose level of significance L.O.S. : α

(4) **Critical region (C.R.) :** is determined by the critical value S_α^* and the kind of A.H. (based on which the test

is Right One-Tailed Test (R.O.T.T) or Left One- Tailed Test (L.O.T.T.).

(5) Compute the test statistic S* using the **sample data.**

(6) **Decision :** Accept or reject N.H. depending on the relation between S* and S^*_α.

2.27.5 Population and Sample

UQ. What is the difference between sample and population. **(SPPU – Dec. 18, 3 Marks)**

(I) Population

(i) **Population** is the set or collection or totality of objects, animate or inanimate, actual or hypothetical, under study.

(ii) Mainly population consists of sets of numbers, measurements or observations which are of interest.

(iii) Size of the population N is the number of objects or observations in the population.

(iv) Population is finite on infinite depending upon the size N being finite or infinite.

(v) When measures like the mean, mode, variance and standard deviation of a population distribution are computed, they are referred to as parameters.

Example

(i) Population of India,

(ii) Engineering colleges recognised by AICTE.

(II) Sample

(i) Since it is impracticable or uneconomical or time consuming to study the entire population, a finite subset of the population known as **sample** is studied.

(ii) Size of the sample is denoted by n.

(iii) Sampling is the process of drawing samples from a given population.

(iv) A sample is a smaller collection of units from a population to study the truths about the population.

(v) The sample should be representative of the general population. This is achieved by random sampling.

Examples

(i) Maruti cars in India,

(ii) Expenditure in a district hospital (sub-sample) etc.

2.27.6 p-Value

(i) The probability of observing a value as extreme as or more extreme than the observed value.

(ii) p-value is the lowest level of significance at which the observed value of the test statistic is significant.

(iii) In the significance testing by p-value approach, α is not pre-determined but the conclusion is based on the size of the p-value and that is computed using the value of test statistic.

(iv) It is the probability of wrongly rejecting the null hypothesis, if it is in fact true.

(v) Small p-values suggest that the null-hypothesis is unlikely to be true. It indicates the strength of evidence for rejecting the null hypothesis, H_0.

(vi) p-value is a probability associated with a critical value. The critical value depends on the probability you are allowing for a type I error.

(vii) The p-value helps to decide whether or not to accept the null-hypthesis. This depends upon how low the p-value should be before null-hypthesis is rejected.

(viii) instead of computing a rejection region, we calculate p-value.

(ix) The p-value is the smallest level of significance at which H_0 can be rejected.

(x) Once the p-value has been determined, the conclusion at any particular level of α results from comparing the p-value to α :

(a) p-value $\leq \alpha \Rightarrow$ reject H_0 at level α

(b) p-value $> \alpha \Rightarrow H_0$ not to be rejected at level α.

2.27.7 Test of Hypothesis Concerning Single Population Mean μ : (with known Variance σ^2 : Large Sample)

Let μ and σ^2 be the mean and variance of a population from which a random sample of size n is drawn. Let $\bar{x}$ be the mean of the sample. Then for large sample ($n \geq 30$), it follows that the sampling distribution of $\bar{x}$ is approximately normally distributed with mean $\mu_{\bar{x}} = \mu$ and

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

Unit II In Sem.

The test statistic for single mean with known variance is $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

To test whether the population mean μ equals to a specified constant μ_0 or not, we formulate the test of hypothesis as :

(1) N.H. : $\mu = \mu_0$ (2) A.H. : $\mu \neq \mu_0$ (3) L.O.S. : α

(4) C.R. : Since A.H. is **not equal to type**, a Two-Tailed Test (T.T.T.) is considered. For given α, critical values $-Z_{\alpha/2}$ and $+Z_{\alpha/2}$ are determined from normal table ($\because$ normal distribution is assumed).

For example, for α = 5% or 0.05, from normal table $-Z_{0.025} = -1.96$ and $Z_{0.025} = 1.96$. Thus the critical region (C.R.) consists of two shaded regions in the Fig. 2.27.4.

i.e., reject N.H. (Null Hypothesis).

H_0 : if $Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$

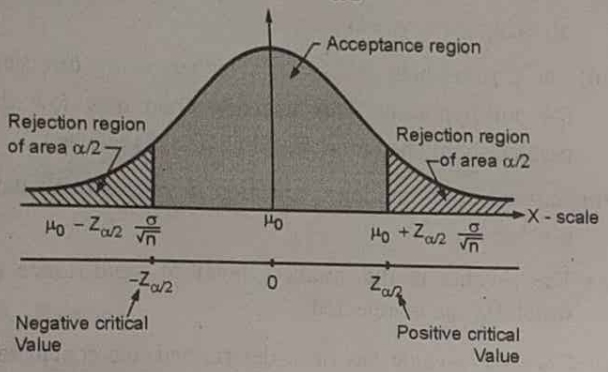


Fig. 2.27.4

5. Compute the test statistic Z, denoted by Z_{cal} or simply Z by

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Here $\bar{x}$, the mean of the sample of size n, is calculated from the simple data.

6. **Conclusion :** Reject H_0 if Z_{cal} or Z falls in the critical region i.e., observed sample. Statistic is probably significant at α – level. Otherwise accept H_0 (if $Z_{-\alpha/2} < Z < Z_{\alpha/2}$)

Note :

(1) Suppose the A.H. is $H_1 : \mu > \mu_0$, then the critical region is given by $Z > Z_\alpha$, since we consider a right one tail test in this case. i.e., reject H_0 if $Z > Z_\alpha$, otherwise accept H_0 if $(Z < Z_\alpha)$, as given in the Fig. 2.27.5.

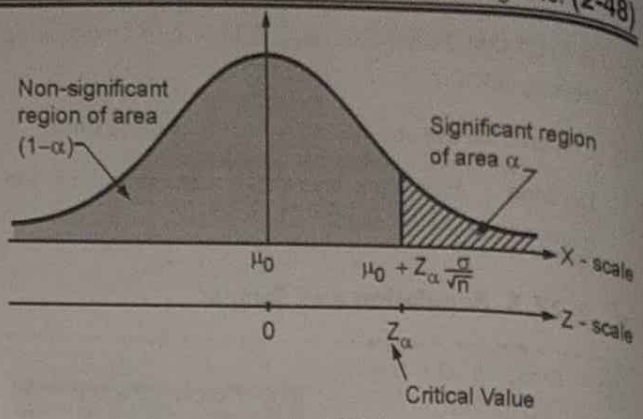


Fig. 2.27.5

(2) If A.H. (Alternative hypothesis) $H_1 : \mu < \mu_0$ then consider a left one tail test with C.R. given by $Z < -Z_\alpha$ as shown in the Fig. 8.1.6, i.e., reject H_0 if $Z < -Z_\alpha$, otherwise accept H_0. (if $Z > -Z_\alpha$).

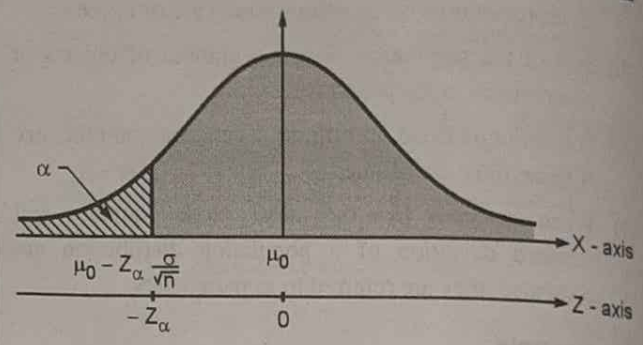


Fig. 2.27.6

(3) We prepare reference table of critical values for a given L.O.S. α, for T.T.T., R.O.T.T., L.O.T.T.

α %	15 %	10 %	5 %	4 %	1%	0.5 %	0.2 %
α	0.15	0.1	0.05	0.04	0.01	0.005	0.002
$-Z_{\alpha/2}$ and $+Z_{\alpha/2}$ for T.T.T.	− 1.44 1.44	− 1.645 1.645	− 1.96 1.96	− 2.06 2.06	− 2.58 2.58	− 2.81 2.81	− 3.08 3.08
$-Z_\alpha$ for L.O.T.T.	− 1.04	− 1.28	− 1.645	− 2.6	− 2.33	− 2.58	− 2.88
Z_α for R.O.T.T.	1.04	1.28	1.645	2.6	2.33	2.58	2.88

Ex. 2.27.1 : The length of life X of certain computers is approximately normally distributed with mean 800 hours and standard deviation 40 hours. If a random sample of 30 computers has an average life of 788 hours, test the null hypothesis that μ = 800 hours against the alternative that $\mu \neq$ 800 hours at (a) 0.5%, (b) 1 % (c) 4% (d) 5% (e) 10% (f) 15% level of significance

Soln. :

Case (a)

(i) Null Hypothesis : $\mu = 800$ hours

(ii) Alternate Hypothesis : $\mu \neq 800$ hours

(iii) α level of significance = 0.5% = 0.005

(iv) **Critical region :** Since alternate hypothesis ≠ type, the test is **two tailed** and critical region is,

$-2.81 < Z < 2.81$

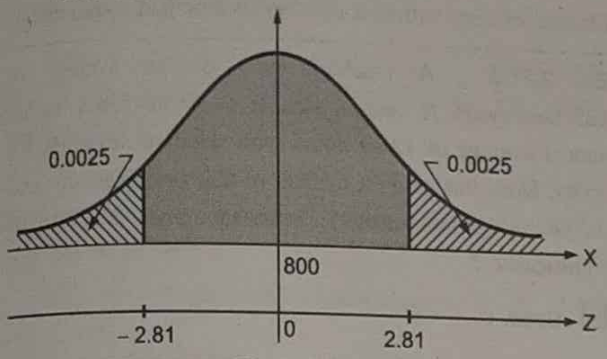


Fig. P. 2.27.1

(v) **Calculation of statistic**

Here, $\bar{x}$ = mean of the sample = 788

n = sample size,

standard deviation, $\sigma = 40$

$$\therefore Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{788 - 800}{40/\sqrt{30}} = -1.643$$

(vi) **Decision**

Accept the null hypothesis H_0,

Since $Z = -1.643 > -2.81 = Z_{\alpha/2} = Z_{0.0025}$;

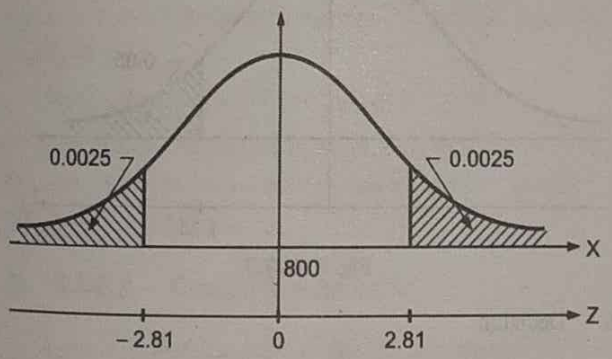


Fig. P. 2.27.1(a)

Case (b)

α = level of significance = 1 % = 0.01

Critical region $-2.58 < Z < 2.58$ (from above table)

Decision

Accept H_0, since

$Z = -1.643 > -2.58 = Z_{\alpha/2} = Z_{0.005}$

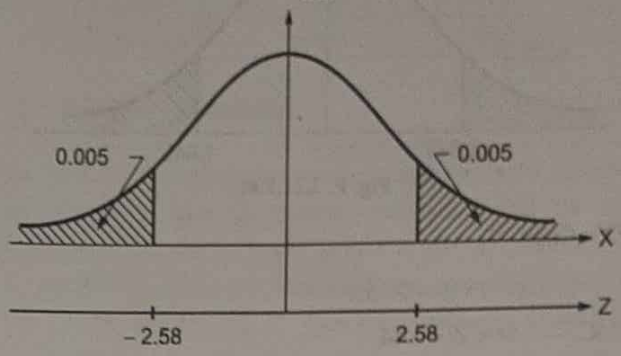


Fig. P. 2.27.1(b)

Case (c)

$\alpha = 4\% = 0.04$

C.R. : $-2.06 < Z < 2.06$

Accept H_0, since $Z = -1.643 > -2.06$

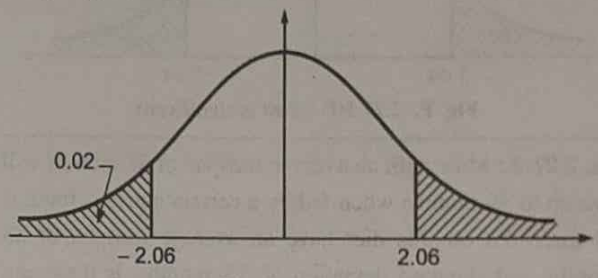


Fig. P. 2.27.1(c)

Case (d)

$\alpha = 5\% = 0.05$

C.R. : $-1.96 < Z < 1.96$ (from table)

Accept H_0, since $Z = -1.643 > -1.96$

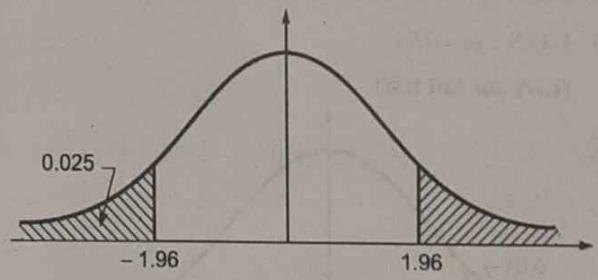


Fig. P. 2.27.1(d)

Case (e)

$\alpha = 10\ \% = 0.10$;

C.R. : $-1.645 < Z < 1.645$

Unit II In Sem.

Accept H_0, since $Z = -1.645 > -1.645$

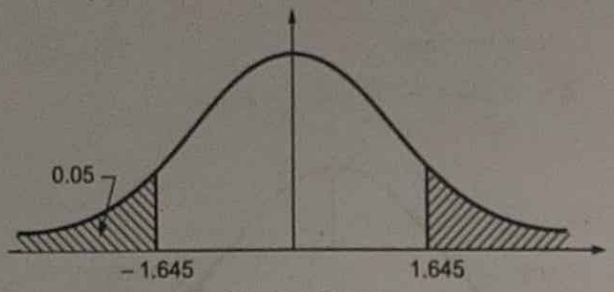


Fig. P. 2.27.1(e)

☞ **Case (f)**

$\alpha = 15\ \% = 0.15;$

C.R. : $-1.44 < Z < 1.44$

Reject H_0, since $Z = -1.643 > -1.44$

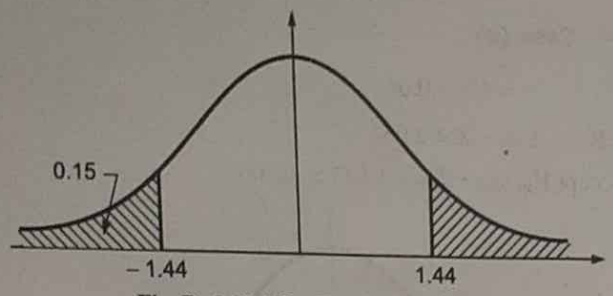


Fig. P. 2.27.1(f) : Test is significant

Ex. 2.27.2 : Mice with an average lifespan of 32 months will live up to 40 months when fed by a certain nutrious food. If 64 mice fed on this diet have an average lifespan of 38 months and standard deviation of 5.8 months, is there any reason to believe that lifespan is less than 40 months.

☑ **Soln. :**

Let μ = average lifespan of mice fed with nutritious food. Use 0.01 level of significance :

(1) N.H. : $H_0 : \mu = 40$ months (2) A.H. : $H_1 : \mu < 40$

(3) L.O.S. : $\alpha = 0.01$

(Left one tail test)

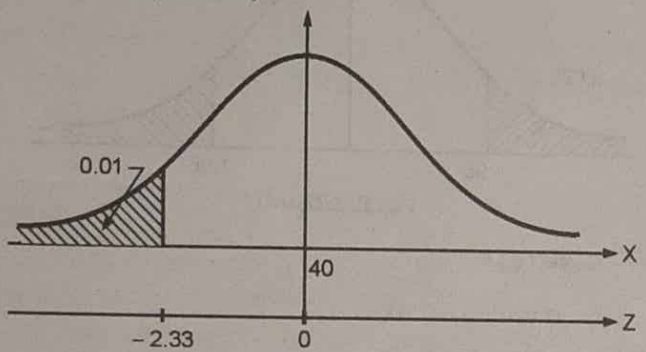


Fig. P. 2.27.2

4. C.R. : $Z < -Z_\alpha = -Z_{0.01} = -2.33$

5. Computation : Here $\bar{x} = 38, \sigma = 5.8, n = 64$

$$\therefore\ Z = \frac{38 - 40}{5.8/\sqrt{64}} = -2.76$$

6. Decision : Reject H_0 :

Since $Z = -2.76 < -2.33 = -Z_\alpha = -Z_{0.01}$

i.e., yes, there is reason to believe that the average lifespan of mice with nutrious food is less than 40 months.

Ex. 2.27.3 : A machine runs on an average of 125 hours/year. A random sample of 49 machines has an annual average of 126.9 hours with standard deviation 8.4 hours. Does this suggest to believe that machines are used on the average more than 125 hours annually at 0.05 level of significance ?

☑ **Soln. :**

Let μ = average number of hours a machine runs in an year.

1. $H_0 : \mu = 125$ hours/year

2. $H_1 : \mu > 125$

3. L.O.S. : $\alpha = 0.05$

4. **C.R. :** $Z > Z_\alpha = Z_{0.05} = 1.64$ (from table)

5. Calculation : $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{126.9 - 125}{8.4/\sqrt{49}} = 1.58$

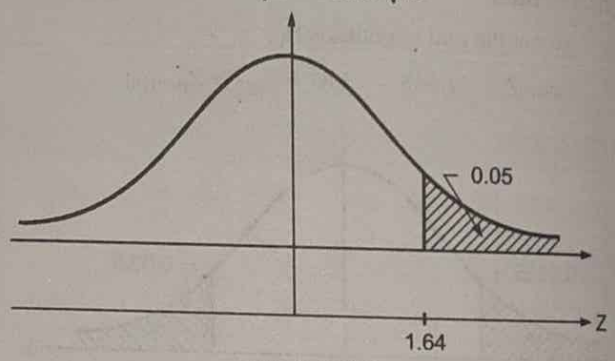


Fig. P. 2.27.3

6. Decision

Accept H_0, since $Z = 1.58 < 1.64 = Z_{0.05}$ i.e. cannot be believed that machine works more than 125 hours in an year.

2.28 TEST OF SIGNIFICANCE OF SMALL SAMPLES

For 'expensive' populations such as satellites, nuclear reactors, super computers etc; the investigation of large samples ($n \geq 30$) is uneconomical and time consuming.

In all such cases, the size of the sample drawn is small (i.e. $n < 30$).

If sample size n is small, then the normal test cannot be applied to the distribution of these standardised statistics.

Hence to deal with small samples, new techniques and tests of significance known as **'Exact Sample Test'** have been developed.

We note that 'Exact Sample Techniques' can be used, even for large samples but **large sample theory cannot be used for small samples.**

The basic fundamental assumptions in all the exact sample tests are:

(i) The parent population (s) from which the sample (S) is (are) drawn is (are) normally distributed.

(ii) The sample (S) is (are) random and independent of each other.

2.28.1 Degree of Freedom (DF)

By degree of freedom we mean the number of classes to which the value can be assigned arbitrarily without restrictions placed.

For example, we want to choose any five numbers whose sum is 40. clearly we can choose any four numbers say 5, 7, 9, 14; but the fifth number '5' is fixed, since the total is 40.

Hence, the freedom of selection is 5 – 1 = 4.

The degree of freedom is denoted by df and it is given by $\nu = n - k$ where n = number of classes and k = number of independent constraints.

2.28.2 Critical Values of 't'

Critical values for various degrees of freedom for the t-distribution are as :

n	Degrees of freedom	$t_{0.025}$
6	5	2·571
16	15	2·131
31	30	2·042
101	100	1·984
1001	1000	1·962
Normal	Infinite	1.960

2.28.3 Examples

Ex. 2.28.1 : An ambulance service company claims that on an average it takes 20 minutes between a call for an ambulance and the patient's arrival at the hospital. If in 6 calls the time taken (between a call and arrival at hospital) are 27, 18, 26, 15, 20, 32. Can the company's claim be accepted?

Unit II In Sem.

☑ Soln. :

Here n = 6 Let X be the time taken between a call and patient's arrival at hospital. From given data :

$\bar{x}$ = average time taken

$$= \frac{27 + 18 + 26 + 15 + 20 + 32}{6} = \frac{138}{6} = 23$$

and standard deviation : $S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$

$$\therefore S^2 = \frac{(27-23)^2 + (18-23)^2 + (26-23)^2 + (15-23)^2 + (20-23)^2 + (32-23)^2}{6-1}$$

$$= 40{\cdot}8$$

$$\therefore S = 6{\cdot}38748$$

Now;

1. **N.H. :** $X = 20$ minutes
2. **A. H. :** $X > 20$
3. **L.O.S. :** $\alpha = 0.10$
4. **Critical Region :** Reject N. H. if $t > t_\alpha = 1.476$ where $t_{0\cdot10}$ with $\nu = n - 1 = 6 - 1 = 5$ degrees of freedom
5. **Calculation :** $t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{23 - 20}{6{\cdot}39\sqrt{6}} = 1.15$

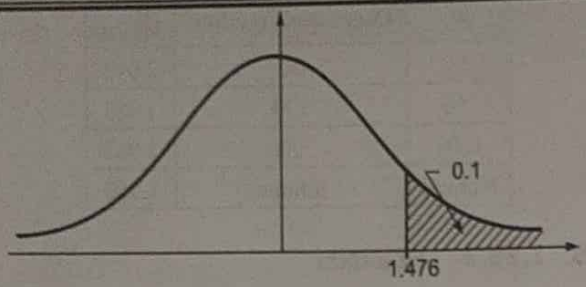


Fig. P. 2.28.1

6. **Decision :** Accept H_0, since $t = 1.15 < 1.476 = t_{0.1}$ with 5 d.f. Accept the claim of the company.

Ex. 2.28.2 : Ten cartons are taken at random from an automatic filling machine. The mean net weight of the 10 cartons is 11.8 ozone and standard deviation is 0.15 oz. Does the sample mean differ significantly from the intended weight of 12 oz. ? You are given that for $\gamma = 9$, $t_{0.05} = 2.26$.

☑ **Soln. :**

We have $n = 10$, $\bar{x} = 11.8$ oz, $S = 0.15$ oz

Null Hypothesis : N. H. : $H_0 : \mu_0 = 12$ oz, i.e. the sample mean $\bar{x} = 11.8$ oz. does not differ significantly from the population mean $\mu = 12$ oz

Alternate Hypothesis : (A· H) : $H_1 : \mu_1 \neq 12$ oz. (Two- tailed)

Test statistic : Under H_0, the test statistic is :

$$t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{\bar{x} - \mu}{\sqrt{s^2/n-1}}$$

$$= \frac{11.8 - 12}{0.15\sqrt{9}} = \frac{-0.2 \times 3}{0.15} = -4.0$$

Tabulated value of t for 9 d.f.at 5% level of significance is 2.26.

Since calculated | t | is much greater than tabulated t, it is highly significant. Hence null hypothesis is rejected, at 5% level of significance and we conclude that the sample mean **differs significantly from the mean $\mu = 12$ oz**

Ex. 2.28.3 : A machine designed to produce insulating washers for electrical devices of average thickness of 0.025 cm. A random sample of 10 workers was found to have an average thickness of 0.024 cm with a standard deviation of 0.0002 cm. Test the significance of the deviation value of t for 9 degrees of freedom at 5% level is 2.262.

☑ **Soln. :** We have : $n = 10$, $\bar{x} = 0.024$ cm; $S = 0.002$ cm

Null Hypothesis $H_0 : \mu = 0.025$ cm, i.e. there is no significant deviation between sample mean $\bar{x} = 0.024$ and population mean $\mu = 0.025$.

Altemate Hypothesis : $H_1 : \mu \neq 0.025$ cm

Test Statistic : $t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} = \frac{0.024 - 0.025}{0.002/\sqrt{9}}$

$= -1.5$

Tabulated $t_{0.05}$ for 9 d.f. = 2.262

Since | t | < 2.262, it is not significant at 5% level of significance.

Hence the deviation $(\bar{x} - \mu)$ is not significant.

⏭ 2.29 't' – TEST FOR SIGNIFICANCE OF SAMPLE CORRELATION COEFFICIENT

Suppose that a random sample (x_i, y_i), i = 1,2, ..., n of size n has been drawn from a bivariate normal distribution. Let r be the observed sample correlation coefficient.

The problem is to find if this sample correlation coefficient r is significant.

Prof. Fisher proved that under null hypothesis $H_0 : \rho = 0$, i.e., the variables are uncorrelated in the population, the statistic

$$t = \frac{r}{\sqrt{1 - r^2}} \cdot (n-2) \qquad ...(2.29.1)$$

Follows student's t-distribution with (n – 2) d.f., n being the sample size.

(I) 95% confidence limits for ρ are :

$$r \pm t_{n-2}(0.025) \cdot \frac{(1 - r^2)}{\sqrt{n}} \qquad ...(2.29.2)$$

(II) 99% confidence limits for ρ are :

$$r \pm t_{n-2}(0.005) \cdot \frac{(1 - r^2)}{\sqrt{n}} \qquad ...(2.29.3)$$

☞ **Method of Solving Problem**

(1) State the null hypothesis and alternate hypothesis

$H_0 : \rho = 0$; $H_1 : \rho \neq 0$

(2) State the significance level.

(3) Find the test statistic of correlation coefficient with the above-defined formula.

(4) Use critical value approach to make a decision.

(5) State the conclusion.

2.29.1 Examples

Ex. 2.29.1 : A random sample of 27 pairs of observations from a normal population gives a correlation coefficient of 0.42. Is it likely that the variables in the population are uncorrelated?

Soln. : We have $n = 27$ and $r = 0.42$

Null Hypothesis : $H_0 : \rho = 0$; i.e. variables are uncorrelated in the population

Alternate Hypothesis : $H_1 : \rho \neq 0$ (Two –tailed test)

Test Statistic : Under H_0, the test statistic is :

$$t = r \cdot \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$\therefore \quad t = 0.42 \frac{\sqrt{25}}{\sqrt{1-(0.42)^2}} = \frac{0.42 \times 5}{0.908} = 2.31$$

The tabulated value of t for 25 d.f and 5% level of significance for a two- tailed test is 2.06.

Since, calculated t > tabulated t, it is significant. Hence, null hypothesis ($\rho = 0$) is rejected at 5% level of significance and we conclude that variables are correlated in the population.

Ex. 2.29.2 : Find the least value of r in a sample of 18 pairs of observations from a bivariate normal population, significant at 5% level of significance.

Soln. :

We have : n = 1.8. The observed value of sample correlation coefficient r will be significant at 5% level of significance if

$$|t| = \left| \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right| > t_{n-2}(0.025) = t_{16}(0.025) \qquad ...(i)$$

But from the table, $t_{16}(0.025) = 2 \cdot 12$

$$\therefore \text{ From, Equation (i) , } \left| \frac{r\sqrt{18-2}}{\sqrt{1-r^2}} \right| > 2 \cdot 12$$

$$\therefore \left| \frac{4r}{\sqrt{1-r^2}} \right| > 2 \cdot 12$$

Squaring and transposing,

$$16 r^2 > (2 \cdot 12)^2 (1-r^2) = 4.4944 (1-r^2)$$

$$\therefore (16 + 4.4944) r^2 > 4.4944$$

$$\therefore \quad r^2 > \frac{4.4944}{20.4944}$$

$$\therefore |r| > \sqrt{0.2193} = 0.4683$$

Ex. 2.29.3 : A coefficient of correlation of 0.2 is derived from a random sample of 625 pairs of observations :

(i) Is this value significant?

(ii) What are the 95% and 99% confidence limits for the correlation coefficient in the population ?

Soln. :

The null hypothesis : $H_0 : \rho = 0$, i.e. the value of r = 0.2 is not significant, against $H_1 : \rho \neq 0$.

The test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{(0.2)\sqrt{625-2}}{\sqrt{1-0.04}} = \frac{0.2\sqrt{623}}{\sqrt{0.96}}$$

$$= 5.09$$

Now, d.f. = 625 – 2 = 623

Since d.f. is greater than 60, the significant values of t are same as in case of normal distribution, i.e.

$t_{0\,05} = 1.96$ and $t_{0\cdot01} = 2.58$ for two-tailed test

Since calculated 't' is much greater than these values, it is significant.

$\therefore H_0 : \rho = 0$, is rejected and we conclude that the sample correlation is significant of correlation.

(I) 95% Confidence Limits for ρ

$$r \pm 1.96\left[\frac{1-r^2}{\sqrt{n}}\right] = 0.2 \pm 1.96\left[\frac{0.96}{\sqrt{625}}\right]$$

$$= 0.2 \pm 1.96 \times 0.0384$$

$$= 0.2 \pm 0.075 = (0.125, 0.275)$$

(II) 99% Confidence Limits for ρ

$$0.2 \pm 2.58 \times 0.0384 = 0.2 \pm 0.099$$

$$= (0.101, 0.299)$$

2.30 t-TEST FOR DIFFERENCE OF MEANS

Suppose we want to test, if two independent samples have been drawn from two normal populations, having the same means, the population variances being equal.

Unit II In Sem.

Let $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ be two independent random samples from the given normal populations.

We set up the **Null Hypothesis :** $H_0 : \mu_x = \mu_y$, i.e. the samples have been drawn from the normal populations with the **same means;** i.e. the **sample means** $\bar{x}$ and $\bar{y}$ **do not differ significantly.**

Under the assumption that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ i.e., population variances are equal but unknown, the test statistic under H_0 is :

$$t = \frac{\bar{x} - \bar{y}}{S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where, $\bar{x} = \frac{\sum x}{n_1}, \bar{y} = \frac{\sum y}{n_2}$

and $S^2 = \frac{1}{n_1 + n_2 - 2} [\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2]$

By comparing the computed value of t with the tabulated value of t for $(n_1 + n_2 - 2)$ d.f. and at desired level of significance, usually 5% or 1%, we reject or retain the null hypothesis H_0.

2.30.1 Assumptions for Difference of Means Test

(i) Parent populations from which samples have been drawn are normally distributed.

(ii) The two samples are random and independent of each other.

(iii) $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown) i.e. population variance are equal and unknown.

2.30.2 Examples

Ex. 2.30.1 : The nicotine content in milligrams of two samples of tobacco were found to be as follows :

Sample A :	24	27	26	21	25	
Sample B :	27	30	28	31	22	36

Can it be said that two samples came from normal population having the same mean?

Soln. :

Null Hypothesis : $H_0 : \mu_1 = \mu_2$, i.e. the two samples have been drawn from the normal populations with the same mean.

Alternative Hypothesis : $H_1 : \mu_1 \neq \mu_0$

Computation of Sample means and S.D'S

	Sample A			Sample B	
x	d = x − A = x − 25	d^2	y	D = y − B = y − 30	D^2
24	−1	1	27	−3	9
27	2	4	30	0	0
26	1	1	28	−2	4
21	−4	16	31	1	1
25	0	0	22	−8	64
			36	6	36
Total	2	22	Total	−6	114

Now, $\bar{x} = A + \frac{\sum d}{n_1} = 25 + \frac{(-2)}{5} = 24.6$

$$\sum (x - \bar{x})^2 = \sum d^2 - \frac{(\sum d)^2}{n_1} = 22 - \frac{4}{5} = 21.2$$

$$\bar{y} = B + \frac{\sum D}{n_2} = 30 + \frac{(-6)}{6} = 29$$

$$\sum (y - \bar{y})^2 = \sum D^2 - \frac{(\sum D)^2}{n_2}$$

$$= 114 - \frac{36}{6} = 108$$

$$\therefore S^2 = \frac{1}{n_1 + n_2 - 2} \left[\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2\right]$$

$$= \left(\frac{21.2 + 108}{9}\right) = 14.36$$

Test statistic : Under H_0, the test statistic is :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{24.6 - 29}{\sqrt{14.36 \left(\frac{1}{5} + \frac{1}{6}\right)}}$$

$$= \frac{-4.4}{\sqrt{14.36 \times \frac{11}{30}}} = \frac{-4.4}{2.2946} = -1.92$$

The tabulated value of 't' for 9 d.f. at 5% level of significance for two- tailed test is 2.262.

Since calculated 't' is less than tabulated 't', it is not significant.

Hence null hypothesis may be accepted at 5% level of significance and we conclude that the samples come from normal populations with the same mean.

Ex. 2.30.2 : A random sample of 20 daily workers of state A was found to have average daily earning of Rs. 44 with sample variance 900. Another sample of 20 daily workers from state B was found to earn on an average Rs. 30 per day with sample variance 400.

Test whether the workers in state A are earning more than in state B.

Soln. :

Let the daily earnings (in Rs.) of the workers in states A and B be denoted by the variables X and Y respectively. Then we have

$n_1 = 20, \bar{x} = 44, S_x^2 = 900; \quad n_2 = 20, \bar{y} = 30, S_y^2 = 400$

Null Hypothesis

$H_0 : \mu_x = \mu_y$, i.e. there is no significant difference in the average daily earnings of the workers in states A and B.

Alternative Hypothesis : $H_1 : \mu_x > \mu_y$ (Right tailed)

Test Statistic : Under H_0,

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad ...(i)$$

where $S^2 = \left(\frac{n_1 S_x^2 + n_2 S_y^2}{n_1 + n_2 - 2}\right)$

$$= \left(\frac{20 \times 900 + 20 \times 400}{38}\right) = \frac{18000 + 8000}{38}$$

$$= 648.21 \quad ...(ii)$$

$$\therefore \quad t = \frac{44 - 30}{\sqrt{648.21\left(\frac{1}{20} + \frac{1}{20}\right)}} = \frac{14}{\sqrt{64.821}}$$

$$= 1.7389 \quad ...(iii)$$

Now, tabulated $t_{0.05}$ for d.f. $= n_1 + n_2 - 2 = 38$, for right tailed test is 1.645 (for d.f. > 30, since significant values of t are same as those of Z for the normal test).

Since calculated t is greater than tabulated t. It is significant at 5% level of significance.

Hence H_0 is rejected; i.e. H_1 is accepted at 5% level of significance.

Hence we conclude that the workers in state A are earning more than those in state B.

2.31 CHI-SQUARE TEST FOR INDEPENDENCE OF ATTRIBUTES

Let us suppose that the given population consisting of N items is divided into 'r' mutually disjoint and exhaustive classes $A_1, A_2, ...A_r$ w.r.t. the attribute A.

Similarly, let us suppose that the same population is divided into 'S' mutually disjoint and exhaustive classes $B_1, B_2, ...B_S$ w.r.t. another attribute B.

Unit II In Sem.

Under the **null hypothesis** that the **two attributes A and B are independent**, the **expected frequency** for $(A_i B_j)$ is given by

$$E[(A_i B_j)] = N.P[A_i B_j] = N P(A_i) \cdot P(B_j)$$

($\because$ attributes are independent)

$$= N \cdot \frac{(A_i)}{N} \times \frac{(B_j)}{N} = \frac{(A_i)(B_j)}{N}$$

2.31.1 The Rule of Expected Frequency

Under the hypothesis of independence of attributes the expected frequencies for each of the cell can be obtained on :

(i) Multiplying the row totals and column totals in which the frequency occurs, and

(ii) Dividing the product by the total frequency N.

2.31.2 Calculation of χ^2

(i) We have a set of $(r \times S)$ observed frequencies $(A_i B_j)$.

(ii) We denote the expected frequency of $(A_i B_j)$ by $(A_i B_j)_0$, i.e.

$$(A_i B_j)_0 = \frac{(A_i)(B_j)}{N}, \; i = 1 \text{ to } r$$

$$j = 1 \text{ to } S$$

(iii) $$\chi^2 = \sum_i \sum_j \left[\frac{[(A_i B_j) - (A_i B_j)_0]^2}{(A_i B_j)_0}\right] = \sum\left[\frac{(O - E)^2}{E}\right]$$

This statistic follows χ^2 distribution with $(r - 1) \times (S - 1)$ degrees of freedom.

(iv) Comparing the calculated value of χ^2 with the tabulated value for $(r - 1) \times (s - 1)$ d.f. and at certain level of significance, we reject or retain the null hypothesis of independence of attributes at that level of significance.

2.31.3 Examples

Ex. 2.31.1 : Suppose that, in a public opinion survey to the questions (i) Do you drink ? (ii) Are you in favour of local opinion on sale of liquor ? were as given in Table. Can you infer that opinion on local option is dependent on whether or not an individual drinks?

Question (b)	Question (a)		Total
	Yes	No	
Yes	56	31	87
No	18	6	24
Total	74	37	111

☑ **Soln. :** We set up

Null Hypothesis : H_0 : Two attributes are independent, i.e. the local option on sale of liquor is independent of whether or not an individual drinks.

Under this hypothesis, we calculate the expected frequencies :

$$E(56) = \frac{87 \times 74}{111} = 58; \; E(31) = \frac{87 \times 37}{111} = 29;$$

$$E(18) = \frac{24 \times 74}{111} = 16, \; E(6) = \frac{24 \times 37}{111} = 8$$

O	E	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
56	58	– 2	4	4/58
31	29	2	4	4/29
18	16	2	4	4/16
6	8	– 2	4	4/8

$$\text{Now, } \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$$

$$= 4\left[\frac{1}{58} + \frac{1}{29} + \frac{1}{16} + \frac{1}{8}\right]$$

$$= 4\,[0.01724 + 0.03448 + 0.06250 + 0.12500]$$

$$= 4\,[0.23922] = 0.95688 \quad \ldots(i)$$

Here, d.f. = (2 – 1) (2 – 1) = 1

From Table, we have $\chi^2_1(0.05) = 3.841$ and

$\chi^2_1(0.01) = 6.635$

Since the calculated value of χ^2, i.e., 0.95688 is much less than the tabulated value of χ^2 for 1 d.f. at both 1% and 5% levels of significance, it is not significant.

Hence H_0 is rejected and we conclude that opinion on local option on sale of liquor is not dependent on **whether or not individual drinks.**

Ex. 2.31.2 : In a certain sample of 2,000 families 1,400 families are consumers of tea. Out of 1,800 Hindu families 1,236 families consume tea. Use χ^2 - test and state whether there is any significant difference between consumption of tea among Hindu and non-Hindu families.

☑ **Soln. :**

We arrange the data in tabular form :

Number of ↓	Hindu	Non- Hindu	Total
Families consuming tea	1236	164	1400
Families not consuming tea	564	36	600
Total	1800	200	2000

Null Hypothesis : H_0 : The two attributes : 'Consumption of tea' and the 'community' are independent;

i.e. there is no significant difference between the consumption of tea among Hindu and non-Hindu families.

under the null-hypothesis of independence :

$$E(1236) = \frac{1800 \times 1400}{2000} = 1260 \ldots$$

We prepare the tables :

Expected frequencies :

1260	140	1400
540	60	600
1800	200	2000

Computation of χ^2 :

O	E	O – E	$(O-E)^2$
1236	1260	– 24	576
564	540	24	576
164	140	24	576
36	60	24	576

$$\text{Now, } \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$$

$$= 576\left[\frac{1}{1260} + \frac{1}{540} + \frac{1}{140} + \frac{1}{60}\right]$$

$$= 576\,[0.000794 + 0.001852 + 0.007143 + 0.016667]$$

$$= 576 \times 0.026456 = 15.2387 \quad \ldots(i)$$

Now, d.f. = (2 – 1) × (2 – 1) = 1

Tabulated value $\chi^2_{0.05}$ for 1 d.f. = 3.841 ...(ii)

Conclusion : Since the calculated value of χ^2 i.e. 15.2387 is much greater than the tabulated value of χ^2 at 5% level of significance, it is highly significant. Hence the null hypothesis H_0 is rejected.

Thus we conclude that with 95% confidence that the two communities differ significantly as regards consumption of tea among them.

Ex. 2.31.3 : In a survey of 200 boys of which 75 were intelligent, 40 had skilled fathers, while 85 of the unintelligent boys had unskilled fathers. Do these figures support the hypothesis that skilled fathers have intelligent boys ? (Value of χ^2 for 1.d.f. is 3.841)

☑ Soln. :

We prepare the data in a tabular form (called as 2 × 2 contingency table).

	Intelligent boys	Unintelligent boys	Total
Skilled Father	40	125 – 85 = 40	40 + 40 = 80
Unskilled Fathers	120 – 85 = 35	85	200 – 80 = 120
Total	75	200– 75 = 125	200

Null Hypothesis : H_0 : The two attributes, 'skill of fathers' and 'intelligence of boys are independent.

In other words : The skill of fathers does not have any effect on the intelligence of boys.

We calculate expected frequencies

$$E(40) = \frac{75 \times 80}{200} = 30, \qquad E(35) = \frac{75 \times 120}{200} = 45$$

$$E(40) = \frac{125 \times 80}{200} = 50, \qquad E(85) = \frac{125 \times 120}{200} = 75$$

O	E	O – E	$(O-E)^2$	$(O-E)^2/E$
40	30	10	100	3.33
35	45	10	100	2.22
40	50	10	100	2.00
85	75	10	100	1.33

$$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$$
$$= 3.33 + 2.22 + 2.00 + 1.33$$
$$= 8.88$$

and d.f. = (2 – 1) (2 – 1) = 1

Tabulated $\chi^2_{0.05}$ for 1 d.f. = 3.841

Conclusion : Since the calculated value of χ^2 i.e. 8.88 is greater than the tabulated value of χ^2 at 5% level of significance, it is significant and hence the null hypothesis H_0 is rejected. Hence we concludes that 'skill' of the fathers has a significant effect on the intelligence of boys.

Unit II In Sem.

2.31.4 2 × 2 Contingency Table

Under the null hypothesis of independence of attributes, the value of χ^2 for 2 × 2 Table is

		Total
a	b	a + b
c	d	c + d
a + c	b + d	N = a + b + c + d

$$\chi^2 = \frac{N(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Where N = a + b + c + d is the total frequency

2.31.5 Yates Correction for Continuity for 2 × 2 Table

When any cell frequency in **2 × 2 table** is less than 5, we apply 'Yates correction for continuity'. This consists in adding 0.5 to the cell frequency which is less than 5 and adjusting the remaining frequencies accordingly, since row and column totals are fixed.

After applying Yates correction the value of χ^2 for 2 × 2 contingency table is $\chi^2 = \frac{\left[|ad - bc| - \frac{N}{2}\right]^2}{(a+c)(b+d)(a+b)(c+d)}$

Where | ad – bc | is modulus value.

Remark : Yates Correction for continuity can be applied only in the case of 2 × 2 table.

Ex. 2.31.4 : In an experiment on the immunisation of goats from Anthrax,. the following results were obtained. Derive your inference on the efficiency of the vaccine.

	Died of Anthrax	Survived	Total
Innoculated with vaccine	2	10	12
Not Innoculated	6	6	12
Total	8	16	24

☑ **Soln. :**

(I) Without Yate's Correction

Null Hypothesis : 'Inooculation with vaccine' and 'the survival from Anthrax' are independent

Now, expected frequencies are

$$E(2) = \frac{8 \times 12}{24} = 4, E(10) = \frac{12 \times 16}{24} = 8$$

$$E(6) = \frac{8 \times 12}{24} = 4, E(6) = \frac{16 \times 12}{24} = 8$$

Expected Frequencies

4	12 – 4 = 8	12
8 – 4 = 4	16 – 8 = 8	12
8	16	24

Computation of χ^2

O	E	O – E	$(O-E)^2$	$(O-E)^2/E$
2	4	2	4	1.0
10	8	2	4	0.5
6	4	2	4	1.0
6	8	2	4	0.5
24	24	0		3.0

$$\text{Now, } \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right]$$

$$= 3.0$$

$$\text{Here d.f.} = (2-1)(2-1) = 1$$

The tabulated value of χ^2 for 1 d.f. at 5% level of significance is 3.841.

Since calculated value of χ^2 is less than tabulated value, it is not significant. Hence null hypothesis H_0 may be accepted. We conclude that vaccine is ineffective in disease

(II) Applying Yate's Correction

Since the cell frequency 2 (in the first cell) is less than 5, we apply Yate's correction. This consists in adding 0.5 to the cell frequency which is less than 5 and adjusting the remaining frequencies as shown.

Under null hypothesis, the expected frequencies are not affected.

2.5	9.5	12
5.5	6.5	12
8	16	24

Computation of χ^2

	O	E	O – E	$(O-E)^2$	$(O-E)^2/E$
	3.5	4.0	1.5	2.25	0.56250
	9.5	8.0	1.5	2.25	0.28125
	5.5	4.0	1.5	2.25	0.56250
	6.5	8.0	1.5	2.25	0.28125
Total	24.0	24.0	0		1.68750

$$\therefore \chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 1.68750$$

Since the value is less than the tabulated value, i.e. 3.841; χ^2 is not significant and hence H_0 is accepted,

2.31.6 Miscellaneous Examples

UEx. 2.31.5 **SPPU – Dec.15, 7 Marks**

A Machine is producing bolts of which a certain fraction is defective. A random sample of 400 is taken from a large batch and is found to contain 30 defective bolts. Does this indicate that the proportion of defectives is larger than that claimed by the manufacturer who claims that only 5% of the products are defective? Find the 95% confidence limits of the proportion of defective bolts in batch. ($Z_\alpha = 1.645$ at 5% level of significance).

☑ **Soln. :**

Null Hypothesis : H_0 : The manufacture's claim is accepted.

We have $\mu_0 = \frac{5}{100} = 0.05$, (Right – tailed test)

and $\bar{x}$ = observed proportion of sample $= \frac{30}{400} = 0.075$

Let $p = 0.05$, $\therefore q = 1 - P = 0.95$

and $S^2 = pq = (0.05)(0.95)$

$$\text{Let } Z = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{0.075 - 0.05}{\sqrt{0.05 \times 0.95/400}} = 2.179$$

Conclusion : The tabulated value of Z at 5% level of significance is $Z_\alpha = 1.645$,

$\because Z_\alpha > 1.645$, $\therefore H_0$ is rejected at 5% level of significance, i.e. the portion of defective bolts is larger than the manufacture claims.

Limits for 95% level of confidence :

$$0.05 \pm 1.645 \frac{\sqrt{0.05 \times (0.95)}}{\sqrt{400}} = (0.0679, 0.0321)$$

Ex. 2.31.6 : An IQ test was administered to 5 persons before and then they were trained. The results are given below :

Candidate	I	II	III	IV	V
IQ before tr	110	120	123	132	125
IQ after tr	120	118	125	136	121

Test whether there is any change in IQ after the training programme (test at 1% level of significance).

☑ Soln. :

Null Hypothesis : $H_0 : \mu_1 = \mu_2$, i.e. there is significant change in IQ after the programme.

Alternate Hypothesis : $\mu_1 \neq \mu_2$ (two-tailed test) level of significance : $\alpha = 0.01$

x	110	120	123	132	125	Total
y	120	118	125	136	121	-
d = x – y	– 10	2	– 2	– 4	4	– 10
d^2	100	4	4	16	16	140

$$\bar{d} = \frac{\sum d}{n} = \frac{-10}{5} = -2$$

$$\text{and } S^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$$

$$= \frac{1}{4}\left[140 - \frac{100}{5}\right] = 30$$

Statistic : Under H_0, the test statistic t is,

$$t_0 = \left|\frac{\bar{d}}{S/\sqrt{n}}\right| = \left|\frac{-2}{\sqrt{30/5}}\right| = \frac{2}{2.45} = 0.816$$

Expected value : $t_e = \left|\frac{\bar{d}}{S^2/n}\right| = 4.604$

with 5 – 1 = 4 d.f.

Conclusion : Since $t_0 < t_e$ at 1% level of significance, we accept the null hypothesis. Hence there is no change in IQ after the training programme.

UEx. 2.31.7 **15, 7 Marks**

The mean lifetime of 100 light bulbs produced by a company is computed to be 1570 hours with a standard deviation of 120 hours. If μ is the mean lifetime of all the bulbs produced by the company, test hypothesis $\mu = 1600$ hours against the alternative hypothesis $\mu \neq 1600$ hours using a 5% level of significance.

☑ Soln. :

Unit II In Sem.

Given data : $\mu = 1600$ hours, S = 120 hours,

$n = 100$, $\bar{X} = 1570$ hours.

Null Hypothesis : $H_0 : \mu = 1600$ hours, i.e. there is no significant difference between the sample mean and population mean.

Alternate Hypothesis : $H_1 : \mu \neq 1600$ (two-tailed alternative)

Level of significance $\alpha = 0.05$

Calculation of Statistic

Under H_0, the test statistic is

$$Z_0 = \left|\frac{\bar{x} - \mu}{S/\sqrt{n}}\right| = \left|\frac{1570 - 1600}{120/\sqrt{100}}\right| = 2.5$$

Expected value : $Z_e = \left|\frac{\bar{x} - \mu}{S/\sqrt{n}}\right| = 1.96$ for $\alpha = 0.05$

Conclusion : Since the calculated value of Z_0 is greater than Z_e, we rejected the null hypothesis at 5% level of significance. Thus there is no significant difference between the sample mean and the population mean.

2.32 CORRELATION

- Correlation is the relationship that exists between two or more variables. Two variables are said to be **correlated** if a change in one variable affects a change in the other variable. Such a data connecting two variables is called bivariate data.
- Correlation measures the closeness of the relationship between the variables.
- Some examples of a relationship are as follows:

  Relationship between heights and weights

  Rainfall and crop-yield correlated.

  Relationship between age of husband and age of wife

2.32.1 Types of Correlation

- Correlation is classified into four types:
  (1) Positive correlation and negative correlation
  (2) Simple correlation and multiple correlation
  (3) Partial correlation and total correlation
  (4) Linear correlation and nonlinear correlation

1. Positive and negative correlations

a. Positive correlation

- If the value of one variable increases, the value of the other variable also increases, or, if value of one variable decreases, the value of the other variable also decreases. This type of correlation is said to be **positive correlation**.

  e.g. The correlation between heights and weights of group of persons

Height (cm)	145	150	160	162	165	175
Weight (kg)	55	60	62	65	67	68

b. Negative correlation

If the value of one variable increases, the value of the other variable decreases, or, if value of one variable decreases, the value of the other variable increases. This type of correlation is said to be **negative correlation**.

e.g. The correlation between the price and demand of a commodity

Price (Rs per unit)	15	10	8	7	6	3
Demand (units)	150	200	220	260	300	320

2. Simple and multiple correlations

a. Simple correlation

The relationship between only two variables is described as simple correlation. e.g. The quantity of money and price level, demand and price

b. Multiple correlation

The relationship between more than two variables is described as multiple correlation. e.g. Relationship between price, demand and supply of a commodity

3. Partial and total correlations

a. Partial correlation

When more than two variables are studied excluding some other variables, the relationship is termed as partial correlation.

b. Total correlation

When more than two variables are studied without excluding any variables, the relationship is termed as total correlation.

4. Linear and nonlinear correlations

a. Linear correlation

If the ratio of change between two variables is constant, the correlation is said to be linear.

The graph of a linear relationship will be a straight line.

e.g.

Milk (l)	5	10	15	20	25	30
Curd (kg)	2	4	6	8	10	12

b. Nonlinear correlation

If the ratio of change between two variables is not constant, the correlation is said to be nonlinear.

The graph of a nonlinear relationship will be a curve.

e.g.

Price (Rs per unit)	15	10	8	7	6	3
Demand (units)	150	200	220	260	300	320

2.32.2 Scatter Diagram

There are various relationship between two variables represented by the following scatter diagrams.

- **Perfect positive correlation :** If all the plotted points lie on a straight line rising from the lower hand corner to the upper righthand corner, the correlation is said to be perfect positive correlation.

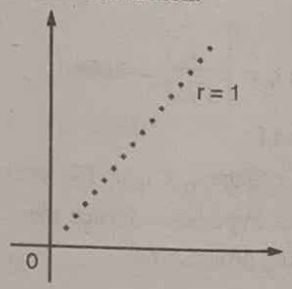


Fig. 2.32.1 : Perfect positive correlation

- **Perfect negative correlation:** If all the plotted points lie on a straight line from the upper left-hand corner to the lower right-hand corner, the correlation is said to be perfect negative correlation.

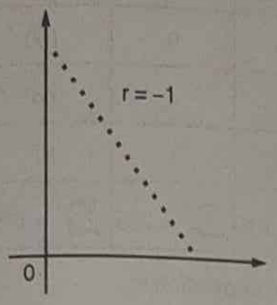


Fig. 2.32.2 : Perfect negative correlation

- **High degree of positive correlation:** If all the plotted points lie in the narrow strip, rising from the lower left-hand corner to the upper right -hand corner, it indicates a high degree of positive correlation.

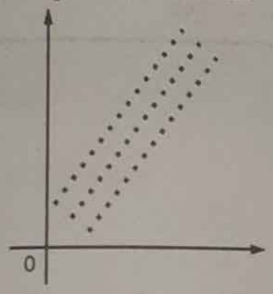


Fig. 2.32.3 : High degree of positive correlation

- **High degree of negative correlation:** If all the plotted points lie in a narrow strip, falling from the upper left-hand corner to the lower right-hand corner, it indicates the existence of a high degree of negative correlation. degree of negative correlation.

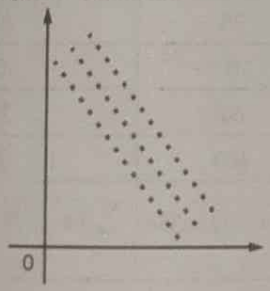


Fig. 2.32.4 : High degree of negative correlation

- **No correlation :** If all the plotted points lie on a straight line parallel to the x-axis or y-axis, it indicates the absence of any relationship between the variables.

2.32.3 Karl Pearson's Coefficient of Correlation

The coefficient of correlation is the measure of correlation between two random variables X and Y, is denoted by r.

$$r = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \qquad \ldots(2.23.3)$$

Where,

cov (X, Y) = the covariance of variables X and Y

$$= \frac{1}{n}\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)$$

σ_X = the standard deviation of variable X

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

σ_Y = the standard deviation of variable Y

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

So,

$$r = \frac{\frac{1}{n}\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n}}\sqrt{\frac{\sum\left(y-\bar{y}\right)^2}{n}}}$$

By simplifying,

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

The above expression r is called Karl Pearson's coefficient of correlation.

2.32.4 Properties of Coefficient of Correlation

(1) The coefficient of correlation lies between – 1and 1. i.e. $-1 \le r \le 1$

(2) Correlation coefficient is independent of change of origin and change of scale. i.e. $r_{xy} = r_{d_x\, d_y}$

Here, $d_x = \frac{x-a}{h}$ and $d_y = \frac{y-b}{k}$

i.e. $$r = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{n\sum d_x^2 - \left(\sum d_x\right)^2}\sqrt{n\sum d_y^2 - \left(\sum d_y\right)^2}}$$

Unit II In Sem.

(3) Two independent variables are uncorrelated. i.e. r = 0.

UEx. 2.32.1 (20, 4 Marks)

Calculate the coefficient of correlation for the following data.

x	9	8	7	6	5	4	3	2	1
y	15	16	14	13	11	12	10	8	9

☑ **Soln. :** Here, n = 9

x	y	x^2	y^2	xy
9	15	81	225	135
8	16	64	256	128
7	14	49	196	98
6	13	36	169	78
5	11	25	121	55
4	12	16	144	48
3	10	9	100	30
2	8	4	64	16
1	9	1	81	9
$\sum x = 45$	$\sum y = 108$	$\sum x^2 = 285$	$\sum y^2 = 1356$	$\sum xy = 597$

The coefficient of correlation is

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

$$= \frac{9(597) - (45)(108)}{\sqrt{9(285) - 45^2}\sqrt{9(1356) - 108^2}}$$

$$r = 0.95$$

UEx. 2.32.2 (19, 3 Marks)

Calculate the correlation of coefficient between the following data.

x	5	9	13	17	21
y	12	20	25	33	35

☑ **Soln. :**

Here, n = 5

$$\bar{x} = \frac{\sum x}{n} = \frac{65}{5} = 13 \quad ; \quad \bar{y} = \frac{\sum y}{n} = \frac{125}{5} = 25$$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
5	12	−8	−13	64	169	104
9	−20	−4	−5	16	25	20
13	25	0	0	0	0	0
17	33	4	8	16	64	32
21	35	8	10	64	100	80
$\sum x = 65$	$\sum y = 125$	$\sum (x - \bar{x}) = 0$	$\sum (y - \bar{y}) = 0$	$\sum (x - \bar{x})^2 = 160$	$\sum (y - \bar{y})^2 = 358$	$\sum (x - \bar{x})(y - \bar{y}) = 236$

The coefficient of correlation is

$$r = \frac{\frac{1}{n}\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\frac{\sum (x - \bar{x})^2}{n}}\sqrt{\frac{\sum (y - \bar{y})^2}{n}}} = \frac{\frac{1}{5}(236)}{\sqrt{\frac{160}{5}}\sqrt{\frac{358}{5}}} = 0.986$$

Ex. 2.32.3 : Calculate the coefficient of correlation for the following pairs of x and y :

x	17	19	21	26	20	28	26	27
y	23	27	25	26	27	25	30	33

Soln. :

Here, n = 8. Let a = 23 and b = 27 be the assumed mean of x and y respectively.

So, $d_x = x - a = x - 23$ and $d_y = y - b = y - 27$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
17	23	– 6	– 4	36	16	24
19	27	– 4	0	16	0	0
21	25	– 2	– 2	4	4	4
26	26	3	– 1	9	1	– 3
20	27	– 3	0	9	0	0
28	25	5	– 2	25	4	– 10
26	30	3	3	9	9	9
27	33	4	6	16	36	24
$\sum x = 184$	$\sum y = 216$	$\sum d_x = 0$	$\sum d_y = 0$	$\sum d_x^2 = 124$	$\sum d_y^2 = 70$	$\sum d_x d_y = 48$

The coefficient of correlation is

$$r = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{n\sum d_x^2 - (\sum d_x)^2}\sqrt{n\sum d_y^2 - (\sum d_y)^2}} = \frac{8(48)}{\sqrt{8(124)}\sqrt{8(70)}} = 0.515$$

Unit II In Sem.

2.32.5 Examples on Correlation Coefficient

Ex. 2.32.4 : From the following data which shows the ages X and systolic B.P. Y of 12 women. Are the two variable ages X and B.P. Y correlated ?

Age (X)	56	42	72	36	63	47	55	49	38	42	68	60
B.P. (Y)	147	125	160	118	149	128	150	145	115	140	152	155

Soln. :

We determine correlation coefficient r to find the association between age and B.P.

Now, $$r = \frac{N\sum XY - \sum X \cdot \sum Y}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum (Y^2) - (\sum Y)^2]}}$$

$$= \frac{12\,(89894) - (628)\,(1684)}{\sqrt{[(12)\,(34416) - (628)^2]\,[(12)\,(238822) - (1684)^2]}}$$

$$= 0.8961$$

∴ Age X and B.P. Y are strongly positively correlated.

Ex. 2.32.5 : From 10 pairs of observations for x and y the following data is obtained :

$n = 10, \sum x = 66, \sum y = 69, \sum x^2 = 476, \sum y^2 = 521,$

$\sum xy = 485$. It was later found that two pairs of (correct) values.

x	y
4	6
9	8

Were copied down as

x	y
2	3
7	5

Calculate the correct value of the coefficient of correlation

☑ **Soln. :** To obtain the correct data, we subtract the incorrect data and add the correct data as :

$$\sum x = 66 - 2 - 7 + 4 + 9 = 70$$

$$\sum y = 69 - 3 - 5 + 6 + 8 = 75$$

$$\sum x^2 = 476 - 4 - 49 + 16 + 81 = 520$$

$$\sum y^2 = 521 - 9 - 25 + 36 + 64 = 587$$

$$\sum xy = 485 - 6 - 35 + 24 + 72 = 540$$

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

$$r = \frac{10(540) - (70)(75)}{\sqrt{[10(520) - (70)^2][10(587) - (75)^2]}}$$

$$r = 0.5533$$

⏭ 2.33 RANK CORRELATION

Sometimes data turns out to be non-numeric i.e. it is qualitative,

For example :

(1) Appearance : Beautiful, ugly

(2) Efficiency : Excellent, good, average, bad etc.

In such cases data is ranked according to that particular character and not according to numeric measurements on them.

And hence correlation coefficient cannot be calculated in the usual manner.

2.33.1 Spearman's Rank Correlation Coefficient

For a given set of n paired observations X_i for i = 1 ; ranks 1, 2, ... are assigned to the X-observations in order of magnitude and similarly to the Y-observations. Then these ranks are substituted for the actual numerical values, and correlation coefficient is calculated, which is called as 'Rank correlation coefficient' and is given by,

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

where d_i = difference between ranks assigned to X_i and Y_i

n = number of pairs of data.

For the sake of simplicity consider $\sum_{i=1}^{n} d_i^2 = \sum d^2$

Remarks

(i) r lies between – 1 and 1.

(ii) If there are ties among either X or Y observations, substitute for each of the tied observations, the mean of the ranks that they jointly occupy.

2.33.2 Tied Ranks

- If there is a tie between two or more individuals ranks, the rank is divided among equal individuals. e.g. if two items have fourth rank, the 4th and 5th rank is divided between them equally and is given as $\frac{4+5}{2} = 4.5^{th}$ rank to each of them.

If three items have the same 4th rank, each of them is given $\frac{4+5+6}{3} = 5^{th}$ rank.

If m is the number of items having equal ranks then the factor $\frac{1}{12}(m^3 - m)$ is added to $\sum d^2$.

If there are more than one cases of this type, this factor is added corresponding to each case. i.e.

$$r = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + ...\right]}{n(n^2 - 1)}$$

Ex. 2.33.1 : Ten students got the following percentage of marks in Mathematics and English:

Mathematics (x)	8	36	98	25	75	82	92	62	65	35
English (y)	84	51	91	60	68	62	86	58	35	49

Find the rank correlation coefficient.

☑ **Soln. :** Here, n = 10

x	y	Rank in Mathematics x	Rank in English y	d = x – y	d^2
8	84	10	3	7	49
36	51	7	8	– 1	1
98	91	1	1	0	0
25	60	9	6	3	9
75	68	4	4	0	0
82	62	3	5	– 2	4

x	y	Rank in Mathematics x	Rank in English y	d = x – y	d^2
92	86	2	2	0	0
62	58	6	7	– 1	1
65	35	5	10	– 5	25
35	49	8	9	– 1	1
				$\sum d = 0$	$\sum d^2 = 90$

The rank correlation coefficient is

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(90)}{10(10^2-1)} = 0.455$$

Ex. 2.33.2 : Ten competitors in a musical test were ranked by three judges A, B and C in the following order:

Rank by A	1	6	5	10	3	2	4	9	7	8
Rank by B	3	5	8	4	7	10	2	1	6	9
Rank by C	6	4	9	8	1	2	3	10	5	7

Using the rank correlation method, find which pair of judges has the nearest approach to common liking in music.

Unit II In Sem.

☑ **Soln. : Here, n = 10**

Rank by Ax	Rank by By	Rank by Cz	$d_1 = x - y$	$d_2 = y - z$	$d_3 = z - x$	d_1^2	d_2^2	d_3^2
1	3	6	– 2	– 3	5	4	9	25
6	5	4	1	1	– 2	1	1	4
5	8	9	– 3	– 1	4	9	1	16
10	4	8	6	– 4	– 2	36	16	4
3	7	1	– 4	6	– 2	16	36	4
2	10	2	– 8	8	0	64	64	0
4	2	3	2	– 1	– 1	4	1	1
9	1	10	8	– 9	1	64	81	1
7	6	5	1	1	– 2	1	1	4
8	9	7	– 1	2	– 1	1	4	1
			$\sum d_1 = 0$	$\sum d_2 = 0$	$\sum d_3 = 0$	$\sum d_1^2 = 200$	$\sum d_2^2 = 214$	$\sum d_3^2 = 60$

The rank correlation coefficient is

$$r_{x,y} = 1 - \frac{6\sum d_1^2}{n(n^2-1)} = 1 - \frac{6(200)}{10(10^2-1)} = -0.21 \; ; \qquad r_{y,z} = 1 - \frac{6\sum d_2^2}{n(n^2-1)} = 1 - \frac{6(214)}{10(10^2-1)} = -0.296$$

$$r_{z,x} = 1 - \frac{6\sum d_3^2}{n(n^2-1)} = 1 - \frac{6(60)}{10(10^2-1)} = 0.64$$

Here, $r_{z,x}$ is maximum, the pair of judges A and C has the nearest common approach.

UEx. 2.33.3 (20, 4 Marks)

The rank correlation coefficient from the following data:

x	10	12	18	18	15	40
y	12	18	25	25	50	25

☑ **Soln. :**

Here, n = 6

x	y	Rank x	Rank y	d = x – y	d^2
10	12	1	1	0	0
12	18	2	2	0	0
18	25	4.5	4	0.5	0.25
18	25	4.5	4	0.5	0.25
15	50	3	6	– 3	9
40	25	6	4	2	4
					$\sum d^2 = 13.5$

- Here, there are two items in the x series having equal values at the rank 4. Each is given the rank 4.5 $\left(\text{i.e. } \frac{4+5}{2} = 4.5 \text{ rank}\right)$
- Similarly, there are three items in the y series having equal values at the rank 3. Each of them is given the rank 4. $\left(\text{i.e. } \frac{3+4+5}{3} = 4 \text{ rank}\right)$

So, $m_1 = 2, m_2 = 3$

The rank correlation coefficient is

$$r = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[13.5 + \frac{1}{12}(8-2) + \frac{1}{12}(27-3) + \cdots.\right]}{6(6^2-1)}$$

$$= 0.5429$$

☞ **Example On Rank Correlation**

UEx. 2.33.4 **(19, 4 Marks)**

Determine rank correlation for the following data which shows the marks obtained in two quizes in mathematics:

Marks in 1st quiz (X) :	6	5	8	8	7	6	10	4	9	7
Marks in 2nd quiz (Y) :	8	7	7	10	5	8	10	6	8	6

☑ **Soln. :**

Assigning ranks to the data of X, we get

X :	4	5	6	6	7	7	8	8	9	10
Rank :	1	2	3	4	5	6	7	8	9	10
or :	1	2	3.5	3.5	5.5	5.5	7.5	7.5	9	10
Similarly Y :	5	6	6	7	7	8	8	8	10	10
Rank :	1	2	3	4	5	6	7	8	9	10
or :	1	2.5	2.5	4.5	4.5	7	7	7	9.5	9.5

☞ **Data assigned with ranks is**

X :	3.5	2	7.5	7.5	5.5	3.5	10	1	9	5.5
Y :	7	4.5	4.5	9.5	1	7	9.5	2.5	7	2.5
D :	– 3.5	–2.5	3	–2	4.5	–3.5	0.5	–1.5	2	3
D^2 :	12.25	6.25	9	4	20.25	12.25	0.25	2.25	4	9

$$\sum D^2 = 79.5$$

$$\text{Rank correlation} = 1 - \left[\frac{6\sum D^2}{n(n^2-1)}\right]$$

$$= 1 - \left[\frac{6(79.5)}{10(99)}\right] = 1 - 0.4818$$

$$\text{Rank correlation} = 0.5182$$

...Chapter Ends

❑❑❑

UNIT III

CHAPTER 3

Big Data Analytics Life Cycle

Syllabus Topics

Introduction to Big Data, sources of Big Data, Data Analytic Lifecycle: Introduction, Phase 1: Discovery, Phase 2: Data Preparation, Phase 3: Model Planning, Phase 4: Model Building, Phase 5: Communication results, Phase 6: Operation alize.

Case study : Global Innovation Social Network and Analysis (GINA).

3.1 DATA ANALYTIC LIFE CYCLE : OVERVIEW

UQ. Explain different phases of data analytics life cycle.
(SPPU – Q. 1(b), Aug. 18, 6 Marks)

UQ. Explain Data Analytic Life cycle.
(SPPU – Q. 1(a), Dec. 18, Q. 2(b), Oct. 19, 8 Marks)

UQ. Draw Data Analytics Lifecycle & give brief description about all phases.
(SPPU – Q. 1(b), May 19, 5 Marks)

- At this level we need to know more deep knowledge of specific roles and responsibilities of the data scientist.
- The data scientist lifecycle is illustrated in Fig. 3.1.1 which gives the high-level overview of the data scientist discovery and analysis process.
- It depicts the iterative behaviour of work performed by the data scientist's with several stages being repetitive in order to make sure that the data scientist is utilizing the "right" analytic model to locate the "right" insights.

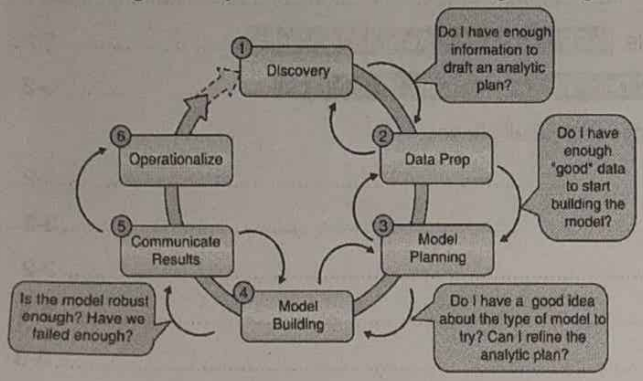


Fig. 3.1.1 : Data Scientist Lifecycle

3.1.1 Phase 1 - Discovery Phase

The following activities of data scientists can be focused by the Discovery :

- Acquisition of a complete understanding of the business process and the business domain. This consists of recognizing the key metrics and KPIs against which the business users will measure success.
- Recognizing the most vital business questions and business decisions that the business users are attempting to answer in support of the targeted business process. This also should contain the occurrence and optimal timeliness of those answers and decisions.
- Evaluating available resources and going through the process of framing the business problem as an analytic hypothesis. At this stage data scientist constructs the initial analytics development plan that will be used to direct and document the resulting analytic models and insights.
- It should be noticed that understanding into which production or operational environments the analytic insights requires to be published is somewhat that should be recognized in the analytics development plan.
- Such information will be essential as the data scientist recognizes in the plan where to "operationalize" the analytic insights and models.
- This is a best opportunity for tight association with the BI analyst who likely has already defined the metrics and processes required to support the business proposal.
- Requirements and the decision making environment of the business users can be well understand by the BI analyst to starts the data scientist's analytics development plan.

3.1.2 Phase 2 - Data Preparation

The following activities of data scientists can be focused by the data preparation :

- Provisioning an analytic workspace, or an analytic sandbox, where the data scientist can work free of the constraints of a production data warehouse environment. Preferably, the analytic environment is set up such that the data scientist can self-provision as much data space and analytic horsepower as required and can fine-tune those requirements throughout the analysis process.
- Obtaining, cleaning, aligning, and examining the data. This contains use of data visualization techniques and tools to get an understanding of the data, recognizing outliers in the data and calculating the gaps in the data to decide the overall data quality; determine if the data is "good enough."
- Transforming and enhancing the data. The data scientist will look to use analytic techniques, such as logarithmic and wavelet transformations, to sort out the potential skewing in the data. The data scientist will

also look to use data enhancement techniques to create new composite metrics such as frequency, recentness, and order. The data scientist will make use of standard tools like SQL and Java, as well as both commercial and open source extract, transform, load (ETL) tools to transform the data.

- After this stage is completed, the data scientist wants to feel comfortable enough with the quality and prosperity of the data to move ahead to the next stage of the analytics development process.

3.1.3 Phase 3 - Model Planning

The following activities of data scientists can be focused by the model planning :

- Determining the numerous analytical models, methods, techniques and workflows to discover as part of the analytic model development. The data scientists knows in advance that which of the analytic models and methods are suitable but it is good thing to plan to check at least one to make sure that the opportunity to build a more predictive model is not missed.
- Determine association and co-linearity between variables in order to select key variables to be used in the model development. The data scientist desires to estimate the cause-and-effect variables as early as possible. Keep in mind, association does not provides assurance causation, so care must be taken in choosing variables that can be calculated while going forward.

3.1.4 Phase 4 - Model Building

The following activities of data scientists can be focused by the model building :

- Manipulating the data sets for testing, training, and production. Whatever new transformation techniques are developed can be tested to observe if the quality, reliability, and predictive capabilities of the data can be enhanced or not.
- Calculating the feasibility and reliability of data to use in the predictive models. Decision calls are depends on quality and reliability of the data to check; is the data "good enough" to be used in developing the analytic models.
- At the end, developing, testing, and filtering the analytic models is done. Testing is carrying out to notice which variables and analytic models deliver the maximum quality, most predictive and actionable analytic insights.
- The model building stage is highly iterative step where manipulation of the data, calculating the reliability of the data, and determining the quality and predictive powers of the analytic model will be modified number of times.
- In this stage the data scientist may be unsuccessful many times in testing different variables and modelling techniques before resolved into the "right" one.

3.1.5 Phase 5 - Communicate Results

The following activities of data scientists can be focused by the communicate results :

- Determining the quality and reliability of the analytic model and statistical implication, ability of measuring and taking the action of the resulting analytic insights. The data scientist wants to make sure that the analytic process and model was successful and accomplishes the required analytic goal of the project.
- To communicate with the insights of analytic model, results and the suggestions requires the use of graphics and charts. It is significant that the business stakeholders such as business users, business analysts, and the BI analysts should realize and obtain the resulting analytic insights.
- The BI analysts are partner in this stage of the data science lifecycle. The BI analysts have the strong understanding of what to present to their business users and how to present it.

3.1.6 Phase 6 - Operationalize

The following activities of data scientists can be focused by the operationalize :

- Providing the final suggestions, reports, meetings, code, and technical documents.
- Optionally, running a pilot or analytic lab to validate the business case, and the financial return on investment (ROI) and the analytic lift.
- Carrying out the analytic models in the production and operational environments. This engross working with

the application and production teams to decide how best to surface the analytic results and insights.

- Combining the analytic scores into management dashboards and operational reporting systems, like sales systems, procurement systems, and financial systems etc.
- The operationalization stage is another area where association between the data scientist and the BI analysts should be very useful.
- Numerous BI analysts have the experience of combining reports and dashboards into the operational systems, as well as establishing centers of excellence to spread analytic learning and skills across the organization.

3.2 CASE STUDY - GINA : GLOBAL INNOVATION NETWORK AND ANALYSIS

UQ. Write a case study on Global Innovation Network & Analysis (GINA).
(SPPU – Q. 2(a), May 19, Q. 2(b), Dec. 19, 5 Marks)

GQ. Write a short note on Case of GINA. **(8 Marks)**

- EMC's GINA (Global Innovation Network and Analytics) team is a group of senior technologists placed in centers of excellence (COEs) all over the world.
- The main goal of team is to connect employees all over the world to drive innovation, research as well as university partnerships.
- The basic consideration of GINA team was that its approach would offer an interface to share ideas globally and enhance sharing of knowledge between GINA members who are not at one place geographically.
- A data repository has been created to store both structured and unstructured data to achieve three important goals :
  1. Store formal as well as informal data.
  2. Keep track of research from technologists all over the world.
  3. To enhance the operations and strategy, extract data for patterns and insights.
- The case study of GINA illustrates an example of the way by which a team applied the Data Analytics Lifecycle for the purpose of analyzing innovation data at EMC.
- Innovation is generally considered as a hard concept to measure, and this team is going to use advanced analytical methods so as to identify key innovators within the company.

3.2.1 Phase 1 - Discovery

- In this phase, identification of data sources is started by the team.
- Even though GINA has technologists which are skilled in several different aspects of engineering, it had few data and ideas regarding what it needs to explore but do not have a formal team which could perform these analytics.
- They consults with various experts and decided to outsource the work to the volunteers within EMC.
- The list of roles is as follows on the working team which were fulfilled :
  (i) **User of Business, Sponsor of Project, Manager of Project :** Vice President
  (ii) **Business Intelligence Analyst :** Representatives from IT Field
  (iii) **DBA (Data Engineer and Database Administrator)** : Representatives from IT
  (iv) **Data Scientist :** Distinguished Engineer who are able to develop social graphs.
- The approach of project sponsor is to influence social media and blogging for the purpose of accelerating the set of innovation as well as research data across the world and to inspire teams of data scientists who can work as "volunteer" globally.
- The data scientists should show passion about data, and the project sponsor should have ability to tap into this passion of greatly talented people to achieve challenging work in a creative way.
- The data regarding the project is divided into two important categories. The first category regards with the idea submissions of near about five years from EMC's internal innovation contests, called as the Innovation Roadmap or Innovation Showcase.

- The Innovation Roadmap is nothing but an organic innovation process in which ideas are submitted by employees globally which are then judged.
- For further incubation, rest out of these ideas are selected.
- Consequently the data is combination of structured data, like idea counts, submission dates, inventor names, and unstructured content, like the textual descriptions regarding the ideas themselves.
- The second category of data consists of encompassed minutes as well as notes which represents innovation and research activity globally
- Additionally it represents combination of structured and unstructured data. The structured data consists of attributes like dates, names as well as geographic locations.
- In the unstructured documents data is regarding "who, what, when, and where" which represents rich data regarding knowledge growth and transfer inside the company.
- There are 10 important IHs which are developed by GINA team :

1. **IH1 :** It is possible to map innovation activity in dissimilar geographic locations to corporate strategic directions.
2. **IH2 :** The delivery time of ideas minimizes by the transfer of global knowledge as part of the idea delivery process.
3. **IH3 :** Innovators participating in global knowledge are able to deliver ideas fast as compared to those who do not.
4. **IH4 :** It is possible to analyze and evaluate an idea submission for the likelihood of receiving funding.
5. **IH5 :** Knowledge invention and increase for a specific topic can be measured as well as compared across geographic locations.
6. **IH6 :** Research-specific boundary can be identified by the knowledge transfer activity spanners in different regions.
7. **IH7 :** It is possible to map strategic corporate themes to geographic locations.
8. **IH8 :** Continuous knowledge growth and transfer events minimize the time required to create a corporate asset from an idea.
9. **IH9 :** Lineage maps get revealed when corporate asset is not generated by the knowledge expansion and transfer.
10. **IH10 :** It is possible to classify and map emerging research topics to particular ideators, innovators, boundary spanners, and assets.

3.2.2 Phase 2 - Data Preparation

- A new analytics sandbox is set up by the team with its IT department for the purpose of storing and experimenting on the data.
- In the process of data exploration exercise, the data scientists and data engineers come to know that specific data require conditioning and normalization.
- Also they come to know that various missing datasets were difficult to testing some of the analytic hypotheses.
- As data is explored by the team, it promptly realized that without good quality data, it would not be able to carry out the subsequent steps in the lifecycle process.
- Consequently it was essential to conclude for project what level of data quality and cleanliness was necessary.
- In the case of the GINA, the team realizes that several of the names of the researchers and people who are communicating with the universities were misspelled or had spaces at leading and trailing side in the data-store.
- Such little problems must be addressed in this phase to enable better analysis as well as data aggregation in subsequent phases.

3.2.3 Phase 3 - Model Planning

- In the GINA project, for large amount of dataset, it looks viable to use social network analysis techniques to observe the networks regarding innovators.
- In other cases, it was hard to provide appropriate methods to test hypotheses because of the lack of data.
- In one case (IH9), a decision is made by the team to begin a longitudinal study to start tracking data points over time about people who are developing new intellectual property.

Unit III End Sem.

- This data collection support the team to test the next two ideas later :
  - (i) **IH8 :** Continuous knowledge growth and transfer events minimize the time required to create a corporate asset from an idea.
  - (ii) **IH9 :** Lineage maps get revealed when corporate asset is not generated by the knowledge expansion and transfer.
- For the longitudinal study being proposed, there is need to team to establish goal criteria for the purpose of study.
- Particularly, it required to decide the end goal of a successful idea which had traversed the entire journey. The parameters regarding the scope of the study consist of the following considerations:
  - (i) Identify the correct milestones for the purpose of accomplishing this goal.
  - (ii) Trace the way by which people shift ideas from each and every milestone towards the goal.
  - (iii) After this, trace ideas which unable to reach the goals, and trace others which are able to reach the goal. Compare the journeys of both types of ideas.
- Make comparison regarding the times and the outcomes with the help of a few different methods based on the way by which data is collected and assembled.

3.2.4 Phase 4 - Model Building

- In this phase several analytical methods are employed by GINA team.
- It contains the work by the data scientist through NLP (Natural Language Processing) techniques on the descriptions in textual format of the Innovation Roadmap ideas.
- Also social network analysis is conducted using R and RStudio, and then developed social graphs and visualizations of the network of communications regarding improvement through R's ggplot2 package.
- Examples of this work are shown in Fig. 3.2.1.
- Fig. 3.2.1 displays social graphs which depict the associations in between idea submitters inside GINA. Innovator from different countries are represented by dots. The large dots with circles around represent hubs.

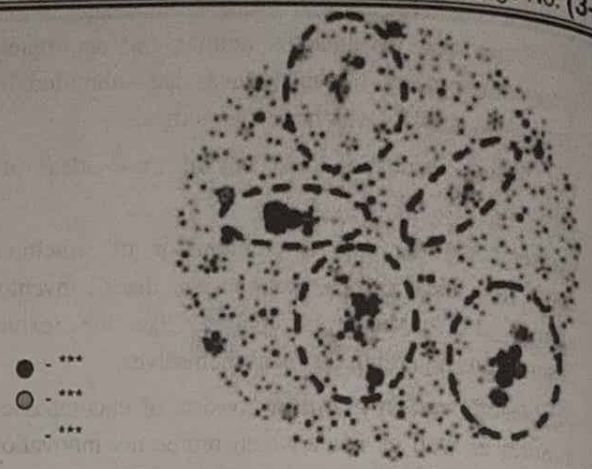


Fig. 3.2.1 : Social graph visualization of idea submitters and finalists

- A hub represents a person having great connectivity.
- The cluster in Fig. 3.2.2 consists of geographic variety, which is hard to show the hypothesis regarding geographic boundary spanners.
- In this graph, one person posses strangely high score when compared to the remaining nodes in the graph.
- This person is identified by the data scientists and they execute a query against his name within the analytic sandbox.

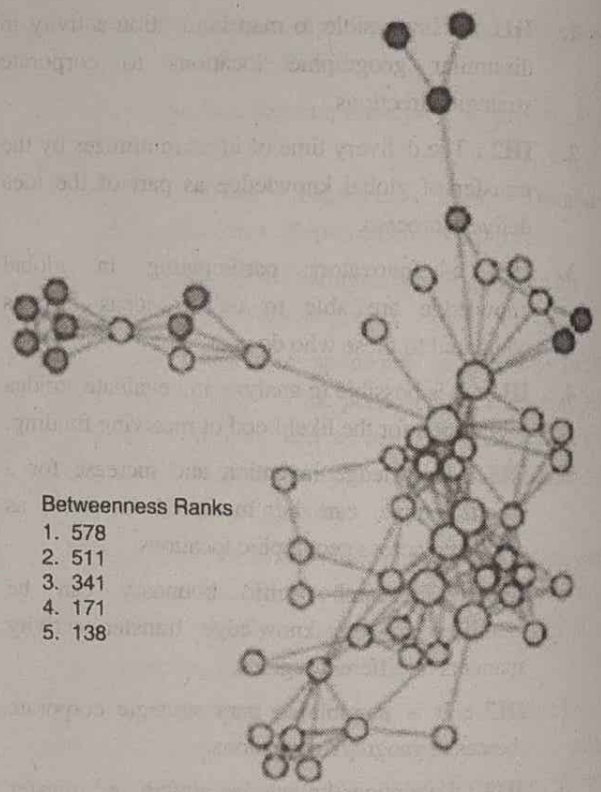


Fig. 3.2.2 : Social graph visualization of top innovation influencers

3.2.5 Phase 5 - Communicate Results

- In Communicate Results phase, the team got various methods to gather results of the analysis and identify the most effective and appropriate findings.
- This project seems to be doing well in the process of identifying boundary spanners and hidden innovators. Consequently, the CTO office establishes longitudinal studies to start data gathering efforts and keep track of innovation outputs for long duration of time.
- The GINA project inspires the concept of knowledge sharing regarding innovation and researchers located at various areas within and outside the company.
- One of the outputs of the project is that there was a strangely great density of innovators in Cork, Ireland.
- Every year, EMC hosts an innovation contest, which was open to all company employees to submit innovation ideas which can drive new value for the company.
- These findings were later on shared internally with the help of presentations and conferences and also promoted using social media and blogs.

3.2.6 Phase 6 - Operationalize

- Implementation of analytics against a sandbox which is basically filled with notes, minutes, and presentations from innovation activities results in high insights into EMC's innovation culture.
- Key findings from the project include :
  - (i) The CTO office and GINA require extra information in the future, containing a marketing initiative for the purpose of convincing people to inform the global community on their innovation/research activities.
  - (ii) Some of the data is comparatively very sensitive, and hence the team requires considering security and privacy regarding the data like who can run the models and see the results.
  - (iii) In addition to running models, there is need of a simultaneous initiative to enhance the basic Business Intelligence activities like dashboards, reporting, and queries on research activities globally.
  - (iv) There is necessity of a mechanism to continually for the purpose of reevaluating the model after deployment. Assessing the benefits is an important goal of this stage, as is defining a process to retrain the model as needed.
- In addition to the actions and findings given in Table 3.2.1, the team also shows how analytics can drive new insights in projects which are basically traditionally hard to measure and quantify.
- Fig. 3.2.1 illustrates an analytics plan for the GINA case study example :

Unit III End Sem.

Table 3.2.1 : Analytic Plan from the EMC GINA Project

Components of Analytic Plan	GINA Case Study
Discovery Business Problem Framed	Tracking the growth of global knowledge ensuring efficient knowledge transfer, and rapidly transforming it into corporate assets.
Data	Five years of innovation idea submissions and history; six months of textual notes from global innovation and research activities
Model Planning Analytic Technique	Social network analysis, social graphs, clustering, and regression Analysis
Result and Key Findings	A) Recognized hidden, high-value innovators and got methods to share their knowledge. B) Informed decisions regarding investment in university research projects. C) Generated tools to help submitters for the purpose of improving ideas with idea recommender systems.

Chapter End...

❑❑❑

UNIT IV

CHAPTER 4

Predictive Big Data Analytics with Python

Syllabus Topics

Introduction, Essential Python Libraries, Basic examples. Data Preprocessing : Removing Duplicates, Transformation of Data using function or mapping, replacing values, Handling Missing Data. Analytics Types : Predictive, Descriptive and Prescriptive.

Introduction to Scikit-learn, Installations, Dataset, mat plotlib, filling missing values, Regression and Classification using Scikit-learn.

Unit IV End Sem

4.1 INTRODUCTION OF PYTHON

GQ. What is Python?

Python is a popular programming language. It was created by Guido van Rossum, and released in 1991.

It is used for :

1. Web development (server-side),
2. Software development,
3. Mathematics,
4. System scripting.

What can Python do?

- Python can be used on a server to create web applications.
- Python can be used alongside software to create workflows.
- Python can connect to database systems. It can also read and modify files.
- Python can be used to handle big data and perform complex mathematics.
- Python can be used for rapid prototyping, or for production-ready software development.

4.1.1 Importance of Python

GQ. Explain importance of python? (2 Marks)

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
- Python has a simple syntax similar to the English language.
- Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.
- Python can be treated in a procedural way, an object-oriented way or a functional way.

Good to know

- The most recent major version of Python is Python 3, which we shall be using in this tutorial. However, Python 2, although not being updated with anything other than security updates, is still quite popular.
- In this tutorial Python will be written in a text editor. It is possible to write Python in an Integrated Development Environment, such as Thonny, Pycharm, Netbeans or Eclipse which are particularly useful when managing larger collections of Python files.

Python Syntax compared to other programming languages

- Python was designed for readability, and has some similarities to the English language with influence from mathematics.
- Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.
- Python relies on indentation, using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for this purpose.

4.1.2 Python Libraries for Data Processing and Modeling

1. Pandas

GQ. Write a short on Pandas? (2 Marks)

- **Pandas** is a free Python software library for data analysis and data handling. It was created as a community library project and initially released around 2008.
- Pandas provides various high-performance and easy-to-use data structures and operations for manipulating data in the form of numerical tables and time series. Pandas also has multiple tools for reading and writing data between in-memory data structures and different file formats.
- In short, it is perfect for quick and easy data manipulation, data aggregation, reading, and writing the data as well as data visualization.

- Pandas can also take in data from different types of files such as CSV, excel etc. or a SQL database and create a Python object known as a data frame. A data frame contains rows and columns and it can be used for data manipulation with operations such as join, merge, groupby, concatenate etc.

2. NumPy

- **NumPy** is a free Python software library for numerical computing on data that can be in the form of large arrays and multi-dimensional matrices. These multidimensional matrices are the main objects in NumPy where their dimensions are called axes and the number of axes is called a rank.
- NumPy also provides various tools to work with these arrays and high-level mathematical functions to manipulate this data with linear algebra, Fourier transforms, random number crunchings, etc.
- Some of the basic array operations that can be performed using NumPy include adding, slicing, multiplying, flattening, reshaping, and indexing the arrays. Other advanced functions include stacking the arrays, splitting them into sections, broadcasting arrays, etc.

3. SciPy

- **SciPy** is a free software library for scientific computing and technical computing on the data. It was created as a community library project and initially released around 2001. SciPy library is built on the NumPy array object and it is part of the NumPy stack which also includes other scientific computing libraries and tools such as Matplotlib, SymPy, pandas etc.
- This NumPy stack has users which also use comparable applications such as GNU Octave, MATLAB, GNU Octave, Scilab, etc. SciPy allows for various scientific computing tasks that handle data optimization, data integration, data interpolation, and data modification using linear algebra, Fourier transforms, random number generation, special functions, etc. Just like NumPy, the multidimensional matrices are the main objects in SciPy, which are provided by the NumPy module itself.

4. Scikit-learn

- **Scikit-learn** is a free software library for Machine Learning coding primarily in the Python programming language. It was initially developed as a Google Summer of Code project by David Cournapeau and originally released in June 2007. Scikit-learn is built on top of other Python libraries like NumPy, SciPy, Matplotlib, Pandas, etc. and so it provides full interoperability with these libraries.
- While Scikit-learn is written mainly in Python, it has also used Cython to write some core algorithms in order to improve performance.
- You can implement various Supervised and Unsupervised Machine learning models on Scikit-learn like Classification, Regression, Support Vector Machines, Random Forests, Nearest Neighbors, Naive Bayes, Decision Trees, Clustering, etc. with Scikit-learn.

5. TensorFlow

- **TensorFlow** is a free end-to-end open-source platform that has a wide variety of tools, libraries, and resources for Artificial Intelligence. It was developed by the Google Brain team and initially released on November 9, 2015.
- You can easily build and train Machine Learning models with high-level API's such as Keras using TensorFlow. It also provides multiple levels of abstraction so you can choose the option you need for your model.
- TensorFlow also allows you to deploy Machine Learning models anywhere such as the cloud, browser, or your own device. You should use TensorFlow Extended (TFX) if you want the full experience, TensorFlow Lite if you want usage on mobile devices, and TensorFlow.js if you want to train and deploy models in JavaScript environments.
- TensorFlow is available for Python and C APIs and also for C++, Java, JavaScript, Go, Swift, etc. but without an API backward compatibility guarantee
- Third-party packages are also available for MATLAB, C#, Julia, Scala, R, Rust, etc.

6. Keras

- **Keras** is a free and open-source neural-network library written in Python. It was primarily created by François Chollet, a Google engineer, and initially released on 27 March 2015. Keras was created to be user friendly, extensible, and modular while being supportive of experimentation in deep neural networks.
- Hence, it can be run on top of other libraries and languages like TensorFlow, Theano, Microsoft Cognitive Toolkit, R, etc. Keras has multiple tools that make it easier to work with different types of image and textual data for coding in deep neural networks.
- It also has various implementations of the building blocks for neural networks such as layers, optimizers, activation functions, objectives, etc. You can perform various actions using Keras such as creating custom function layers, writing functions with repeating code blocks that are multiple layers deep, etc.

4.1.3 Python Libraries for Data Visualization

GQ. Explain Different python Libraries? **(4 Marks)**

1. Matplotlib

- **Matplotlib** is a data visualization library and 2-D plotting library of Python It was initially released in 2003 and it is the most popular and widely-used plotting library in the Python community. It comes with an interactive environment across multiple platforms.
- Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers etc. It can be used to embed plots into applications using various GUI toolkits like Tkinter, GTK+, wxPython, Qt, etc.
- So you can use Matplotlib to create plots, bar charts, pie charts, histograms, scatterplots, error charts, power spectra, stemplots, and whatever other visualization charts you want! The Pyplot module also provides a MATLAB-like interface that is just as versatile and useful as MATLAB while being totally free and open source.

2. Seaborn

- **Seaborn** is a Python data visualization library that is based on Matplotlib and closely integrated with the numpy and pandas data structures. Seaborn has various dataset-oriented plotting functions that operate on data frames and arrays that have whole datasets within them.
- Then it internally performs the necessary statistical aggregation and mapping functions to create informative plots that the user desires. It is a high-level interface for creating beautiful and informative statistical graphics that are integral to exploring and understanding data.
- The Seaborn data graphics can include bar charts, pie charts, histograms, scatterplots, error charts, etc. Seaborn also has various tools for choosing color palettes that can reveal patterns in the data.

3. Plotly

- **Plotly** is a free open-source graphing library that can be used to form data visualizations. Plotly (plotly.py) is built on top of the Plotly JavaScript library (plotly.js) and can be used to create web-based data visualizations that can be displayed in Jupyter notebooks or web applications using Dash or saved as individual HTML files.
- Plotly provides more than 40 unique chart types like scatter plots, histograms, line charts, bar charts, pie charts, error bars, box plots, multiple axes, sparklines, dendrograms, 3-D charts, etc.
- Plotly also provides contour plots, which are not that common in other data visualization libraries. In addition to all this, Plotly can be used offline with no internet connection.

4. GGplot

- **Ggplot** is a Python data visualization library that is based on the implementation of ggplot2 which is created for the programming language R. Ggplot can create data visualizations such as bar charts, pie charts, histograms, scatterplots, error charts, etc. using high-level API. It also allows you to add different types of data visualization components or layers in a single visualization.

- Once ggplot has been told which variables to map to which aesthetics in the plot, it does the rest of the work so that the user can focus on interpreting the visualizations and take less time in creating them.
- But this also means that it is not possible to create highly customised graphics in ggplot. Ggplot is also deeply connected with pandas so it is best to keep the data in DataFrames.

4.2 DATA PREPROCESSING

GQ. Explain Data preprocessing? (4 Marks)

- **Data Preprocessing** is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.
- Therefore, certain steps are executed to convert the data into a small clean data set. This technique is performed before the execution of the **Iterative Analysis**. The set of steps is known as Data Preprocessing. It includes

4.2.1 Removing Duplicates

- Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric python packages. **Pandas** is one of those packages and makes importing and analyzing data much easier.
- An important part of Data analysis is analyzing Duplicate Values and removing them. Pandas **drop_duplicates()** method helps in removing duplicates from the data frame.
- **Syntax :** DataFrame.drop_duplicates(subset=None, keep='first', inplace=False)
- **Parameters :**
  - **Subset :** Subset takes a column or list of column label. It's default value is none. After passing columns, it will consider them only for duplicates.
  - **Keep :** keep is to control how to consider duplicate value. It has only three distinct value and default is 'first'.

4.2.2 Transformation of Data using Function or Mapping

- The Transform function in Pandas (Python) can be slightly difficult to understand, especially if you're coming from an Excel background. Honestly, most data scientists don't use it right off the bat in their learning journey.
- But Pandas' transform function is actually quite a handy tool to have as a data scientist! It is a powerful function that you can lean on for feature engineering in Python.
- *To learn the basics of Python and Pandas for data science, check out these popular courses:*
  - Python for Data Science
  - Pandas for Data Analysis in Python

1. What is the Transform Function in Python?
2. Why is the Transform Function Important?
3. Apply vs. Transform Function in Python

1. What is the Transform Function in Python?

- Python's Transform function returns a self-produced dataframe with transformed values after applying the function specified in its parameter. This dataframe has the same length as the passed dataframe.
- That was a lot to take in so let me break it down using an example.
- Let's say we want to multiply 10 to each element in a dataframe:

```
#import library
importpandasaspd
importnumpyasnp
view rawT1.py hosted with by GitHub
#creating a dataframe
df=pd.DataFrame(np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]]),
columns=['a', 'b', 'c'])
view rawT2.py hosted with by GitHub
```

The original dataframe looks like this:

a	b	c
1	2	3
4	5	6
7	8	9

Unit IV End Sem

```
#applying the transform function
df.transform(func=lambdax : x*10)
```

view rawT3.py hosted with by GitHub

This is the dataframe we get after applying Python's Transform function:

a	b	c
10	20	30
40	50	60
70	80	90

2. Why is Python's Transform Function Important?

- Transform comes in handy during feature extraction. As the name suggests, we extract new features from existing ones. Let's understand the importance of the transform function with the help of an example.
- Here, we have a dataset about a department store:

User_ID	Product_ID	Purchase
1001	P1	100
1001	P2	200
1001	P3	300
1001	P4	500
1002	P2	200
1003	P3	400
1004	P1	200
1004	P2	300
1004	P3	400
1004	P4	500
1005	P1	100
1005	P2	200
1005	P3	300
1005	P4	400
1005	P5	500

- We can see that each user has bought multiple products with different purchase amounts. We would like to know what is the mean purchase amount of each user. This helps us in creating a new feature for the model to understand the relationship better. This is the desired output:

User_ID	Product_ID	Purchase	User_Mean
1001	P1	100	275
1001	P2	200	275
1001	P3	300	275
1001	P4	500	275
1002	P2	200	200
1003	P3	400	400
1004	P1	200	350
1004	P2	300	350
1004	P3	400	350
1004	P4	500	300
1005	P1	100	300
1005	P2	200	300
1005	P3	300	300
1005	P4	400	300
1005	P5	500	300

There are multiple approaches to do this:

1. Using Groupby followed by merge()
2. Transform function approach

☞ Approach 1 : Using Groupby followed by merge()

- The first approach is using ***groupby*** to aggregate the data then merge this data back into the original dataframe using the merge() function. **Let's do it!**

Step 1 : Import the libraries and read the dataset

```
importpandasaspd
df=pd.read_csv("purchase.csv") #Can be any csv of your choice.
```

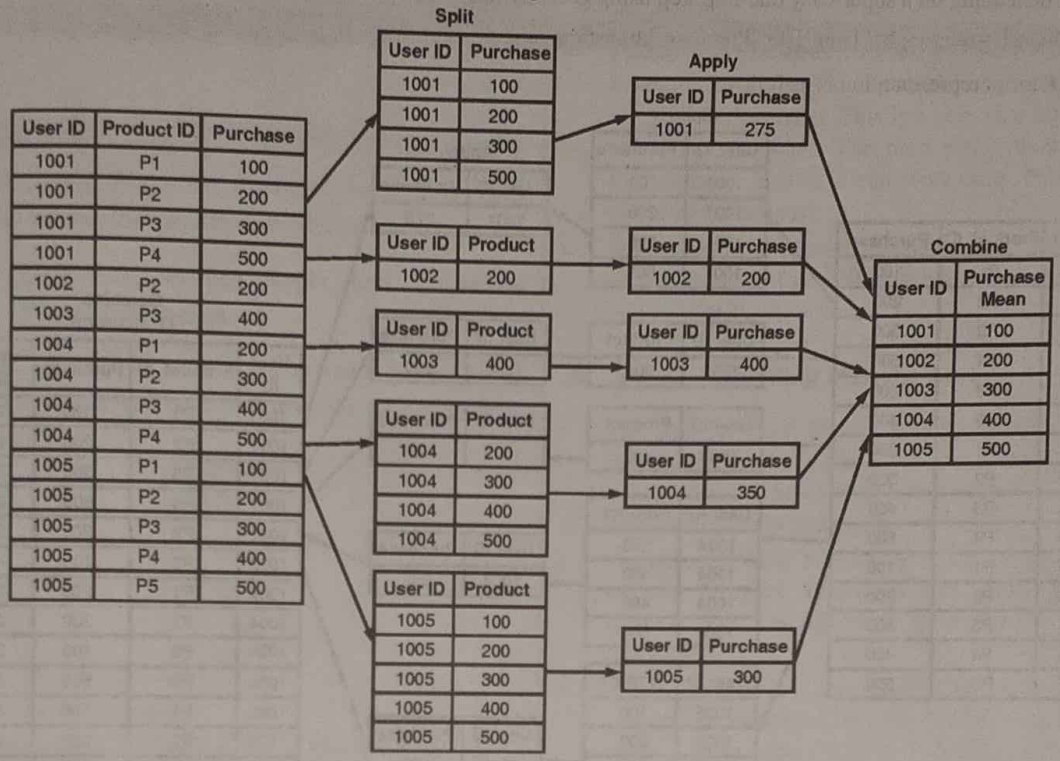
view rawT4.py hosted with by GitHub

Step 2 : Use groupby to calculate the aggregate

```
df.groupby('User_ID')["Purchase"].mean()
```

view rawT5.py hosted with by GitHub

Here is a pictorial representation of how *groupby* puts together the mean of each user



(1D1)Fig. 4.2.1

Unit IV End Sem.

Step 3 : Using merge() function to recombine

Now the tough part. How do we combine this data back to the original dataframe? We'll be using the merge() function for this task. You can read more about joins and merges in Python using Pandas here and here, respectively.

```
mean_purchase=df.groupby('User_ID')["Purchase"].mean().rename("User_mean").reset_index()
df_1=df.merge(mean_purchase)
```

view rawT6.py hosted with by GitHub

Our original dataframe looks like this:

User_ID	Product_ID	Purchase	User_Mean
1001	P1	100	275
1001	P2	200	275
1001	P3	300	275
1001	P4	500	275
1002	P2	200	200
1003	P3	400	400
1004	P1	200	350
1004	P2	300	350
1004	P3	400	350
1004	P4	500	300
1005	P1	100	300
1005	P2	200	300
1005	P3	300	300
1005	P4	400	300
1005	P5	500	300

- This certainly does our work. But it is a multistep process and requires extra code to get the data in the form we require. This multistep process can be resource-consuming in hackathons where time is a major constraint.

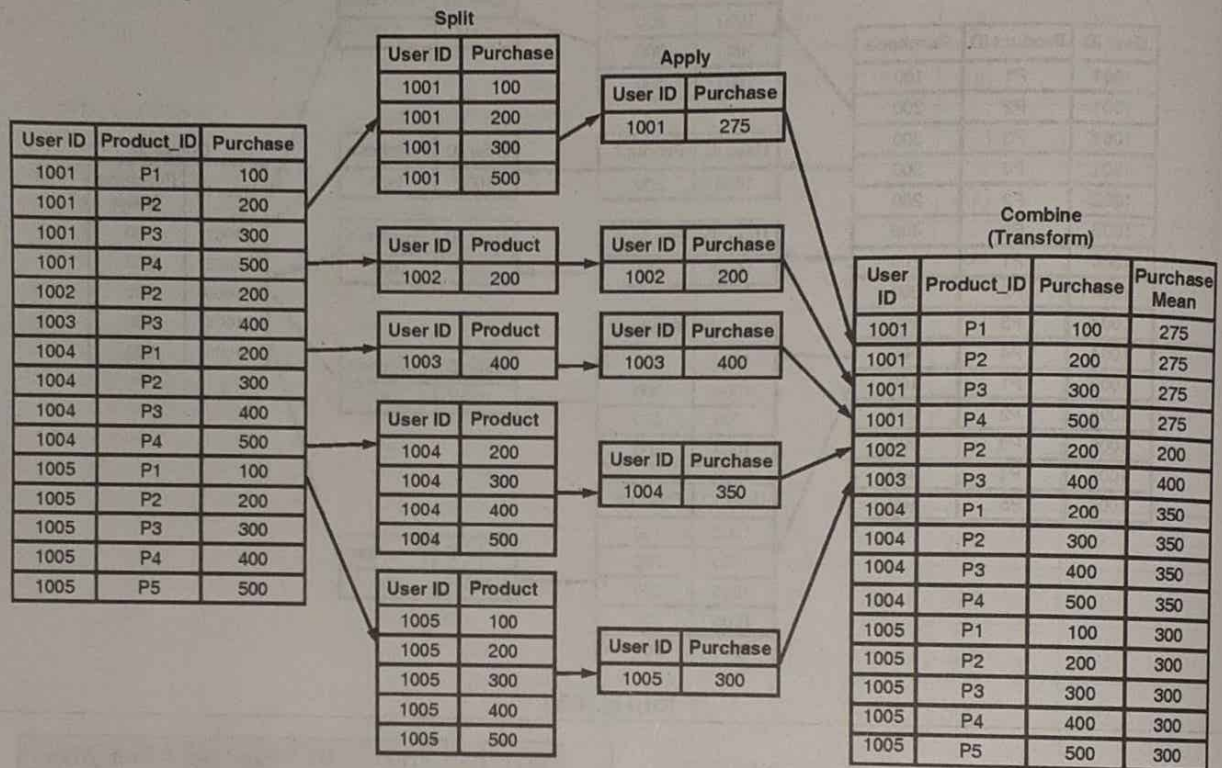
We can solve this effectively using the transform function in Pandas

☞ Approach 2 : Using Python's Transform Function

- This an important function for creating features. Trust me, it can be game-changer!
- The transform function retains the same number of items as the original dataset after performing the transformation. We'll be leaning on a super-easy one-line step using groupby followed by a transform:

```
df["User_Mean"] =df.groupby('User_ID')["Purchase"].transform('mean')
```

The pictorial representation is as follows:



(1D2)Fig.4.2.2

- Couldn't be simpler, right? The original dataframe looks similar to the above one in the last step.
- The time taken by the transform function to perform the above operation is comparatively less over a large dataframe. That's a significant advantage as compared to the first approach we used.
- Let me demonstrate the Transform function using Pandas in Python.
- Suppose we create a random dataset of 1,000,000 rows and 3 columns. Now we calculate the mean of one column based on groupby (similar to mean of all purchases based on groupbyuser_id).

▶ **Step 1 :** Import the libraries

```
#importing libraries
importpandasaspd
importrandom
```

view rawT8.py hosted with by GitHub

▶ **Step 2 :** Create the dataframe

```
data=pd.DataFrame({
'C' : [random.choice(('a','b','c')) foriinrange(1000000)],
'A' : [random.randint(1,10) foriinrange(1000000)],
'B' : [random.randint(1,10) foriinrange(1000000)]
})
```

▶ **Step 3 :** Use the merge procedure

```
%%timeit
data.groupby('C')["A"].mean()
mean=data.groupby('C')["A"].mean().rename("N").reset_index
()
df_1=data.merge(mean)
```

view rawT10.py hosted with by GitHub

Output

```
230 ms per loop
```

▶ **Step 4 :** Use the transform function

```
%%timeit

data['N3'] =data.groupby(['C'])['A'].transform('mean')
```

Output

```
35.1 ms per loop
```

This clearly shows the transform function is much faster than the previous approach. Well done!

▶ 3. Difference Between Apply And Transform Function in Python

Now, let's say we want to create a new column based on the values of another column. This is the dataframe we're working with:

a	b	c
1	2	3
4	5	6
7	8	9

With the apply function:

```
df['d'] =df.apply(lambdarow: row.a+row.b+row.c, axis=1)
```

This is what the output looks like using the Apply function:

a	b	c	d
1	2	3	6
4	5	6	15
7	8	9	24

The apply function sends a whole copy of the dataframe to work upon so we can manipulate all the rows or columns simultaneously.

☞ With the Transform function

This feature is not possible in the Transform function. This just manipulates a single row or column based on axis value and doesn't manipulate a whole dataframe. So, we can use either Apply or the Transform function depending on the requirement.

4.2.3 Replacing values in Python

- Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric python packages. ***Pandas*** is one of those packages and makes importing and analyzing data much easier.
- Pandas `dataframe.replace()` function is used to replace a string, regex, list, dictionary, series, number etc. from a dataframe. This is a very rich function as it has many variations. The most powerful thing about this function is that it can work with Python regex (regular expressions).
- **Syntax :** DataFrame.replace(to_replace=None, value=None, inplace=False, limit=None, regex=False, method='pad', axis=None)

4.2.4 Handling Missing Data

- The way in which Pandas handles missing values is constrained by its reliance on the NumPy package, which does not have a built-in notion of NA values for non-floating-point data types.
- Pandas could have followed R's lead in specifying bit patterns for each individual data type to indicate nullness, but this approach turns out to be rather unwieldy.
- While R contains four basic data types, NumPy supports *far* more than this: for example, while R has a single integer type, NumPy supports *fourteen* basic integer types once you account for available precisions, signedness, and endianness of the encoding.
- Reserving a specific bit pattern in all available NumPy types would lead to an unwieldy amount of overhead in special-casing various operations for various types, likely even requiring a new fork of the NumPy package.
- Further, for the smaller data types (such as 8-bit integers), sacrificing a bit to use as a mask will significantly reduce the range of values it can represent.
- NumPy does have support for masked arrays – that is, arrays that have a separate Boolean mask array attached for marking data as "good" or "bad." Pandas could have derived from this, but the overhead in both storage, computation, and code maintenance makes that an unattractive choice.
- With these constraints in mind, Pandas chose to use sentinels for missing data, and further chose to use two already-existing Python null values: the special floating-point `NaN` value, and the Python None object.

- This choice has some side effects, as we will see, but in practice ends up being a good compromise in most cases of interest.

4.2.5 None : Pythonic missing data

- The first sentinel value used by Pandas is None, a Python singleton object that is often used for missing data in Python code.
- Because it is a Python object, None cannot be used in any arbitrary NumPy/Pandas array, but only in arrays with data type 'object' (i.e., arrays of Python objects):

```
importnumpyasnp
importpandasaspd
vals1=np.array([1,None,3,4])
vals1
array([1, None, 3, 4], dtype=object)
```

- This dtype=object means that the best common type representation NumPy could infer for the contents of the array is that they are Python objects.
- While this kind of object array is useful for some purposes, any operations on the data will be done at the Python level, with much more overhead than the typically fast operations seen for arrays with native types:

```
fordtypein['object','int']:
print("dtype =",dtype)
%timeitnp.arange(1E6, dtype=dtype).sum()
print()
dtype = object
10 loops, best of 3: 78.2 ms per loop

dtype = int
100 loops, best of 3: 3.06 ms per loop
```

- The use of Python objects in an array also means that if you perform aggregations like sum() or min() across an array with a None value, you will generally get an error:

```
vals1.sum()

TypeError                 Traceback (most recent call last)
<ipython-input-4-749fd8ae6030> in <module>()
----> 1vals1.sum()

/Users/jakevdp/anaconda/lib/python3.5/site-
packages/numpy/core/_methods.py in _sum(a, axis, dtype, out,
keepdims)
  30
  31def          _sum(a,          axis=None,dtype=None,
out=None,keepdims=False):
---> 32returnumr_sum(a, axis,dtype, out,keepdims)
  33
  34def          _prod(a,          axis=None,dtype=None,
out=None,keepdims=False):

TypeError: unsupported operand type(s) for +: 'int' and
'NoneType'
```

- This reflects the fact that addition between an integer and None is undefined.

4.2.6 NaN: Missing numerical data

- The other missing data representation, NaN (acronym for *Not a Number*), is different; it is a special floating-point value recognized by all systems that use the standard IEEE floating-point representation:

```
In: vals2=np.array([1,np.nan,3,4])
vals2.dtype

out: dtype('float64')
```

- Notice that NumPy chose a native floating-point type for this array: this means that unlike the

4.3 ANALYTICS TYPES

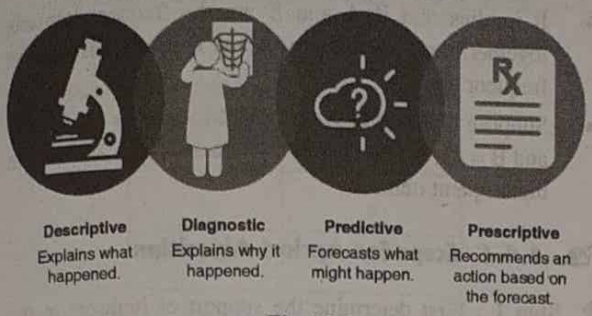
GQ. Explain Analytics Types? **(4 Marks)**

- The four types of analytics are usually implemented in stages and no one type of analytics is said to be better than the other. They are interrelated and each of these offers a different insight.
- With data being important to so many diverse sectors from manufacturing to energy grids, most of the companies rely on one or all of these types of analytics. With the right choice of analytical techniques, big data can deliver richer insights for the companies
- Before diving deeper into each of these, let's define the four types of analytics:

1) **Descriptive Analytics :** Describing or summarising the existing data using existing business intelligence tools to better understand what is going on or what has happened.
2) **Diagnostic Analytics :** Focus on past performance to determine what happened and why. The result of the analysis is often an analytic dashboard.
3) **Predictive Analytics :** Emphasizes on predicting the possible outcome using statistical models and machine learning techniques.

4) **Prescriptive Analytics :** It is a type of predictive analytics that is used to recommend one or more course of action on analyzing the data.

Let's understand these in a bit more depth.



(1D3)Fig.4.3.1

4.3.1 Descriptive Analytics

- This can be termed as the simplest form of analytics. The mighty size of big data is beyond human comprehension and the first stage hence involves crunching the data into understandable chunks. The purpose of this analytics type is just to summarise the findings and understand what is going on.
- Among some frequently used terms, what people call as advanced analytics or business intelligence is basically usage of descriptive statistics (arithmetic operations, mean, median, max, percentage, etc.) on existing data.
- It is said that 80% of business analytics mainly involves descriptions based on aggregations of past performance. It is an important step to make raw data understandable to investors, shareholders and managers.
- This way it gets easy to identify and address the areas of strengths and weaknesses such that it can help in strategizing.
- The two main techniques involved are data aggregation and data mining stating that this method is purely used for understanding the underlying behavior and not to make any estimations.
- By mining historical data, companies can analyze the consumer behaviors and engagements with their businesses that could be helpful in targeted marketing, service improvement, etc. The tools used in this phase are MS Excel, MATLAB, SPSS, STATA, etc.

4.3.2 Diagnostic Analytics

- Diagnostic analytics is used to determine why something happened in the past. It is characterized by techniques such as drill-down, data discovery, data mining and correlations.
- Diagnostic analytics takes a deeper look at data to understand the root causes of the events.
- It is helpful in determining what factors and events contributed to the outcome. It mostly uses probabilities, likelihoods, and the distribution of outcomes for the analysis.
- In a time series data of sales, diagnostic analytics would help you understand why the sales have decrease or increase for a specific year or so.
- However, this type of analytics has a limited ability to give actionable insights. It just provides an understanding of causal relationships and sequences while looking backward.
- A few techniques that uses diagnostic analytics include attribute importance, principle components analysis, sensitivity analysis, and conjoint analysis. Training algorithms for classification and regression also fall in this type of analytics

4.3.3 Predictive Analytics

- As mentioned above, predictive analytics is used to predict future outcomes. However, it is important to note that it cannot predict if an event will occur in the future; it merely forecasts what are the probabilities of the occurrence of the event.
- A predictive model builds on the preliminary descriptive analytics stage to derive the possibility of the outcomes.
- The essence of predictive analytics is to devise models such that the existing data is understood to extrapolate the future occurrence or simply, predict the future data.
- One of the common applications of predictive analytics is found in sentiment analysis where all the opinions posted on social media are collected and analyzed (existing text data) to predict the person's sentiment on a particular subject as being- positive, negative or neutral (future prediction).
- Hence, predictive analytics includes building and validation of models that provide accurate predictions. Predictive analytics relies on machine learning algorithms like random forests, SVM, etc. and statistics or learning and testing the data.
- Usually, companies need trained data scientists and machine learning experts for building these models. The most popular tools for predictive analytics include Python, R, RapidMiner, etc.
- The prediction of future data relies on the existing data as it cannot be obtained otherwise. If the model is properly tuned, it can be used to support complex

Unit IV End Sem.

forecasts in sales and marketing. It goes a step ahead of the standard BI in giving accurate predictions.

4.3.4 Prescriptive Analytics

- The basis of this analytics is predictive analytics but it goes beyond the three mentioned above to suggest the future solutions. It can suggest all favorable outcomes according to a specified course of action and also suggest various course of actions to get to a particular outcome.
- Hence, it uses a strong feedback system that constantly learns and updates the relationship between the action and the outcome.
- The computations include optimisation of some functions that are related to the desired outcome. For example, while calling for a cab online, the application uses GPS to connect you to the correct driver from among a number of drivers found nearby.
- Hence, it optimises the distance for faster arrival time. Recommendation engines also use prescriptive analytics.
- The other approach includes simulation where all the key performance areas are combined to design the correct solutions. It makes sure whether the key performance metrics are included in the solution.
- The optimisation model will further work on the impact of the previously made forecasts. Because of its power to suggest favorable solutions, prescriptive analytics is the final frontier of advanced analytics or data science, in today's term.

4.4 APRIORI ALGORITHM

UQ. Explain Apriori association rule mining algorithm.
(SPPU – Q. 2(c), Dec. 18, Q. 5(a), Oct. 19 7 Marks)

- The Apriori algorithm uses frequent item-sets to generate association rules, and it is designed to work on the databases that contain transactions. With the help of these association rule, it determines how strongly or how weakly two objects are connected.
- This algorithm uses a **breadth-first search** and **Hash tree** to calculate the itemset associations efficiently. It is the **iterative process** for finding the frequent itemsets from the large dataset.
- This algorithm is mainly used for market analysis and helps to find those products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

Frequent itemset

- Frequent itemsets are those items whose support is greater than the threshold value or user-specified minimum support.
- It implies that if A and B are the frequent itemsets together, then individually A and B should also be the frequent itemset.
- Suppose there are the two transactions : A = {2, 3, 4, 5} and B = {3, 4, 8}, in these two transactions, 3 and 4 are the frequent itemsets.

4.4.1 Steps for Apriori Algorithm

- **Step 1 :** First determine the support of itemsets in the transactional database, and select the minimum support and confidence.
- **Step 2 :** Consider all supports in the transactions with higher support value than the minimum or selected support value.
- **Step 3 :** Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.
- **Step 4 :** Sort the rules in the decreasing order of lift.

4.4.2 Working of Apriori Algorithm

- We consider an example to understand the apriori algorithm and mathematical calculation :

Example : Suppose we have the following dataset. It has various transactions. Now we find the frequent itemsets and generate the association rules using the Apriori algorithm.

TID	ITEMSETS
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

Given : Minimum support = 2, minimum confidence = 50 %

Step 1 : Calculating C1 and L1

(i) In the first step, we prepare a table that contains support count (The frequency of each item-set individually in the dataset) of each itemset in the given dataset.

This table is called the candidate set or C1

ITEMSETS	Support-count
A	6
B	7
C	5
D	2
E	1

(ii) Now we consider all the itemsets that have the greater support count. It gives us the table for the frequent itemset L1.

Since all the itemsets have greater or equal support count than the minimum support, except the E, so E itemset is to be removed.

Itemset	Support-count
A	6
B	7
C	5
D	2

Step 2 : Candidate generation C2 and L2 :

(i) Here, we generate C2 with the help of L1. In C2, we create the pair of the itemsets of L1 in the form of subsets.

(ii) After creating subsets, we again find the support count from the main transaction table of datasets, i.e., how many times these pairs have occurred together in the given dataset.

So we obtain the table for C2

Itemset	Support-count
{A, B}	4
{A, C}	4
{A, D}	1
{B, C}	4
{B, D}	2
{C, D}	0

(ii) Now we compare C2 support count with the minimum support count, and after comparing, the itemset with less support count will be eliminated from the table C2. It gives us the table for L2.

Itemset	Support-count
{A, B}	4
{A, C}	4
{B, C}	4
{B, D}	2

A, B, C, D

Step 3 : Candidate generation C3 and L3 :

(i) For C3, we will repeat the same two processes, but now we will form the C3 table with subsets of three itemsets together, and will calculate the support count from the dataset. It gives in Table 4.4.1

Table 4.4.1

Itemset	Support-count
{A, B, C}	2
{B, C, D}	1
{A, C, D}	0
{A, B, D}	0

(ii) Now we create L3 table. Here we note from the above C3 table that there is only one combination of itemset that has support count equal to the minimum support count.

Unit IV End Sem.

Hence, the L3 will have only one combination, i.e. {A, B, C}

▶ **Step 4 :** To find the association rules for the subsets.

To generate the association rules, we create a new table with the possible occurred combination {A, B, C}.

For all the rules, we calculate the **confidence** using formula sup (A ^ B) / A. After calculating the confidence value for all rules, we will exclude the rules that have less confidence that the minimum threshold (50 %).

We prepare the following **table** :

From the table, we observe that threshold minimum confidence is 50 %, so the first three rules A ^ B → C, B^ C → A, and A ^ C → B can be considered as the strong association rules for the given problem.

Rules	Support	Confidence
A ^ B → C	2	Sup { (A ^ B) ^ C } / Sup (A ^ B) = $\frac{2}{4}$ = 0.5 = 50 %
B ^ C → A	2	Sup { (B ^ C) ^ A } / Sup (B ^ C) = $\frac{2}{4}$ = 0.5 = 50 %
A ^ C → B	2	Sup { (A ^ C) ^ B } / Sup (A ^ C) = $\frac{2}{4}$ = 0.5 = 50 %
C → A ^ B	2	Sup { C ^ (A ^ B) } / Sup (C) = $\frac{2}{5}$ = 0.4 = 40 %
A → B ^ C	2	Sup { A ^ (B ^ C) } / Sup (A) = $\frac{2}{6}$ = 0.33 = 33.33 %
B → B ^ C	2	Sup { B ^ (B ^ C) } / Sup (B) = $\frac{2}{7}$ = 0.28 = 28 %

☞ Advantages of Apriori algorithm

(i) It is easy to understand the algorithm.

(ii) The join and prune steps of the algorithm can be easily implemented on large datasets.

☞ Disadvantages of Apriori algorithm

(i) The apriori algorithm works slow compared to other algorithms.

(ii) The overall performance can be reduced as it scans the databse for multiple times.

(iii) The time complexity and space complexity of the apriori algorithm is O (2^D), which is very high. Here D represents the horizontal width present in the database.

Remarks :

☞ Confidence level in Email A/B Testing

- Email A/B testing is generally used to select between two different variations of an email message so that the winning version can be sent to the broader population.
- We do the A/B testing with the two emails variations, and we get the following results :

Table 4.4.2 : A/B testing results

Variation	Number of Emails sent	Number of clicks	Conversion
A	5,000	100	2.0 %
B	6,000	150	2.5 %

We carry few calculations to find the confidence level statistic :

1. Conversion

Conversion, represented by P is calculated as :

P = Number of Click throughs / Number of Emails sent

∴ P (Variation - A) = 2.0 % = 0.02

P (Variation - B) = 2.5 % = 0.025

2. Standard error

Standard error (SE) represents the statistical accuracy of an estimate (in this case, the conversion rates that we have calculated).

Expressed mathematically, it is :

$$S.E. = SQRT\left[\frac{P(1-P)}{\text{Sample size}}\right];$$

SQRT = square root of

$$= \sqrt{\frac{P(1-P)}{\text{Sample size}}}$$

Here

$$\text{S.E. (Variation - A)} = \sqrt{\frac{0.02\,(0.98)}{5000}} = 0.00198;$$

$$\text{S.E. (Variation - B)} = \sqrt{\frac{0.025\,(0.975)}{6000}} = 0.0020$$

Now, we use these two formulae in another formula to get an important number.

3. Significance : the Z score

A statistic, referred to as the "Z score" helps to determine whether the conversions in the two variations are really different.

It is defined as

$$\text{Z score} = \frac{\{P(\text{variation B}) - P(\text{variation A})\}}{\sqrt{[\text{SE (variation A)}]^2 + [\text{SE (variation B)}]^2}}$$

$$= \frac{\{0.025 - 0.02\}}{\sqrt{(0.002)^2 + (0.00198)^2}} = 1.77$$

Note : The greater the Z score is, the more confident we are that the conversions we calculated in the two variables are actually different from each other.

4.4.3 Examples

Ex. 4.4.1 Mention the Apriori algorithm for finding frequent item sets based on the All Electronics transaction database D of the given Table. Transactional data for an All Electronics Branch

TID	List of item IDs
T 100	11, 12, 15
T 200	12, 14
T 300	12, 13
T 400	11, 12, 14
T 500	11, 13
T 600	12, 13
T 700	11, 13
T 800	11, 12, 13, 15
T 900	11, 12, 13

$L_1 \bowtie L_1$ is equivalent to $L_1 \times L_1$

☑ **Soln. :**

(1) **First iteration :** Each item is a member of the set of candidate 1-itemsets, C_1. The algorithm scans all the transactions to count the number of occurrences of each item.

(2) Let the minimum support count required be 2, that is min.sup = 2. (Note that here we are referring to absolute support because we are using a support count. The corresponding relative support is $\frac{2}{9} = 0.22 = 22\%$).

The set of frequent 1-itemsets, L_1, can then be determined. It consists of the candidate 1-itemsets satisfying minimum support.

(3) Now, we discover the set of frequent 2-itemsets L_2, we use the algorithm $L_2 \bowtie L_2$ to generate a candidate set of 2-itemsets, C_2. C_2 consists of $\left(\frac{|L_1|}{2}\right)$ 2-itemsets.

Observe that no candidates are removed from C_2 during the prune step because each subset of the candidates is also frequent.

(4) Next, the transactions in D are scanned and the support count of each candidate itemset in C_2 is accumulated as shown in the middle table of second row.

Scan D for count of each candidate →

Itemset	Sup. count
{11}	6
{12}	7
{13}	6
{14}	2
{15}	2

Compare candidate support count with minimum support count →

Itemset	Sup. count
{11}	6
{12}	7
{13}	6
{14}	2
{15}	2

Unit IV End Sem.

Generate C_2 candidates from L_1 →

C_2

Item set
{11,12}
{11,13}
{11,14}
{11,15}
{12,13}
{12,14}
{12,15}
{13,14}
{13,15}
{14,15}

Scan D for count of each candidate →

C_2

Item set	Sup. count
{11,12}	4
{11,13}	4
{11,14}	1
{12,15}	2
{12,13}	4
{12,14}	2
{12,15}	2
{13,14}	0
{13,15}	1
{14,15}	0

Compare candidate support count with minimum support count →

L_2

Item set	Sup. count
{11,12}	4
{11,13}	4
{11,15}	2
{12,13}	4
{12,14}	2
{12,15}	2

Generate C_3 candidates from L_2 →

C_3

Itemset
{11,12,13}
{11,12,15}

Scan D for count of each candidate →

C_3

Itemset	Sup. count
{11,12,13}	2
{11,12,15}	2

Compare candidate support count with minimum support count →

L_3

Item set	Sup. count
{11,12,13}	2
{11,14,15}	2

Fig. P.4.4.1 : Generation of the candidate itemsets and frequent itemsets, where the minimum support count is 2.

(5) The set of frequent 2-itemsets, L_2, is then found out, it consists of those candidate 2-itemsets in C_2 having minimum support.

(6) From the join step, we get

$C_3 = L_2 \bowtie L_2 = \{11,12,13\}, \{11,12,15\}, \{11,13,15\}, \{12,13,14\}, \{12,13,15\}, \{12,14,15\}.$

Since all subsets of a frequent itemset must also be frequent, based on the apriori property, we note that the four later candidates cannot be frequent.

We remove them from C_3.

$\therefore C_3 = \{11,12,13\}, \{11,12,15\}$ after pruning search strategy.

Ex. 4.4.2 Using apriori algorithm, generate frequent item sets (min.sup. >= 33.3%) for the following transaction database.

Transaction_ID	Item list
T_1	{A, B, D, K}
T_2	{A, B, C, D, E}
T_3	{A, B, C, E}
T_4	{B, D}
T_5	{A, C}
T_6	{B, D}

☑ **Soln. :** We carry out scan 1 and scan 2

▸ **Step 1 :**

T ID	Item list
T_1	{A, B, D, K}
T_2	{A, B, C, D, E }
T_3	{A, B, C, E}
T_4	{B, D}
T_5	{A, C}
T_6	{B, D}

C_1 → 1^{ST} Scan →

Item set	Sup.
A	4
B	5
C	3
D	4
E	2
K	1

L_1

Itemset	Supp
A	4
B	5
C	3
D	4

C_2

Itemset
{A, B}
{A, C}
{A, D}
{B, C}
{B, D}
{C, D}

2^{nd} Scan →

Itemset	Supp
{A, B}	3
{A, C}	3
{A, D}	2
{B, C}	2
{B, D}	4
{C, D}	1

L_2

Item set	Supp
A, B	3
A, C	3
B, D	4

C_3 →

Item set
A, B,C, D

Itemset	Supp
A, B, D	2

Ex. 4.4.3 A database has five transactions. Let the minimum support be 60%. Find all the frequent item sets using Apriori algorithm. Show each step.

TID	Items
1	Butter, milk
2	Butter, Dates, Balloon, eggs
3	Milk, dates, Balloon, Cake
4	Butter, milk, dates, balloon
5	Butter, milk, Dates, Cake

☑ **Soln. :**

We use absolute support, where duplicate values are counted only once per TID.

▸ **Step 1:**

Itemset	Support	% Support
Butter	4	80 %
Milk	4	80 %
Dates	4	80 %
Balloon	3	60 %
Eggs	1	20 %
Cake	2	40 %

Since the total number of TID is S, given minimum support is 60 %; i.e. equivalent to 3 : 5

It implies that itemsets with 1 or 2 support counts are to be discarded.

Itemset	Support	% Support
Butter	4	80 %
Milk	4	80 %
Dates	4	80 %
Balloon	3	60 %

▸ **Step 2 :**

We generate frequent 2-itemsets. We scan the database. Each combination is counted per TID

Item set	Support	Support %
{Butter, Milk}	3	60 %
{Butter, Dates}	3	60 %
{Butter, Balloon}	2	40 %
{Butter, Cake}	1	20 %
{Butter, eggs }	1	20 %
{Milk, dates}	3	60 %

Unit IV End Sem.

Item set	Support	Support %
{Milk, Balloon }	2	40 %
{Milk, cake}	2	40 %
{Dates, Balloon}	3	60 %
{Dates, eggs}	1	20 %
{Dates, Cake}	1	20 %
{Balloon, Cake}	1	20 %
{Balloon, Eggs}	1	20 %
{Eggs, Cake}	0	0

▶ **Step 3 :**

Results itemset

Item set	Support	Support %
{Butter, Milk}	3	60 %
{Data, Milk}	3	60 %
{Dates, Balloon}	3	60 %

▶ **Step 4 :**

To generate frequentset 3- itemset

Itemset	Support	Support %
{Butter, Milk dates}	2	40 %
{Milk, dates, balloons}	2	40 %

⏭ 4.5 FP GROWTH

4.5.1 Data Mining Process

- Association rules mining is an important technology in data-mining. FP - Growth (frequent pattern growth) algorithm is a classical algorithm in association rules mining.
- Data mining is a process to obtain potentially useful, previously unknown, and ultimately recognisable knowledge from the data.

4.5.2 FP - Growth Algorithm

FP - Growth Algorithm

(i) Compresses data sets to a FP - tree,

(ii) Scans the database twice,

(iii) Does not produce the candidate itemsets in mining process, and

(iv) Greatly improves the mining efficiency.

- FP - Growth algorithm compresses the database into a frequent pattern tree and still maintains the information of associations between item sets.
- Then the compressed database is divided into a set of condition database (it is a special type of projection database).
- Each condition database is dug, and associates with a frequent item. Transaction database is in Table 4.5.1 (support count is 2); mining process using FP - Growth.

Table 4.5.1: Transaction database

TID	Items
001	A, B, C, D, E
002	B, C, E
003	C, E, D
004	A, C, D

- Scanning the database for the first time, we can obtain a set of frequent items and their support count.
- The collection of frequent items is ordered by decreasing sequence of support count. The result set or list writes for L. In this way, we have L = [C:4, D:3, E:3, A:2, B:2]

4.5.3 Building FP - tree

- First, the algorithm creates the root node of the tree, with the tag "null". Then it scans the database for the second time.
- Each item in a transaction is ordered by the sequence of L.
- Later it creates a branch for each transaction.

  For example, the first transaction
- "001 : A, B, C, D, E" contains five items {C, D, E, A, B} according to the sequence of L, generating the first branch. < (C:1), (D:1), (E:1), (A:1), (B:1) > for building FP-tree.

- The branch has five nodes. In it, C is the children link of root, D links to C, E links to D, A links to E, and B links to A.
- The second transaction "002: B, C, E" contains three items {C, E, B} according to the sequence L and it generates a branch.
- In it, C links to the root, E links to C, and B links to E.
- This branch shares the prefix <C> with the existing path of transaction "001".
- In this way, the algorithm makes the count of node C increase by 1 and creates two new nodes < (E:1), (B:1) > as a link of (C:2).
- Generally, the algorithm considers increasing a branch for a transaction and when each node follows common prefix, its count increases by 1; algorithm creates node for the item following the prefix and linking.
- The algorithm creates an item header table, for convenience of tree traversed.
- Each item through a node link points to itself in FP-tree.
- After scanning all transactions, we get the FP-tree displayed in Fig. 4.5.1

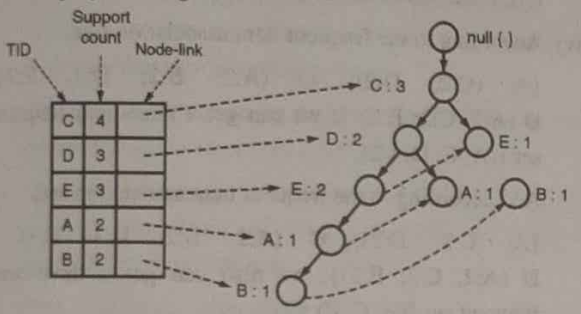


Fig. 4.5.1: Generating FP-tree

4.5.4 FP-tree mining processing

- The algorithm starts by the frequent patterns length of 1 (initial suffix pattern) and builds its conditional pattern base (a "subdatabase", consisting of prefix path set which appears with the suffix pattern).
- Then, algorithm builds a (conditional) FP-tree for the conditional pattern base and recursively digs the tree.
- The achievement of pattern growth gets through the link between frequent patterns generated by conditional FP-tree and suffix pattern. The mining of FP-tree is summarized in the Table 4.5.2.

Table 4.5.2 : Dig FP-tree through creating conditional subpattern base

Item	Conditional pattern base	Conditional FP-tree	Frequent pattern
B	{ (CDEA : 1), (CE : 1) }	<(C : 2, E:2)>	CB : 2 EB : 2 CE B:2
A	{ (CDE :1), (CD:1) }	<C:2, D:2>	CA : 2, D A : 2, CD A : 2
E	{(C,D:2) (C:1)}	<C:2, D:2>	CE : 2, D E:2, CD E:2
D	{(C:2)}	<C:2>	CD:2

System model

- Algorithms of frequent patterns mining have been applied in many fields. Researching their system models can facilitate a better understanding.
- Fig. 4.5.2 is a system model of the improved algorithms.

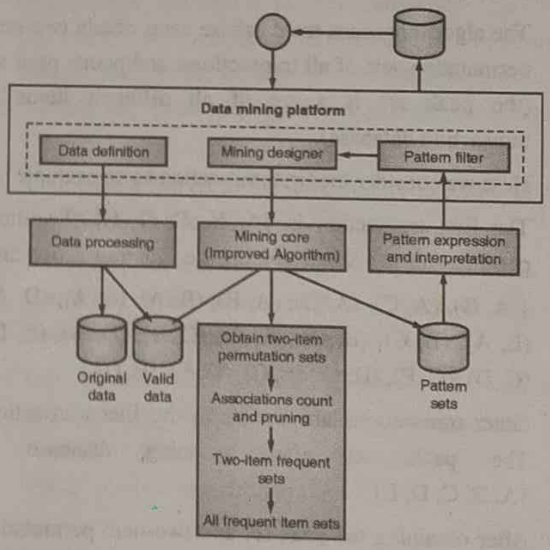


Fig. 4.5.2 : Associating rules mining system model

Unit IV End Sem.

- The user can get needed knowledge which passes data mining through the data-mining platform.
- Data mining platform includes data definition, mining designer and pattern filter.
- Through the data definition, we can do a pretreatment for data and make incomplete data usable; through the mining designer.
- We can use the improved algorithm to dig data and get useful patterns (here are given frequent item sets); through the pattern filter, we can select interesting patterns from obtained patterns.

☞ Improved algorithms based on the FP-Growth algorithm

- FP-Growth algorithm requires scanning database twice. Its algorithm efficiency is not high. So we put forward two improved algorithms :
  (I) Painting-Growth algorithm and
  (II) N Painting-Growth algorithm
- Which use two-item permutation sets to dig. Both algorithms scan database only once to obtain the results of mining.

(I) Painting-Growth algorithm and

We take the transaction database in table as an example, the mining process with Painting-Growth algorithm is as follows :

(i) The algorithm scans the database once, obtain two-item permutation sets of all transactions, and paints peak set (the peak set is a set of all different items in transaction database).

Here we consider the first transaction as an example.

The first transaction is {A, B, C, D, E}. Two-item permutation sets after scanning the first transaction are

{(A, B), (A, C), (A, D), (A, E), (B, A), (C, A), (D, A), (E, A), (B, C), (B, D), (B, E), (C, B), (D, B), (E, B), (C, D), (C, E), (D, C), (E, C), (D, E), (E, D) }

Other transactions are similar to the first transaction. The peak set after scanning database is {A, B, C, D, E}.

(ii) After obtaining the peak set and two-item permutation sets of all transactions, the algorithm paints the association picture according to two-item permutation sets and peak set.

It links the two items appearing in each two-item permutation. When the permutation appear again, the link count increases by 1. The association picture is as shown in the Fig. 4.5.3

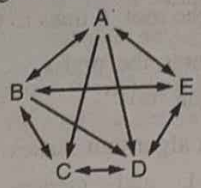


Fig. 4.5.3: The association picture

(iii) According to the association picture, algorithm exploits the support count to remove unfrequented associations. We can get the frequent item association sets as follows :

{A (C:2, D:2); B (C:2, E:2); C (A:2, B:2, D:3, E:3); D (A:2, C:3, E:2); E (B:2, C:3, D:2)}

Here we consider the item A as an example, A (C:2, D:2) shows that the support count of two-item set (AC) is 2 and the support count of two item set (AD) is 2. Other items are similar to item A.

(iv) According to the frequent item association sets,

{A (C:2, D:2); C (A:2, B:2; D:3, E:3}; D (A:2, C:3; E:2) }; we can get a three-item frequent set {(A, C, D) : 2}.

And according to the frequent item association sets

{A (C:2, D:2); C (A:2, B:2; D:3, E:3) ; D (A:2, C:3; E:2)}; we also can get a three-item frequent set {(B, C, E) : 2}.

Similarly, according to the frequent item association sets

{C (A:2, B:2, D:3, E:3); D (A:2, C:3; E:2) ; E (B:2, C:3; D:2) }; we get a three-item frequent set {(C, D, E) : 2}.

At this point, we get all frequent item sets.

(II) N Painting-Growth algorithm

N Painting-Growth algorithm is similar to the painting-Growth algorithm, but with different implementation method.

N Painting-Growth algorithm removes the painting steps. The mining process of N painting-growth is as follows :

(i) The algorithm scans the database once and gets two-item permutation sets of all transactions.

(ii) Then, the algorithm counts each permutation in two-item permutation sets getting all item association sets.

(iii) Later, the algorithm removes infrequent associations according to the support count and gets frequent item association sets.

(iv) Finally, it gets all frequent item sets according to the frequent item association sets,

Mining ends.

Thus we observe that from the above processes it can be seen that the N painting-growth algorithm is the removing of painting steps version of painting-Growth N. Painting-Growth algorithm only passes instantiation of a class in main function to implement.

☞ Experimental results analysis

- To improved algorithms-painting-growth and N painting-Growth algorithm-the biggest advantage is reducing database scanning to once.
- Comparing with scanning database twice of FP-Growth algorithm, it has improved time efficiency.
- Another advantage is that improved algorithms are simple, completing all mining-needing transactions, two-item permutation sets.
- But improved algorithms have disadvantages. In painting-Growth algorithm, the algorithm needs to build the association picture, leading to a large-memory overhead.
- In N Painting-Growth algorithm, the implementation method is less vivid than painting growth algorithm.
- When using the two improved algorithms to dig multi-item frequent sets, they scan the frequent item association sets repeatedly for count. This reduces the time efficiency.
- In large data, the performance of the N Painting-Growth is disappointing.

4.5.5 Example on FP growth

Ex. 4.5.1 : Using frequent pattern growth approach, we examine the transaction database D of the table.

Item_ID	Support Count
12	7
11	6
13	6
14	2
15	2

☑ **Soln.**

(i) The first scan of the database is the same as Apriori, which derives the set of frequent items (1 - itemsets) and their support counts (frequencies). Let the minimum count be 2.

(ii) The set of frequent items is sorted in the order of descending support count

This resulting set is denoted by

L = {{12 : 7}, {11 : 6}, {13 : 6}, {14 : 2}{15 : 2}}

(iii) We construct an FP-tree

We create the root of the tree and label it with "null"

(iv) Scan database D a second time. The items in each transaction are sorted in descending support count, and processed in L-order

(v) For example, the scan of the first transaction, " T 100 : 11, 12, 15", which contains three items (12, 11, 15 in L order), leads to the construction of the first branch of the tree with three nodes, < 12 : 1 >, < 11 : 1 >, and < 15 : 1 >,

Where 12 is linked to the root, and 11 is linked to 12, and 15 is linked to 11.

(vi) The second transaction, T200, contains the items 12 and 14 in L order, which results in a branch where 12 is linked to the root and 14 is linked to 12.

(vii) This branch shares a common prefix, 12, with the existing path for T100. Therefore, we increment the count of 12 node by 1, and create a new node (14 : 1), which is linked as a child to (12 : 2)

Unit IV End Sem.

(viii) In general, when considering the branch to be added for a transaction, the cont of each node along a common prefix is incremented by 1.

(xi) To create free traversal, an item header table is created so that each item points to its occurrences in the tree. Via a chain of node-links. We exhibit this in the diagram. This way, we transform the problem of mining frequent patterns in data bases into that of mining the FP-tree.

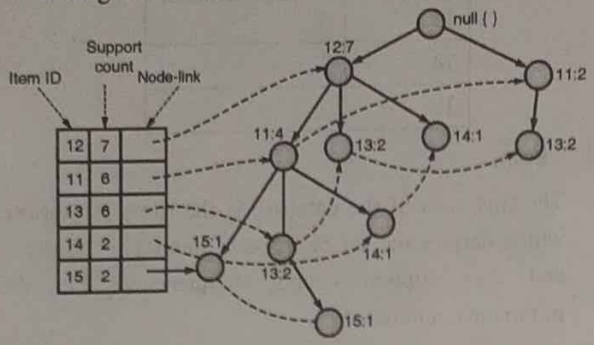


Fig. P.4.5.1 : An FP-tree registers compressed, frequent pattern information

Method of mining FP-tree

(i) Start from each frequent length-1 pattern (as an initial suffix pattern), construct its conditional pattern base;

(ii) A "sub-database" which consists of the set of prefix paths in the FP-tree with the suffix pattern, then construct its conditional FP-tree and

(iii) perform mining recursively on the free.

The pattern growth is achieved with the frequent patterns generated from a conditional FP-tree.

Item	Conditional pattern base	Conditional FP- tree	Frequent patterns Generated
15	{{12, 11 : 1}, {12, 11, 13 : 1}}	< 12 : 2, 11 : 2 >	{12, 15 : 2}, {11, 15 : 2}, {12, 11, 15 : 2}
14	{{12, 11 : 1}, {12 : 11}}	< 12 : 2 >	{12, 14 : 2}
13	{{12, 11 : 2}, {12 : 2}, {11, 2}}	< 12 : 4, 11: 2 >, < 11 : 2 >	{12, 13 : 4}, {11, 13, : 4}, {12, 11, 13 : 2}
11	{{12 : 4}}	< 12 : 4 >	{12, 11 : 4}

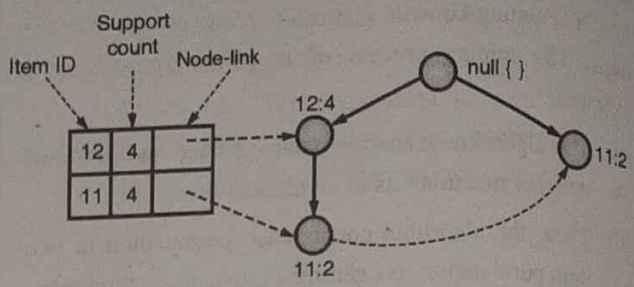


Fig. P.4.5.1(a)

The conditional FP-tree associated with the conditional node 13.

4.6 REGRESSION

UQ. What is regression ? Explain any one type of regression in detail. **(SPPU – Q. 6(b), Aug. 18, 4 Marks)**

- Regression is defined as a method of estimating the value of one variable when that of the other is known and the variables are correlated.
- Regression analysis is used to predict or estimate one variable in terms of the other variable.
- It is useful in statistical estimation of demand curves, supply curves, production function, cost function etc.

4.6.1 Type of Regression

UQ. Explain linear regression with example. **(SPPU – Q. 1(c), Dec. 18, Q. 5(b), Oct. 19, 6 Marks)**

Regression is classified into two types:

1. Simple regression and multiple regression
2. Linear regression and nonlinear regression

1. Simple and multiple regression

- **Simple regression :** The regression analysis for studying only two variables at a time is known as simple regression.
- **Multiple regression :** The regression analysis for studying more than two variables at a time is known as multiple regression.

2. Linear and nonlinear regression

- **Linear regression :** If the regression curve is straight line, then the regression is said to be linear.
- **Nonlinear regression :** If the regression curve is not a straight line i.e. not a first-degree equation in the variables x and y, the regression is said to be nonlinear regression.

4.6.2 Lines of Regression

Definition

1. Line of regression of y on x is the line which gives the best estimate for the value of y for any specified value of x.

We have already seen that the equation of line of regression of y on x is

$$y - \bar{y} = r\left(\frac{\sigma y}{\sigma x}\right)(x - \bar{x}) \quad \ldots(i)$$

2. The line of regression of x on y is the line which gives the best estimate of x for any given value of y. And the equation of line of regression of x on y is

$$x - \bar{x} = r\left(\frac{\sigma x}{\sigma y}\right)(y - \bar{y}) \quad \ldots(ii)$$

Remarks

1. Equation (i) implies that the line of regression of y on x passes through the point $(\bar{x}, \bar{y})$.

Similarly Equation (ii) implies that the line of regression of x on y passes through the point $(\bar{x}, \bar{y})$.

Hence both the lines pass through the point $(\bar{x}, \bar{y})$. In other words, the mean values $(\bar{x}, \bar{y})$ can be obtained as the point of intersection of the two regression lines.

2. Why two lines of regression ?

- There are always two lines of regression on of y on x and the other of x on y.
- The line of regression of y on x is used to estimate or predict the value of y for any given value of x, i.e., when y is dependent variable and x is independent variable. The estimate so obtained will be best in the sense that it will have the minimum possible error as defined by the principle of least squares.
- We can also obtain an estimate x for any given value of y by using Equation (i) but the estimate so obtained will not be best since Equation (i) is obtained on minimising the sum of squares of errors of estimates in y and not in x.
- Hence to estimate or predict x for any given value of y, we use the regression equation of x on y i.e. Equation (ii), which is derived on minimising the sum of squares of errors of estimates in x. Here x is dependent variable and y is an independent variable.
- The two regression equations are not reversible because the basis and assumptions for deriving these equations are quite different.
- The regression equation of y on x is obtained or minimising the sum of square of the errors parallel to Y-axis while the regression equation of x on y is obtained on minimising the sum of squares of the errors parallel to X-axis.

3. In case of perfect correlation, i.e., $r = \pm 1$ (positive r negative), the equation of line of regression of y on x becomes

$$y - \bar{y} = \pm\frac{\sigma y}{\sigma x}(x - \bar{x})$$

$$\text{i.e. } \frac{y - \bar{y}}{\sigma y} = \pm\frac{(x - \bar{x})}{\sigma x} \quad \ldots(iii)$$

Similarly the equation of the line of regression of x on y becomes

$$x - \bar{x} = \pm\frac{\sigma x}{\sigma y}(y - \bar{y})$$

$$\text{i.e. } \frac{x - \bar{x}}{\sigma x} = \pm\frac{(y - \bar{y})}{\sigma y} \quad \ldots(iv)$$

Thus (iii) and (iv) are identical

Hence, in case of perfect correlation ($r = \pm 1$), both the lines of regression coincide and we get only one line.

4.6.3 Using Regression Lines for Prediction

The equation of the regression line is commonly used to predict the value for the dependent variable Y for a given value of the independent variable X.

For example the predicted value of Y, written as $\hat{y}_i$ when X_i given by

$$\hat{y}_i = a + b X_i$$

Unit IV End Sem.

Where a and b are least squares estimates given by the normal equations.

The regression equation should be used for prediction with utmost care.

Before using the lines of regression, we should test for 'goodness of fit'.

Following points to be noted while using equations of lines of regression :

(i) If the value of 'r' is significant, we can use lines of regression for estimation and prediction.

(ii) If 'r' is not significant then the linear model is not a good fit and lines of regression equations should not be used.

(iii) Even if 'r' is significant, we should not use the linear regression model to make prediction for Y corresponding to far distant values of X.

4.7 COEFFICIENT OF REGRESSION

- Let us consider the line of regression of y on x :

$$y = a + bx$$

- The coefficient 'b' which is the slope of the line of y on x is called 'coefficient of regression of y on x'. For convenience the slope b, i.e. coefficient of regression of y o7n x is written as 'by x'.
- Similarly in the regression equation of x on y.

$$x = A + By;$$

- Again the coefficient 'b' is called 'coefficient of regression of x on y. For convenience it is written as 'b_{xy}'.
- Thus : Notation :

  b_{yx} = coefficient of regression of y on x and

  b_{xy} = coefficient of regression of x on y.
- Now, the coefficient of regression of y on x is given by

$$b_{yx} = \frac{\text{cov}(x, y)}{\sigma_x^2} = r\frac{\sigma y}{\sigma x}$$

$[\because \text{cov}(x, y) = r\sigma x \cdot \sigma y]$ ...(i)

- Similarly, the coefficient of regression of x on y is given by

$$b_{xy} = \frac{\text{cov}(x, y)}{\sigma_y^2} = r\frac{\sigma x}{\sigma y} \quad ...(ii)$$

- Hence, the equation of line of regression of y on x is given by

$$y - \bar{y} = b_{xy}(x - \bar{x}) \quad ...(iii)$$

and the equation of the line of regression of x on y become

$$x - \bar{x} = b_{xy}(y - \bar{y}) \quad ...(iv)$$

☞ Remarks

1. We develop the formulae for regression coefficients b_{yx} and b_{xy} :

We have $\text{cov}(x, y) = \frac{1}{n}\sum(x - \bar{x})(y - \bar{y})$

$$= \frac{1}{n}\sum xy - \bar{x}\cdot\bar{y}$$

(on simplification)

$$= \frac{1}{n^2}[n\sum xy - (\sum x)(\sum y)]$$

and $\sigma_x^2 = \frac{1}{n}\sum(x - \bar{x})^2 = \frac{1}{n}\sum x^2 - \left(\frac{\sum x}{n}\right)^2$

$$= \frac{1}{n^2}\left[n\sum x^2 - \left(\frac{\sum x}{n}\right)^2\right]$$

$$\therefore b_{yx} = \frac{\text{cov}(x, y)}{\sigma_x^2} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

$$= \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \quad ...(v)$$

Similarly, $\sigma_y^2 = \frac{1}{n}\sum(y - \bar{y})^2$

$$= \frac{1}{n^2}[n\sum y^2 - (\sum x)^2]$$

$$\therefore b_{xy} = \frac{\text{cov}(x, y)}{\sigma_y^2}$$

$$\therefore b_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum y^2 - (\sum y)^2} \quad ...(vi)$$

Formula (v) and (vi) can be used conveniently to find regression coefficient.

Also, we can use :

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \quad \text{and} \quad b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad ...(vii)$$

2. Correlation coefficient between two variables x and y is a symmetrical function between x and y, i.e. $r_{yx} = r_{xy}$.

But the regression coefficients are not symmetric functions of x and y;

i.e. $b_{xy} \neq b_{yx}$

$\because b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$;

While $r_{xy} = \frac{cov(x, y)}{\sigma_x \cdot \sigma_y} = \frac{cov(y, x)}{\sigma_y \cdot \sigma_x} = r_{yx}$

3. Since $b_{yx} = \frac{cov(x, y)}{\sigma_x^2}$ and $b_{xy} = \frac{cov(x, y)}{\sigma_y^2}$

and $r_{xy} = \frac{cov(x, y)}{\sigma_x \cdot \sigma_y}$

$\because \sigma_x > 0, \ \sigma_y > 0,$

$\therefore$ 'sign' of b_{yx} and b_{xy} depend on cov (x, y)

If cov(x, y) is positive, then both b_{yx} and b_{xy} are +ve. And if cov(x, y) is negative then both b_{xy} and b_{yx} are negative.

Thus the sign of correlation coefficient is same as that of regression coefficients. If regression coefficients are positive, r is positive and if regression coefficients are negative r is negative.

4.7.1 Theorems on Regression coefficients

▶ **Theorem 1 :** The correlation coefficient is the geometric mean between the regression coefficients, i.e.

$$r^2 = b_{yx} \cdot b_{xy}$$

Proof : We have

$$b_{yx} = \frac{cov(x, y)}{\sigma_x^2}; \ b_{xy} = \frac{cov(x, y)}{\sigma_y^2}$$

$$= r\frac{\sigma x}{\sigma y} = r\frac{\sigma x}{\sigma y}$$

$$\therefore b_{yx} \cdot b_{xy} = \left(r\frac{\sigma y}{\sigma x}\right) \cdot \left(r\frac{\sigma x}{\sigma y}\right) = r^2$$

$$\therefore r = \pm\sqrt{byx \cdot bxy};$$

Hence the result

Remark

If regression coefficient are positive, we take r +ve and if regression coefficients are negative, we take r – ve.

▶ **Theorem 2 :** The arithmetic mean of the modulus value of the regression coefficients is greater than the modulus value of the correlation coefficients.

Proof : We recall the result : If a and b are any two distinct positive real numbers them

i.e. $\frac{a+b}{2} > \sqrt{ab}$

$$\therefore \frac{1}{2}\left[|b_{xy}| + |b_{yx}|\right] > \sqrt{|b_{xy}| + |b_{yx}|}$$

$$\therefore \frac{1}{2}\left[|b_{xy}| + |b_{yx}|\right] > \sqrt{|b_{xy}| \cdot |b_{yx}|} = |r|$$

$$\therefore \frac{1}{2}\left[|b_{xy}| + |b_{yx}|\right] > |r|$$

▶ **Theorem 3 :** If one of the regression coefficents is greater than unity (one), the other must be less than unity.

Proof :

since $r^2 = b_{yx} \cdot b_{xy}$

and $0 \leq r^2 \leq 1. \quad \therefore b_{yx} \cdot b_{xy} \leq 1$

If $b_{yx} > 1$, then $\frac{1}{byx} < 1$ ...(i)

$\because$ Now, $b_{xy} \leq \frac{1}{byx} < 1$ ...from (i)

▶ **Theorem 4 :** Regression coefficients are independent of change of origin but not of scale.

Proof :

Let $u = \frac{x-a}{h}, \quad V = \frac{y-b}{k};$

Where a, b, h (> 0) and k (> 0) are constants.

Since the correlation coefficient is independent of change of origin and scale, we have

$$r_{xy} = r_{uv} \quad ...(i)$$

Also, $\sigma_x = h\,\sigma_u, \quad \sigma_y = k\,\sigma_v$ ...(ii)

Since standard deviation is independent of change of origin but not of scale

$$\therefore b_{xy} = r_{xy}\frac{\sigma x}{\sigma y} = r_{uv}\frac{h\,\sigma_u}{k\,\sigma_v} = \frac{h}{k}\left[r_{uv}\frac{\sigma u}{\sigma v}\right]$$

$$= \frac{h}{k}b_{uv} \quad ...(iii)$$

$$\therefore b_{yx} = r_{yx}\frac{\sigma y}{\sigma x} = r_{vu}\frac{k\,\sigma_v}{h\,\sigma_u} = \frac{k}{h}r_{vu}\ \frac{\sigma_v}{\sigma_u} = \frac{k}{h}b_{vu}$$

...(iv)

Unit IV End Sem.

$$\therefore b_{xy} = \frac{h}{k} b_{uv} \quad \text{and}$$

$$b_{yx} = \frac{k}{h} b_{vu}$$

Ex. 4.7.1 : The regression lines of a sample are $x + 6y = 6$ and $3x + 2y = 10$. Find (i) sample means $\bar{x}$ and $\bar{y}$ (ii) the coefficient of correlation between x and y (iii) Also, find the value of y at $x = 12$.

☑ **Soln. :**

(i) The regression lines pass through the point $(\bar{x}, \bar{y})$.

So, the regression lines of as sample are

$$\bar{x} + 6\bar{y} = 6$$

$$3\bar{x} + 2\bar{y} = 10$$

To solve the above equations, we get $\bar{x} = 3, \bar{y} = \frac{1}{2}$.

(ii) Consider the line $x + 6y = 6$ be the regression line of y on x. So,

$$y = -\frac{1}{6}x + 1$$

Compare with general form of regression line of y on x.

$$b_{yx} = -\frac{1}{6}$$

Again, consider the line $3x + 2y = 10$ be the regression line of x on y. So,

$$x = -\frac{2}{3}y + \frac{10}{3}$$

Compare with general form of regression line of x on y,

$$b_{xy} = -\frac{2}{3}$$

Thus,

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{\left(-\frac{1}{6}\right)\left(-\frac{2}{3}\right)} = \frac{1}{3}$$

Here, both b_{yx} and b_{xy} are negative. So, r is also negative.

Therefore, the coefficient of correlation is $r = -\frac{1}{3}$.

(iii) From Equation (ii), At $x = 12$,

$$y = -\frac{1}{6}x + 1$$

$$\therefore y = -\frac{1}{6}(12) + 1$$

$$\therefore y = -1$$

Ex. 4.7.2 : From the following results, obtain the two regression equations and estimate the yield when the rainfall is 29 cm and the rainfall, when the yield is 600 kg :

	Yield in kg	Rainfall in cm
Mean	508.4	26.7
SD	36.8	4.6

The coefficient of correlation between yield and rainfall is 0.52.

☑ **Soln. :**

Let x be the rainfall in cm and y be the yield in kg.

Here,

$\bar{x} = 26.7, \sigma_x = 4.6, \bar{y} = 508.4, \sigma_y = 36.8$ and $r = 0.52$

The regression coefficients are

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.52\frac{36.8}{4.6} = 4.16$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.52\frac{4.6}{36.8} = 0.065$$

Now, the regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 508.4 = 4.16(x - 26.7)$$

$$\therefore y = 4.16x + 397.328$$

And the regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 26.7 = 0.065(y - 508.4)$$

$$\therefore x = 0.065y - 6.346$$

When the rainfall x is 29 cm, estimated yield y is

$y = 4.16(29) + 397.328 = 517.968$ kg

When the yield y is 600 kg, estimated rainfall x is

$x = 0.065(600) - 6.346 = 32.654$ cm

Ex. 4.7.3 : The following data give the experience of machine operators and their performance rating as given by the number of good parts turned out per 100 pieces.

Operators	1	2	3	4	5	6
Performance rating (x)	23	43	53	63	73	83
Experience (y)	5	6	7	8	9	10

Calculate the regression line of performance rating on experience and also estimate the probable performance if an operator has 11 years of experience.

☑ **Soln. :**

Here, n = 6

x	y	y^2	xy
23	5	25	115
43	6	36	258
53	7	49	371
63	8	64	504
73	9	81	657
83	10	100	830
$\sum x = 338$	$\sum y = 45$	$\sum y^2 = 355$	$\sum xy = 2735$

The regression coefficient of x on y is

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2} = 11.429$$

Here, $\bar{x} = \frac{\sum x}{n} = 56.33$ and $\bar{y} = \frac{\sum y}{n} = 7.5$

So, the equation of regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$
$$x - 56.33 = 11.429(y - 7.5)$$
$$\therefore x = 11.429\,y - 29.3875$$

When the experience is 11 years of an operator, estimated performance is x = 96.33

UEx. 4.7.4 (19, 7 Marks)

The number of bacterial cells (y) per unit volume in a culture at different hours (x) is given below :

x	0	1	2	3	4	5	6	7	8	9
y	43	46	82	98	123	167	199	213	245	272

Fit lines of regression of y on x and x on y. Also, estimate the number of bacterial cells after 15 hours.

☑ **Soln. :**

Here, n = 10

x	y	x^2	xy	y^2
0	43	0	0	1849
1	46	1	46	2116
2	82	4	164	6724
3	98	9	294	9604
4	123	16	492	15129
5	167	25	835	27889
6	199	36	1194	39601
7	213	49	1491	45369
8	245	64	1960	60025
9	272	81	2448	73984
$\sum x = 45$	$\sum y = 1488$	$\sum x^2 = 285$	$\sum xy = 8924$	$\sum y^2 = 282290$

Here, $\bar{x} = \frac{\sum x}{n} = 4.5$ and $\bar{y} = \frac{\sum y}{n} = 148.8$

The regression coefficients are

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2} = 0.0366$$

and

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = 27.0061$$

The regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$y - 148.8 = 27.0061(x - 4.5)$$
$$\therefore y = 27.0061x + 27.2726$$

The regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$
$$x - 4.5 = 0.0366(y - 148.8)$$
$$\therefore x = 0.366y - 0.9461$$

Thus, at x = 15 hours, y = 432.3641

Unit IV End Sem.

UEx. 4.7.5 (19, 7 Marks)

Find the two lines of regression from the following data:

Age of Husband (x)	25	22	28	26	35	20	22	40	20	18
Age of wife (y)	18	15	20	17	22	14	16	21	15	14

Hence, estimate (i) the age of the husband when the age of the wife is 19, and (ii) the age of the wife when the age of the husband is 30.

☑ **Soln. :**

Let a = 26 and b = 17 be the assumed means of x and y series respectively.

$d_x = x - a = x - 26$ and $d_y = y - b = y - 17$

Here, n = 10

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
25	18	−1	1	1	1	−1
22	15	−4	−2	16	4	8
28	0	2	3	4	9	6
26	17	0	0	0	0	0
35	22	9	5	81	25	45
20	14	−6	−3	36	9	18
22	16	−4	−1	16	1	4
40	21	14	4	196	16	56
20	15	−6	−2	36	4	12
18	14	−8	−3	64	9	24
$\sum x = 256$	$\sum y = 172$	$\sum d_x = -4$	$\sum d_y = 2$	$\sum d_x^2 = 450$	$\sum d_y^2 = 78$	$\sum d_x d_y = 172$

Means of x and y are

$$\bar{x} = \frac{\sum x}{n} = 25.6 \text{ and } \bar{y} = \frac{\sum y}{n} = 17.2$$

The regression coefficients are

$$b_{xy} = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{n\sum d_y^2 - (\sum d_y)^2} = 2.227$$

and

$$b_{yx} = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{n\sum d_x^2 - (\sum d_x)^2} = 0.385$$

The regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$y - 17.2 = 0.385(x - 25.6)$$
$$\therefore y = 0.385x + 7.344$$

The regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$
$$x - 25.6 = 2.227(y - 17.2)$$
$$\therefore x = 2.227y - 12.704$$

When the age of the wife is 19, estimated age of the husband is

$$x = 2.227(19) - 12.704 = 29.601 \approx 30$$

So, Age of the husband is 30 years

When the age of husband is 30, estimated age of the wife is

$$y = 0.385(30) + 7.344 = 18.894 \approx 19$$

So, Age of the wife is 19 years.

4.7.2 Solved Examples on Coefficients of Regression

Ex. 4.7.6 : From the following data, obtain the two regression equations :

Sales :	91	97	108	121	67	124	51	73	111	57
Purchases :	71	75	69	97	70	91	39	61	80	47

Soln. :

Let us denote the sales by the variable x and the purchases by the variable y :

x	y	$u = x - \bar{x}$	$v = y - \bar{y}$	u^2	v^2	uv
91	71	1	1	1	1	1
97	75	7	5	49	25	35
108	69	18	– 1	324	1	– 18
121	97	31	27	961	729	837
67	70	– 23	0	529	0	0
124	91	34	21	1156	441	714
51	39	– 39	– 31	1521	961	1209
73	61	– 17	– 9	289	81	153
111	80	21	10	441	100	210
57	47	– 33	– 23	1089	529	759

We have $\sum x = 900$, $\sum y = 700$, $\sum u = 0$, $\sum v = 0$,

$\sum u^2 = 6360$, $\sum v^2 = 2868$, $\sum uv = 3900$

We have, $\bar{x} = \frac{\sum x}{n} = \frac{900}{10} = 90$,

$\bar{y} = \frac{\sum y}{n} = \frac{700}{10} = 70$,

Now, $b_{yx} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum u \cdot v}{\sum u^2}$

$b_{yx} = \frac{3900}{6360} = 0.6132$

$b_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2} = \frac{\sum u \cdot v}{\sum v^2}$

$b_{xy} = \frac{3900}{2868} = 1.361$

Now, Regression equations

1. Equation of line of regression of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$\therefore y - 70 = 0.6132 (x - 90)$$
$$= 0.6132x - 55.188$$
$$\therefore y = 0.6132 x + 14.812$$

2. Equation of line of Regression of x on y is :

$$x - \bar{x} = b_{xy}(y - \bar{y})$$
$$\therefore x - 90 = 1.361 (y - 70)$$
$$= 1.361 y - 95.27$$
$$\therefore x = 1.361 y - 5.27$$

3. Again, $r^2 = b_{yx} \cdot b_{xy} = (0.6132)(1.361) = 0.8346$

$$\therefore r = \pm\sqrt{0.8346} = \pm 0.9135$$

$\because b_{yx}$ and b_{xy} both are positive

$$\therefore \mathbf{r = 0.9135}$$ **...Ans.**

Ex. 4.7.7 : From the data, given below :

Marks in Economics	25	28	35	32	31	36	29	38	34	32
Marks in statistics	43	46	49	41	36	32	31	30	33	39

Find : (a) The two regression coefficients

(b) two regression equation

(c) coefficients of correlation between marks in Economics states

(d) The most likely marks in statistics when marks in economics are 30.

Soln. :

Let marks in Economics be denoted by the variable x and in statistics by y.

x	y	u	v	u^2	v^2	uv
25	43	– 7	5	49	25	– 35
28	46	– 4	8	16	64	– 32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	– 1	– 2	1	4	2
36	32	4	– 6	16	36	– 24
29	31	– 3	– 7	9	49	21
38	30	6	– 8	36	64	– 48
34	33	2	– 5	4	25	– 10
32	39	0	1	0	1	0

Unit IV End Sem

We have

$$\sum x = 320, \quad \sum y = 380,$$

Also, $u = x - \bar{x}$, $v = y - \bar{y}$

$$\text{and } \bar{x} = \frac{\sum x}{n} = \frac{320}{10} = 32;$$

$$\bar{y} = \frac{\sum y}{n} = \frac{380}{10} = 38$$

$$\therefore u = x - 32, \quad v = y - 38$$

and, $\sum u^2 = 140$, $\sum v^2 = 398$, $\sum vu = -93$

(a) Regression coefficients

Coefficient of Regression of y on x

$$\frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{\sum u \cdot v}{\sum u^2}$$

$$= \frac{-93}{140} = \mathbf{-0.6643}$$

Coefficients of regression of x on y

$$= \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2} = \frac{\sum uv}{\sum v^2}$$

$$= \frac{-93}{398} = \mathbf{-0.2337}$$

(b) Regression Equations

1. Equation of line of regression of y on x :

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\therefore \quad y - 38 = -0.6643\,(x - 32)$$

$$\therefore y = -0.6643x + 0.6643 \times 32 + 38$$

$$= -0.6643x + 59.2576$$

2. Equation of line regression of x on y :

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\therefore \quad x - 32 = -0.2337\,(y - 38)$$

$$= -0.2337\,y + 0.2337\,(38)$$

$$\therefore x = -0.2337y + 32 + 0.2337 \times (38)$$

$$\therefore x = -0.2337y + 40.8806$$

(c) Correlation coefficient

We have

$$r^2 = b_{yx} \cdot b_{xy} = (-0.643)\,(-0.2377)$$

$$= 0.1552$$

$$= r = \pm\sqrt{0.1552} = \pm 0.394$$

Since, both the regression coefficient are negative r must be negative

$$\therefore r = -0.394$$

(d) To estimate the most likely marks in statistics (y) when marks in Economics (x) are 30; we use the line of regression of y on x, i.e. $y = 0.6643\,x + 59.2576$

When $x = 30$; $\quad y = -0.6643\,(30) + 59.2576$

$$\therefore y = 39.3286$$

∴ Most likely marks in statistics are 39. ...Ans.

Ex. 4.7.8 : A panel of judges A and B graded seven debators and independently awarded the following marks

Debator	1	2	3	4	5	6	7
Marks by A	40	34	28	30	44	38	31
Marks by B	32	19	26	30	38	34	28

An eight debator was awarded 36 marks by Judge A while judge b was not present.

If judge B was also present, how many marks would you expect him to award to eight debator assuming some degree of relationship exists in judgement ?

☑ **Soln. :**

Let the marks awarded by judge 'A' be denoted by the variable X and the marks awarded by judge 'B' be the variable Y.

Debator	x	y	$u = x - \bar{x}$	$v = y - \bar{y}$	u^2	v^2	uv
1	40	32	5	2	25	4	10
2	34	39	−1	9	1	81	−9
3	28	26	−7	−4	49	16	28
4	30	30	−5	0	25	0	0
5	44	38	9	8	81	64	72
6	38	34	3	4	9	16	12
7	31	28	−4	−2	16	4	8
Total			$\sum u = 0$	$\sum v = 17$	206	185	121

$\sum u = 0, \sum v = 17, \sum u^2 = 206, \sum v^2 = 185, \sum uv = 121$

To find the marks obtained by eighth debator, we use the equation of line of regression of y on x

We have $\bar{x} = 35, \bar{y} = 30 + \frac{17}{7} = 32.4286$

$$\text{and } b_{yx} = b_{vu} = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum u^2 - (\sum u)^2}$$

$$= \frac{7 \times 121 - 0 \times 17}{7 \times 206} = \frac{121}{206} = 0.5874$$

$\therefore$ Equation of line of regression of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\therefore \quad y - 32.4286 = 0.5874\,(x - 35)$$

$$\therefore \quad y = 0.5874\,x - 0.5874\,(35) + 32.4286$$

$$= 0.5874x + 11.8696$$

When $x = 36$,

$$y = 0.5874\,(36) + 11.8696 = 33.016$$

$\therefore$ Judge B would have given 33 marks to the eight debator. ...Ans.

Ex. 4.7.9 : A departmental store gives in-service training, to its salesman which is followed by a test. It is considering whether it should terminate the service of any salesman who does not do well in the test. The following data give the test scores and sales made by nine salesmen during a certain period

Test scores	14	19	24	21	26	22	15	20	19
Sales (1000 Rs.)	31	36	48	37	50	45	33	41	39

Calculate the coefficient of correlation between the test scores and the sales. Does it indicate that the termination of services of low test scores is justified ? It the firm wants a minimum sales volume of Rs. 30,000, what is the minimum test score that will ensure continuation of service ? Also estimate the most probable sales volume of a salesman making a score of 28.

☑ **Soln. :**

Let x denote the test scores of the salesmen and y denote their corresponding sales (in ′000). Rs.

To calculate regression lines, we prepare the table

Let $u = x - \bar{x} = x - 20, \quad v = y - \bar{y} = y - 40$

x	y	$u = x - \bar{x}$ $= x - 20$	$v = y - \bar{y}$ $= y - 40$	u^2	v^2	uv
14	31	− 6	− 9	36	81	54
19	36	− 1	− 4	1	16	04
24	48	4	8	16	64	32
21	37	1	− 3	1	9	− 03
26	50	6	10	16	100	60
22	45	2	5	4	25	10
15	33	− 5	− 7	25	49	35
20	41	0	1	0	1	00
19	39	− 1	−1	1	1	01
Total 180	360	0	0	120	346	193

We have

$$\bar{x} = \frac{180}{9} = 20, \quad \bar{y} = \frac{360}{9} = 40,$$

$$\sum u^2 = 120, \quad \sum v^2 = 346, \sum uv = 193$$

Now, 1. coefficient of regression of y on x

$$b_{yx} = \frac{\sum u \cdot v}{\sum u^2} = \frac{193}{120} = 1.6083$$

2. and $b_{xy} = \frac{\sum u \cdot v}{\sum v^2} = \frac{193}{346} = 0.5578$

3. Now, $r^2 = b_{yx} \cdot b_{xy}$

$$= 1.6083 \times 0.5578 = 0.8971$$

$$\therefore \quad r = \sqrt{0.8971} = 0.9471$$

$\because$ correlation coefficient are positive, $\therefore r = 0.9471$

To find Regression Equations

1. To obtain the test score (x) for given sales (y), we use the equation of the line of regression of x on y

$$\text{i.e., } x - \bar{x} = b_{x/y}(y - \bar{y})$$

$$\therefore x - 20 = 0.5578\,(y - 40)$$

$$= 0.5578\,y - 0.5578 \times 40$$

$$\therefore x = 0.5578y - 0.5578 \times 40 + 20$$

$$= 0.5578y - 2.312$$

To ensure the continuation of service, the minimum test-score (x) corresponding to a minimum sales volume (y) of Rs. 30 and is given by

$$x = 0.5578\,(30) - 2.312$$

$$= 14.422 = 14$$

Unit IV End Sem

2. To estimate the sales volume (y) of a salesman with given test score (x), we use the line of regression of y on x :

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$\therefore y - 40 = 1.6083(x - 20)$$
$$\therefore y = 1.6083x - 1.6083(20) + 40$$
$$= 1.6083x + 7.8340$$

Hence the estimated sales volume of a salesman with test score of 28 is

$$y = 1.6083(28) + 7.8340 = 52.866$$
$$\text{i.e., } y = 52{,}866$$

Ex. 4.7.10 : The adjoining table shows the number of motor registrations in a certain territory for a term of 5 years and the sale of motor tyres by a firm in that territory for the same period

Find the regression equation to estimate the sale of tyres when motor registration is known.

Estimate sale of tyres when registration is 850.

Year	Motor Registration	No of tyres sold
1	600	1,250
2	630	1,100
3	720	1,300
4	750	1,350
5	800	1,500

☑ **Soln. :**

We denote the number of registration by the random variable x and the number of tyres sold by y.

Now to find regression equation of y on x :

Hence we change the origin as well as scale to both the variables x and y.

Let $u = \frac{x - A}{h} = \frac{x - 720}{10}$

$v = \frac{y - B}{k} = \frac{y - 1350}{50}$

We prepare the table :

x	y	u	v	u2	uv
600	1250	– 12	– 2	144	24
630	1100	– 9	– 5	81	45
720	1300	0	– 1	0	0
750	1350	3	0	9	0
800	1500	8	3	64	24

$\therefore \sum u = -10, \quad \sum v = -5$

$\sum u^2 = 298, \quad \sum uv = 93$

Now, $\bar{x} = A + h\left(\frac{\sum u}{n}\right)$

$= 720 + 10\left(\frac{-10}{5}\right) = 700$

$\bar{y} = B + k\left(\frac{\sum v}{n}\right)$

$= 1350 + 50\left(\frac{-5}{5}\right) = 1300$ And,

$$b_{yx} = \frac{k}{h}(b_{vu}) = \frac{k}{h}\,\frac{n\sum uv - (\sum u)(\sum v)}{n\sum u^2 - (\sum u)^2}$$

$$= \frac{50}{10}\left[\frac{5 \times 93 - (-10)(-5)}{5 \times 298 - (-10)^2}\right]$$

$$= 5\left[\frac{465}{1490} - \frac{50}{100}\right] = 1.4928$$

$\therefore$ Equation of line of regression of sale of tyres (y) on the motor registration is :

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$\therefore y - 1300 = 1.4928(x - 700)$$
$$= 1.4928x - (1.4928)(700)$$
$$\therefore y = 1.4928x - (1.4928)(700) + 1300$$
$$\therefore y = 1.4928x + 255.04$$

When $x = 850$;

$$y = 1.4928(850) + 255.04$$
$$= 1523.92 = \mathbf{1524}$$ **...Ans.**

is the estimate of sale of tyres

Ex. 4.7.11 : The data about the sales and advertisement expenditure of a firm is given below :

	Sales (in crores of Rs.)	Advt. expenditure (in crores of Rs.)
Means	40	6
Standard deviations	10	1.5

Coefficient of correlation = r = 0.9

(i) estimate the likely sales for a proposed advertisement expenditure of Rs. 10 corers.

(ii) What should be the advt. expenditure if the firm propose a sales target of 60 crores of rupees ?

☑ **Soln. :**

Let x denote the sales (in crores of Rs.) and the variable y denote the advertisement expenditure (in crores of Rs.). Then from the given data,

$$\bar{x} = 40, \quad \sigma_x = 10;$$

$$\bar{y} = 6, \quad \sigma_y = 1.5,$$

$$r_{xy} = r = 0.9$$

(i) To estimate the likely sales (x) for a proposed advt. Expenditure (y), we write the regression equation of x on y :

$$x - \bar{x} = r\frac{\sigma x}{\sigma y}\left(y - \bar{y}\right)$$

$$x - 40 = (0.9)\left(\frac{10}{1.5}\right)(y - 6)$$

when $y = 10$;

$$\therefore x - 40 = (0.9)\left(\frac{10}{1.5}\right)(10 - 6) + 40$$

$$= 6 \times 4 + 40 = 64 \text{ crores of Rs.}$$

(ii) To estimate The advt. expenditure (y) for proposed sales (x), we need the equation of line of regression of y on x which is given by

$$y - \bar{y} = r\frac{\sigma y}{\sigma y}\left(x - \bar{x}\right)$$

$$\therefore y - 6 = (0.9)\left(\frac{1.5}{10}\right)(x - 40)$$

$$\therefore y = 0.135\,(x - 40) + 6$$

∴ likely advt. expenditure (y) for the proposed sale target (x) of 60 crores is

$$y = 0.135\,(60 - 40) + 6$$

$$y = 2.7 + 6 = 8.7 \text{ crores of Rs.} \quad \textbf{...Ans.}$$

Ex. 4.7.12 : Point out the inconsistency in the following statement :

The regression equation of y on x is $2y + 3x = 4$ and the correlation coefficient between x and y is 0.8.

☑ **Soln. :**

The regression line of y on x is

$$2y + 3x = 4$$

$$\therefore y = -\frac{3}{2}x + 2$$

$\therefore$ by $x = -\frac{3}{2}$ = coefficient of correlation of y on x

Also, $r = 0.8$ (given)

$\because b_{yx}$ and r have different signs, the given statement is inconsistent.

Ex. 4.7.13 : The following is an estimated supply regression for sugar : $Y = 0.025 + 1.5\,X$

Where y is supply in kilos and X is price (Rs.) per kilo.

(i) Interpret the coefficient of variable X.

(ii) Predict the supply when the price is Rs. 20 per kilo,

(iii) Given that r (x, y) = 1 in the above case, interpret the implied relationship between price and quantity supplied.

☑ **Soln. :**

The regression equation of Y (supply in kgs) on X (price in Rupees per kg) is given to be

$$Y = 0.025 + 1.5\,X = a + bX \text{ (say)}$$

(i) The coefficient of the variable X is b = 1.5 is the coefficient of regression of Y on X.

It reflects the unit change in the value of Y, for a unit change in the corresponding value of X.

This implies that if the price of the sugar goes up by Re. 1 per kg; the estimated supply of sugar goes up by 1.5 kg.

(ii) From (i), the estimated supply of sugar when its price is Rs. 20 per kg. is given by

$$Y = 0.025 + 1.5\,(20) = 30.025 \text{ kg} \textbf{...Ans.}$$

(iii) $\because$ r (X, Y) = 1, implies that the relationship between X and Y is exactly linear. This means that all the observed values (X, Y) lie on a straight line. **...Ans.**

Ex. 4.7.14

(a) On each of 30 items, two measurements are made. The following summations are given :

$\sum X = 15$, $\sum Y = 6$, $\sum XY = 56$, $\sum X^2 = 61$ and $\sum Y^2 = 90$.

(b) Calculate the product moment correlation coefficient and the slope of the regression line of Y on X.

Unit IV End Sem

(c) How would your results be affected if X is replaced by $u = \frac{X-1}{2}$.

Soln. :

(a) The product moment correlation coefficient

$R = r(x, y)$ is,

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2 \left[n\sum Y^2 - (\sum Y)^2\right]}$$

$$= 0\ \frac{1680 + 90}{\sqrt{(1830 - 225)(2700 - 36)}}$$

$$= \frac{1770}{\sqrt{(1605)(2664)}} = 0.856$$

(b) Coefficient of correlation of Y on X

$$= \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

$$= \frac{1680 + 90}{1605} = \frac{1770}{1605} = 1.1028$$

Slope of line of regression of Y on X is 1.1028.

(c) Now, $u = \frac{X-1}{2}$

Since r (X, Y) is independent of change of origin and scale,

$\therefore r(X, Y) = r(u, Y) = 0.856$

But the regression coefficient is independent of change of origin but not of scale.

$$b_{YX} = \frac{1}{2} b_{Yu}$$

$$\therefore b_{yu} = 2\, b_{YX}$$

$$= 2\,(1.1028) = \mathbf{2.2056} \quad \textbf{...Ans.}$$

Ex. 4.7.15 : By using the following data, find out the two lines of regression and from them compute coefficient of correlation.

$\sum X = 250$, $\sum Y = 300$, $\sum XY = 7900$, $\sum X^2 = 6500$ and $\sum Y^2 = 10{,}000$, and $N = 10$.

Soln. :

We have $\bar{X} = \frac{\sum X}{N} = \frac{250}{10} = 25$,

and $\bar{Y} = \frac{\sum Y}{N} = \frac{300}{10} = 30$

Now, b_{yx} = coefficient of regression of Y on X

$$= \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

$$= \frac{10 \times 7900 - (250)(300)}{10(6500) - (250)^2}$$

$$= \frac{7900 - 75000}{65000 - 62500} = \frac{4000}{2500} = 1.6$$

and b_{xy} = coefficient regression of X on Y

$$= \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2}$$

$$= \frac{10 \times 7900 - (250)(300)}{10(1000) - (300)^2} = \frac{4000}{10000} = 0.4$$

$\therefore$ Correlation coefficient r_{XY} between X and Y is given by

$$r_{XY}^2 = b_{yx} \cdot b_{xy} = 1.6 \times 0.4 = 0.64$$

$$\therefore r_{XY} = \pm\sqrt{0.64} = \pm 0.8$$

Since the regression coefficients are positive,

$$\therefore r_{XY} = 0.8$$

Regression Equations

1. Regression equation of Y on X is

$$Y - \bar{Y} = b_{yx}\left(X - \bar{X}\right)$$

$$\therefore Y - 30 = 1.6\,(X - 25)$$

$$\therefore Y = 1.6\,X - 40 + 30$$

$$\therefore Y = 1.6 \times - 10$$

2. Regression equation of X on Y is :

$$X - \bar{X} = b_{YX}\left(Y - \bar{Y}\right)$$

$$\therefore Y - 25 = 0.4\,(Y - 30)$$

$$\therefore X = 0.4\,Y + 25 - 12$$

$$\therefore \mathbf{X = 0.4Y + 13} \quad \textbf{...Ans.}$$

Ex. 4.7.16 : In the estimation of regression equations of two variables X and Y, the following results were obtained :

$\sum X = 900$, $\sum Y = 700$, $\sum XY = 10$;

and $\sum X^2 = 6360$ and $\sum Y^2 = 2860$, $= \sum XY = 3900$

where x and y are deviations from respective means. Obtain the two regression equations.

☑ **Soln. :**

The coefficients of regression of Y on X and X on Y are given by :

$$b_{YX} = \frac{\text{cov}(x, y)}{\sigma_x^2} = ; \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$= \frac{\sum xy}{\sum x^2} = \frac{3900}{6360} = 0.6132$$

and $$b_{XY} = \frac{\text{cov}(x, y)}{\sigma_Y^2}$$

$$= \frac{\sum xy}{\sum y^2} = \frac{3900}{2860} = 1.3636$$

$$\bar{X} = \frac{\sum X}{n} = \frac{900}{10} = 90;$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{700}{10} = 70;$$

(1) Regression equations of Y on X

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\therefore Y - 70 = 0.6132 (X - 90)$$

$$\therefore Y = 0.6132X - 0.6132 (90) + 70$$

$$\therefore Y = (0.6132 X) + 14.812$$

(2) Regression Equation of X on Y

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$\therefore Y - 90 = 1.3636 (Y - 70)$$

$$\therefore X = 1.3636 Y - 1.3636 \times 70 + 90$$

$$\therefore \mathbf{X = 1.3636Y - 5.452}$$ **...Ans.**

Ex. 4.7.17 : Following information regarding a distribution is given :

$n = 5, \bar{X} = 0, \bar{Y} = 20, \sum (X - 4)^2 = 100,$

$\sum (Y - 10)^2 = 160, \sum (X - 4)(Y - 10) = 80$

Find the two regression coefficients and hence the coefficient of correlation.

☑ **Soln. :**

Let $u = X - 4$, $v = Y - 10$, then we have

$n = 5, \bar{X} = 10, \bar{Y} = 20, \sum (X - 4)^2 = \sum u^2 = 100,$

$\sum (Y - 10)^2 = \sum v^2 = 160, \sum (X - 4)(Y - 10) = \sum uv = 80$

Also, $u = X - 4$, $\therefore \bar{u} = \bar{X} - 4 = 10 - 4 = 6$

$\sum u = n\bar{u} = 5 \times 6 = 30,$

$V = Y - 10$, $\therefore \bar{V} = \bar{Y} - 10 = 20 - 10 = 10,$

$\therefore \sum V = n\bar{V} = 5 \times 10 = 50,$

Since, the regression coefficients are independent of the change of origin, the regression coefficient are given by,

$$b_{YX} = b_{vu} = \frac{n \sum uv - (\sum u)(\sum v)}{n \sum u^2 - (\sum u)^2}$$

$$= \frac{5 (80) - (30) (50)}{5(100) - (30)^2} = \frac{-1100}{-400} = \frac{11}{4}$$

and $$b_{XY} = b_{uv} \frac{n \sum uv - (\sum u)(\sum v)}{n \sum v^2 - (\sum v)^2}$$

$$= \frac{5 (80) - (30) (50)}{5(160) - (50)^2} = \frac{-1100}{-700} = \frac{11}{17}$$

$\therefore$ correlation coefficients $= \pm\sqrt{b_{YX} \cdot b_{XY}}$

$$\therefore r = \pm\sqrt{\frac{11}{4} \cdot \frac{11}{7}} = \pm 1.33$$

$\because$ regression coefficients are +ve

$\therefore r = 1.33 > 1$, which is impossible,

$\because |r| \leq 1$ $\therefore$ Given data is inconsistent

4.8 SUPPORT VECTOR MACHINE

- Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms that analyse data for classification and regression analysis.
- It is a **linear model** for classification and regression problems. It can solve linear and non-linear problems and also many practical problems.

Unit IV End Sem.

- The idea of SVM is very simple : the algorithm creates a line or a hyperplane which separates the data into classes.
- Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximise the margin of the classifier. It support reactors are deleted, then there will be change in the position of hyperplane. These are the points that help us build our SVM.

4.8.1 Hyperplane

- Hyperplanes are decision boundaries that help classifying the data points. Data points falling on either side of the hyperplane can be attributed to different classes.
- The dimension of the hyperplane depends upon the number of features.
- A hyperplane is a subspace whose dimension is one less than that of its given space. If the space is 2-dimensional, then the hyperplane is 1 dimensional, i.e. a line. If the space is 1 dimensional, then the hyperplane is a point.
- Plane is a flat surface extending to infinity in all directional, while hyperplane is an n-dimensional generalisation of a plane.
- A hyperplane can also be considered as a curve and it has dimension (at least) 1.
- A hyperplane is a subspace of one dimension less than the given space. Actually hyper means 'over' (in Greek) and it implies excess or exaggeration. Since it has less dimension than the original space, it is called a hyperplane. If the original space has dimension 'n', then hyperplane is of dimension '(n – 1)'.

4.8.2 Equation of Hyperplane

- **Coordinate Hyperplane :** Each pair of axes defines a coordinate hyperplane. These hyperplanes divide the space into 8 trihedral, (∵ there are 8 octants).
- To define the equation of a hyperplane, we need either a point in the plane and a unit vector orthogonal to the plane; two vectors lying in the plane ; or three coplanar points in the hyperplane.
- The equation of the hyperplane is $w \cdot x + b = 0$, where w is a vector normal to the hyperplane and b is an offset.

4.8.3 Applications of SVM

- SVM is a machine learning model that is able to generalise between two different classes if the data is provided in the training set to the algorithm.
- The main function of SVM is to check for that hyperplane that is able to distinguish between the two classes.
- SVM is a supervised machine learning algorithm that can be used for both classification or regression problems.
- SVM is a very good algorithm for classification. It is a system learning algorithm that is mainly used to classify data into different classes.
- SVMs take care of outliers better than KNN. If training data is much larger than the number of features (m >> n), KNN is better than SVM. But SVM outperforms KNN when there are large features and lesser training data.

Remark : The K-nearest neighbors (KNN) algorithm is a simple and easy-to-implement supervised machine learning algorithm. It can be used to solve both classification and regression problems.

4.8.4 Support Vectors

- The optimal (maximum margin) hyperplane remains unchanged if we remove all training instances **but** the support vectors. Hence they are given the name 'support vectors.'
- Support vectors are data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane. Hence they are considered as the **'critical elements'** of a data set.

4.8.5 Types of SVM

1. **Admin SVM :** It is the cluster set-up process automatically creates the admin SVM for the cluster.

2. **Node SVM :** A node SVM is created when the node joins cluster, and the node SVM represents the individual nodes of the cluster.

4.8.6 Functions of SVM

1. SVM separates the classes with a decision surface that maximizes the margin between the classes. A decision tree, having its origin in machine learning theory, is an efficient tool for the solution of classification and regression problems.
2. The biggest difference between the models that are built from a "feature" point of view is that Naïve Bayes treats them as independent, whereas SVM looks at the interactions between them to a certain degree, as long as a nonlinear kernel is used.
3. SVM is a learning algorithm mainly used at classification problems, which considers the data as support vectors and generates a hyperplane to classify them.

- Under these support vectors, we maximize the margin of the classifier.
- Kernel function is a method used to take data as input and transform into the required form of processing data.

4.8.7 Optimal Hyperplane

- In a binary classification problem, given a linearly separable data set, the optimal separating hyperplane is the one that correctly classifies all the data, which being farthest away from the data points. New test points are drawn according to the same distributions as the training data.
- To define an optimal hyperplane, we need to maximise the width of margin (w). The beauty of SVM is that if the data is linearly separable, there is a unique global minimum value.

4.8.8 Optimal Hyperplane for Linearly Separable Patterns

- Let $(\{x_i, d_i\})_{i=1}^{n}$ be the training sample, where x_i is the input pattern for i^{th} example and d_i is the corresponding desired response.
- Here we assume that the patterns represented by subsets are linearly separable.
- We assume that the pattern represented by subset $d_i = +1$ and the pattern represented by the subset $d_i = -1$ are 'linearly separable'
- Now the equation of a decision surface in the form of a hyperplane that does the separation is

$$w^T x + b = 0 \quad \ldots(i)$$

- Where x is input vector, w is adjustable weight vector and b is a bias.

Thus we can now write

$$w^T x_i + b \geq 0 \text{ for } d_i = +1 \quad \ldots(ii)$$

and $$w^T x_i + b < 0 \text{ for } d_i = -1$$

- The separation between the hyperplane from Equation (i) and the closest data point is called the 'margin of separation'.
- Our aim is to find that particular hyperplane for which the margin of separation is maximum. The 'decision surface' so obtained (under this condition) is called as optimal hyperplane. Fig. 4.8.1 gives the geometric construction of an optimal hyperplane for two-dimensional space.
- Let us w_0 and b_0 be the optimum values of the weight vector and bias respectively, then the optimal hyperplane is given by in Fig. 4.8.1.

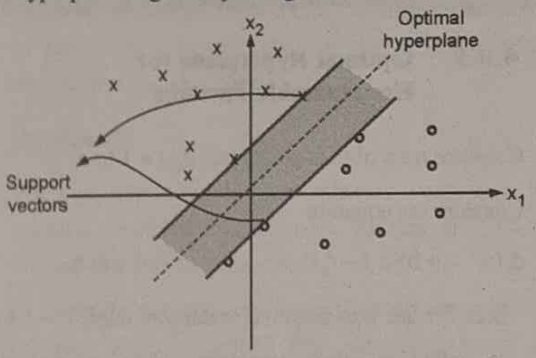


Fig. 4.8.1

$$w_0^T x + b_0 = 0 \quad \ldots(iii)$$

- Consider $g(x) = w_0^T x + b_0$; this is discriminant function and gives the 'distance' from x to the optimal hyperplane

Unit IV End Sem.

We can write x as

$$x = x_p + r\frac{w_0}{\| w_0 \|}$$

where x_p is normal projection of x onto the optimal hyperplane and r is the distance.

$\left(\text{Note that } \frac{w_0}{\| w_0 \|} \text{ is a unit vector}\right)$

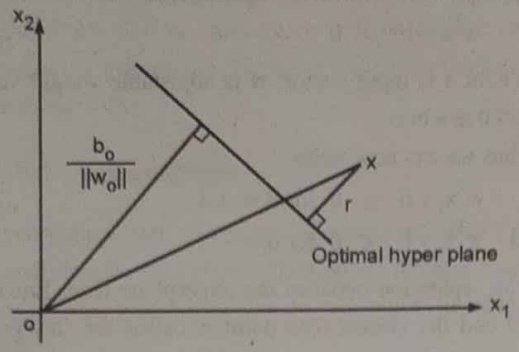


Fig. 4.8.2 : Algebraic distance (geometrically)

Since $g(x_p) = 0$,

$$\therefore g(x) = w_0^T x + b_0 = r \| w_0 \|$$

$$\therefore r = \frac{g(x)}{\| w_0 \|}$$

Remark : (i) r is positive is if x is on the positive side of the optimal hyperplane.

(ii) r is negative if x is on the negative side of the optimal hyperplane.

4.8.9 Optimal Hyperplane for Nonseparable Patterns

Consider a set of data points (x_i, d_i) i = 1 to n

Consider the equation

$$d_i (w^T x_i + b) \geq 1 - \xi i \qquad i = 1 \text{ to } n,$$

where ξ i are non-negative scalar variables, i = 1 to r and are called as slack variables. They measure the deviation of a data point from the ideal condition of pattern separability.

The support reactors are the points which satisfy

$$d_i (w^T x_i + b) \geq 1 - \xi i \qquad i = 1 \text{ to } n$$

The data point falls inside the region of separation

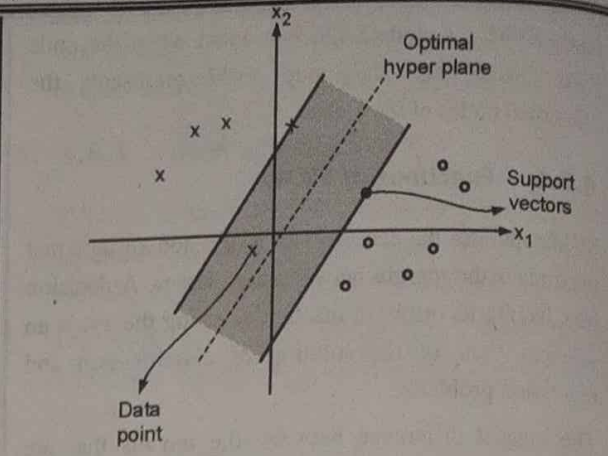


Fig. 4.8.3

4.9 SUPPORT VECTOR REGRESSION

- Support Vector Regression (SVR) is a supervised learning algorithm, that is used to predict discrete values. S.V.R. uses the same principle as the support vector machines (SVMs). The basic idea behind SVR is to find the best fit line. In SVR, the best fit lines is hyperplane, that has the maximum number of points.
- SVR is slightly different from SVM. SVR is a regression algorithm, so we can use SVR for working with continuous values instead of classification which is SVM.
- For regression problems, SVR is the counterpart of SVM. SVR acknowledges the presence of non-linearity in the data and provides a proficient prediction model.
- SVR uses the same principle of SVM. The approach of using SVMs to solve regression problems is called SVR.

4.9.1 Kernel

- It is a function that maps a lower dimensional data into a higher dimensional data.
- Kernel function is a method used to take data as input and transform into the required form of processing data.
- After the transformation, the equations of linear SVM are applicable as is leading to optimal classification.
- SVR gives the flexibility to define how much error is acceptable in our model and will find an appropriate line or hyperplane in higher dimensions to fit the data.

4.9.2 Regression Tree

It is built through a processor known as binary recursive partitioning, which is an iterative process that splits the data into partitions or branches, and then continues splitting each partition into smaller groups as the method moves up each branch.

4.9.3 Disadvantages of Decision Tree

1. Decision tree cannot be used with continuous numerical variables.
2. A small change in data tends to cause a big difference in the structure, which causes instability.
3. Calculations involved can also become complex compared to other algorithms.
4. and it takes a longer time to train the model.

4.9.4 When to Use Decision Tree

When the decision tree has a continuous target variable, e.g. a regression tree would be used for the price of a newly launched product because price can be anything depending on various constraints.

4.9.5 Classification Algorithm Regression Tree (CART)

UQ. Explain following Decision Tree Algorithms
i) CART **SPPU – Q. 5(a), May 19, 9 Marks)**

- It is a predictive algorithm used in machine learning. It explains how a target variable's values can be predicted based on other values. It is a decision tree where each fork is split in a predictor variable and each node at the end has a prediction for the target values.
- Decision trees are among the most popular machine learning algorithm given their intelligibility and simplicity.

4.9.6 Difference between Classification Tree and Regression Tree

- The primary different between them is : The classification decision trees are built with unordered values with dependent variables.
- The regression decision trees take ordered values with continuous values.
- If the response variable is something like the price of a property or the temperature of the day, a regression tree is used. Regression trees are used for prediction-type problems while decision trees are used for classification-type problems.
- Regression trees can be used for a newly launched product because price can be anything depending on various constraints.
- Both types of decision trees fall under classification and Regression (CART) designation.

4.9.7 Terminology of Regression Tree

1. Root

This is the beginning of a decision tree, which also represents the population sample, e.g. if we want to decide the employee, who can carry out a perfect successfully, then the entire population of employees is at the root of the 'decision tree'.

2. Leaf

The terminal node is called the leaf node. In our example, the final best employee would be the leaf node or the terminal node.

3. Decision node

The other nodes are divided into the further categories. In the above example, the various criterion would determine a decision node.

4. Child node

- When a node is divided into other subparts, then the subparts are called as child nodes. And the node that is divided is called the parent node.
- Consider the above example of a Regression tree where the root node is divided into sub-nodes based on desired (continuous) values. It is further divided into other subparts before getting to the leaf node or the terminal node.

4.9.8 Advantages of Regression Tree

1. A user should visualize each step, which can help with making rational decisions.

Unit IV End Sem.

2. One should give priority to a decision criterion, .e.g. in our employee example, one can put the attendance criterion on the top of the 'decision tree' if that is the most important criterion.
3. Making a decision based on regression is much easier than most other methods. Since most undesired data will be filtered outlier at each step, one has to work on less data as one goes further in the tree.
4. It is easy to prepare regression tree. A user can present it to the higher authorities in a much easier way as it can be represented on a simple chart or diagram.

4.10 LOGISTIC REGRESSION

UQ. Explain logistic regression. Explain use cases of logistic regression.
(SPPU – Q. 5(b), Aug. 18, 4 Marks)

Logistic Regression is supervised learning classification algorithm used to predict the probability of an output variable. The nature of dependent variable is such that there would be only two possible classes.

4.10.1 L.R. Classification

In a classification problem output or target variable y, can take any discrete values for given set of features or inputs X. L.R. is a regression model. Steps of L-R are :

1. Data Pre-processing step
2. Fitting logistic regression to the training set.
3. Predicting the test results,
4. Test accuracy of the result
5. Visualising the test set result.

4.10.2 Sigmoid Function

- It is a powerful machine learning algorithm that utilises a sigmoid function and works best an binary classification problems, although it can be used on multi-class classification problems through the 'one versus all' method.
- In spite of its name, logistic regression is not fit for 'regression tasks'.
- The idea of L.R is to find a relationship between features and probability of particular outcome,
- e.g. when we have to predict if a student passes or fails in an examination when the number of hours spent studying is given as a feature, the response variable has two values pass and fail.
- Logistic Regression is basically a statistical analysis method used to predict a data value based on prior observations of a data set, e.g. L.R. can be used to predict whether a student from a village will be admitted to a particular college.

4.10.3 Advantages of L.R.

1. Logistic regression is better than linear regression. Linear regression is used to handle regression problems whereas Logistic Regression is used to handle the classification problems.
2. Linear regression provides a continuous output but logistic regression provides discrete output.
3. There are two reasons why linear regression is not suitable for classification. The first one is that linear regression deals with continuous values whereas classification problems require discrete values.
4. When new data points are added, there is a shift in threshold value].
5. Logistic reasoning is one of the most important supervised learning classification method.
6. It is a fast, versatile extension of a generalised linear model.
7. It is easier to implement and interpret. And it is very effective to train. If number of observations is lesser than the number of features, Logistic regression is not to be used, otherwise it may lead to overfitting.
8. It makes no assumptions about distribution of classes in feature space.

4.10.4 Disadvantages of L.R.

1. The limitation of logistic reasoning is the assumption of linearity between the dependent variable and the independent variable.
2. It not only provides a measure of how appropriate a predictor is, but also its direction of association as (positive) or (negative)

4.10.5 Calculation of L.R.

- L.R. is calculated as the odds ratio denoted by OR. It is simply the odds of being a case for one group divided by the odds of being a case for another group. Logistic reasoning takes the natural logarithms of the odds (referred to as the 'logit' or log-odds) to create a continuous criterion.
- The log function has the effect of removing the floor restriction, i.e. 'logit' function transforms values in the range 0 to 1 to values over the entire real no. range $(-\infty, \infty)$.

4.10.6 Examples on Logistic Regression

Type : Probability of passing an examination versus hours of study :

Ex. 4.10.1 : group of 20 students spends between 0 and 6 hours studying for an examination. How does the number of hours spent studying affect the probability of the student passing the exam ?

☑ **Soln. :**

▶ **Step (I) :** The given table shows the number of hours each student spent studying and whether they passed (1) or failed (0)

Hours x_i	0.50	0.75	1.00	1.25	1.50	1.75	2.0	2.25	2.75	3.0
Pass y	0	0	0	0	0	1	0	1	1	0

Hours x_i	3.0	3.25	3.50	4.0	4.25	4.50	4.75	5.00	5.50
Pass y_i	1	0	1	1	1	1	1	1	1

▶ **Step (II) :** We fit a logistic function to the data consisting of hours studied (x_i) and outcome of the test ($y_i = 1$ for pass, 0 for fail)

The x variable is called the "explanatory variable" and the y variable is called the "categorical variable" consisting of two categories "pass" or "fail" corresponding the values 1 and 0 respectively.

The logistic function is of the form

$$p(x) = \frac{1}{[1 + e^{-\lambda(x-\mu)}]}$$

▶ **Step (III) :** Let the 'fit' to y_i at a given xi as :

$$p_i = p(x_i)$$

We find λ and μ which give the "best fit" to the data

The "best fit" is given by where the L is maximised and L is given by

$$L = \prod_{i\,:\,y_i=1} (P_i) \cdot \prod_{i\,:\,y_i=0} (1-P_i)$$

Taking log and let log L = l, then

$$l = \sum_{i\,:\,y_i=1} \log (P_i) + \sum_{i\,:\,yi=0} \log (1 - P_i)$$

▶ **Step (IV) :** Differentiate l w.r.t. μ and λ;

$$\therefore \frac{\partial l}{\partial \mu} = \lambda \sum_i (y_i - p_i) = 0 \quad \ldots(i)$$

$$\text{and} \quad \frac{\partial l}{\partial \lambda} = \lambda \sum_i (y_i - p_i)(x_i - \mu) = 0 \quad \ldots(ii)$$

on simplification

$$\sum_i (y_i - p_i) = 0 \quad \ldots(iii)$$

$$\text{and} \sum_i (y_i - p_i)\, x_i = 0 \quad \ldots(iv)$$

on solving, $\mu = 2.71008$

and $\lambda = 1.50465$

$\therefore \lambda\mu = (2.71008)(1.50465) = 4.07771$

We have $p = 0.0167$ and

Thus for a student who studies 4 hours, the estimated probability of passing the exam is 0.87

We prepare the table

▶ **Step (V) :** We prepare the Table :

Hours of study	Probability of passing exam
1	0.07
2	0.26
2.71	$\frac{1}{2}$
3	0.61
4	0.87
5	0.97

Unit IV End Sem

Ex. 4.10.2 : Logistic function is defined by

$$\phi(v) = \frac{1}{1 + \exp(-av)}$$

Whose limiting values are 0 and

1. Show that the derivative of $\phi(v)$ w.r.t. v is given by

$$\frac{d\phi}{dv} = a\,\phi(v)\,[1 - \phi(v)]$$

What is the value of this derivative at origin?

☑ **Soln. :**

▶ **Step (I) :** We have $\phi(v) = \frac{1}{1 + \exp(-av)}$

Different w.r.t. v.; we get

$$\frac{d\phi}{dv} = \frac{-}{[1 + \exp(-av)]^2}\,[-a \exp(-av)]$$

$$= \frac{a \exp(-av)}{[1 + \exp(-av)]^2}$$

$$= \frac{a\,[1 + \exp(-av) - 1]}{[1 + \exp(-av)]^2}$$

$$= a\left[\frac{1}{(1 + \exp(-av))} - \frac{1}{\{1 + \exp(-av)\}^2}\right]$$

$$= a\,[\phi(v) - \{\phi(v)\}^2]$$

$$= a\,\phi(v)\,[1 - \phi(v)] \quad ...(i)$$

▶ **Step (II) :** At origin, V = 0,

$\therefore \exp(-av) = 1$

From (i), at v = 0;

$$\frac{d\phi}{dv} = a\,(1)[1 - 1] = 0$$

Ex. 4.10.3 : The logistic function is given by

$$\phi_j(v_j(n)) = \frac{1}{1 + \exp(-av_j(n))},$$

$a > 0$, and $-\infty < v_j(n) < \infty$

Where $v_j(n)$ is the induced local field of neuron j. And the amplitude of the output lies inside the range $0 \le y_j \le 1$.

Show that $\frac{d\phi_j}{dv_j} = a y_j(n)[1 - y_j(n)]$

☑ **Soln. :**

▶ **Step (I) :** Note that $y_i(n) = \phi_j(v_j(n))$ . ...(i)

Differentiating v_j w.r.t $v_j(n)$; we get

$$\frac{d\phi_j}{dv_j} = \frac{-1}{[1 + \exp(-av_j(n))]^2}\,[-a \exp(-av_j(n))]$$

$$= \frac{a \exp(-a\,v_j(n))}{[1 + \exp(-av_j(n))]^2}$$

$$= \frac{a\,[1 + \exp(-av_j(n)) - 1]}{[1 + \exp(-av_j(n))]^2}$$

$$= a\left[\frac{1}{\{1 + \exp(-av_j(n))\}} - \frac{1}{\{1 + \exp(-av_j(n))^2\}}\right]$$

$$= a\left[\phi_j(v_j(n)) - \phi_j^2(v_j(n))\right] = a\,\phi_j(v_j(n))\,[1 - \phi_j(v_j(n))]$$

...(ii)

▶ **Step (II) :** Using ii, $y_i = \phi_j(v_j(n))$; we get

$$\frac{d\phi_j(n)}{dv_j(n)} = a\,y_i(n)[1 - y_i(n)]$$

4.11 NAIVE BAYES THEOREM

To develop a system of problem solving is to collect evidence as the system goes along and to modify its behaviour on the basis of evidence. To model this behaviour, we need a statistical theory of evidence. Bayesian statistics is such a theory.

4.11.1 Naive Bayes Classifiers

In statistics, naive Bayes classifiers are a family of simple 'probabilistic classifiers' based on applying Bayes' theorem with independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with density estimation they can achieve higher accuracy levels.

Bayes' theorem states : Given class variable y and dependent feature vectors $x_1, x_2, ,,,,, x_n$

$$P\left(\frac{y}{x_1,x_2,\ldots,x_n}\right) = \frac{P(y) \cdot \frac{(x_1,x_2,\ldots x_n,)}{y}}{P(x_1,x_2,\ldots,x_n)} \quad ...(a)$$

Using naïve conditional independence assumption,

$$P\left(\frac{x_i}{y}, x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots x_n\right) = P\left(\frac{x_i}{y}\right), \forall\, i$$

$\therefore$ (a) becomes

$$P\left(\frac{y}{x_1,x_2,\ldots,x_n}\right) = \frac{P(y)\cdot\prod_{i=1}^{n} P\left(\frac{x_i}{y}\right)}{P(x_1,x_2,\ldots,x_n)}$$

Since $P(x_1, x_2,\ldots,x_n)$ is constant, given the input, we can use the following classification rule :

$$P\left(\frac{y}{x_1,x_2,\ldots,x_n}\right) \propto P(y)\prod_{i=1}^{n} P\left(\frac{x_i}{y}\right),$$

And we can use maximum posteriori (MAP) estimation to obtain P(y) and $P\left(\frac{x_i}{y}\right)$; the former is then the relative frequency of class y in the training set.

4.11.2 Examples on Naive Bayes

Ex. 4.11.1 : A company has two plants to manufacture scooters Plan I manufactures 80% of the scooters and plant II manufactures 20%. At plant I, 85 out of 100 scooters are rated standard quality or better. At plant II, only 65 out of 100 scooters are rated standard quality or better.

(i) What is the probability that scooter selected at random came from plant I, if it is known that the scooter is of standard quality.

(ii) What is the probability that the scooter came from plant II, if it is known that the scooter is of standard quality?

☑ **Soln. :**

We define the following events :

E_1 : Scooter is manufactured by plant I

E_2 : Scooter is manufactured by plant II

E : Scooter is rated at standard quality

Then we are given

$P(E_1) = 0.80$, $P(E_2) = 0.20$

$P(E/E_1) = 0.85$, $P(E/E_2) = 0.65$

Tree diagram

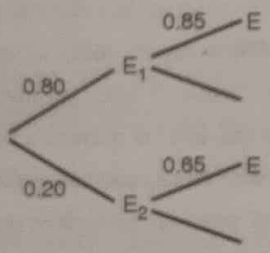


Fig. P. 4.11.1

Events	Probability
$E \cap E_1$	$0.80 \times 0.85 = 0.68$
$E \cap E_2$	$0.20 \times 0.65 = 0.13$
	Total $= 0.81$

Now,

(i) $P(E_1/E) = \frac{0.68}{0.81} = 0.84$ and

(ii) $P(E_2/E) = \frac{0.13}{0.81} = 0.16$

Ex. 4.11.2 : Classify whether a given person is a male or a female based on the measured features. The features include height, weight, and foot size.

Training set :

Person	Height (feet)	Weight (1bs)	Foot size (increase)
Male	6	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5	100	6
Female	5.5	150	8
Female	5.42	130	7
Female	5.75	150	9

☑ **Soln. :**

▶ **Step (I) :** The classifier is

person	Mean height	variance height	Mean weight	variance weight	Mean foot size	Variance (foot size)
Male	5.855	3.5033×10^{-2}	176.25	1.2292×10^{2}	11.25	9.1667×10^{-1}
Female	5.4175	9.7225×10^{-2}	132.5	5.5833×10^{2}	7.5	1.6667

Step (II) : Using normal distribution and Baye's rule, we have

P (height/female) $= 2.23 \times 10^{-1}$

P(weight/female) $= 1.6789 \times 10^{-2}$

P (foot size/female) $= 2.8669 \times 10^{-1}$

And

P(height/male) $= 1.5789$

P (weight/ male) $= 5.9881 \times 10^{-6}$

Unit IV End Sem

P (foot size/male) = 1.3112×10^{-3}

▶ **Step (III) :** since numerator (posterior) is greater in the female case, the prediction is that the sample is female.

Ex. 4.11.3 : The contents of urns I, II and III are as follows:

1 white, 2 black and 3 red balls;

2 white, 1 black and 1 red balls;

4 white, 5 black and 3 red balls;

One urn is chosen at random and two balls drawn. They happen to be white and red. What is the probability that they come from urns I, II, III?

☑ **Soln. :**

▶ **Step (I) :** Let E_1, E_2, E_3 be the events of choosing 1st, 2nd and 3rd urn respectively and let E be the event that the two balls drawn from the selected urn are white and red. Then we have :

	E_1	E_2	E_3
$P(E_i)$	1/3	1/3	1/3
$P\left(\frac{E}{E_i}\right)$	$\frac{1\times3}{6C_2}=\frac{1}{5}$	$\frac{2\times1}{4C_2}=\frac{1}{3}$	$\frac{4\times3}{12C_2}=\frac{2}{11}$
$P(E\cap E_i)=P(E_i)\cdot P\left(\frac{E}{E_i}\right)$	$\frac{1}{3}\cdot\frac{1}{5}=\frac{1}{5}$	$\frac{1}{3}\cdot\frac{1}{3}=\frac{1}{9}$	$\frac{1}{3}\cdot\frac{2}{11}=\frac{2}{33}$

▶ **Step (II) :** We have

$$\sum_i P(E_i)\cdot P\left(\frac{E_1}{E_i}\right)=\frac{1}{15}+\frac{1}{9}+\frac{2}{33}=\frac{118}{495}$$

By Baye's rule, the probability that the two white and red balls drawn are from 1st urn is

$$P\left(\frac{E_1}{E}\right)=\frac{P(E_1)\cdot P\left(\frac{E}{E_1}\right)}{\sum_i P(E_i)\cdot P\left(\frac{E}{E_i}\right)}=\frac{1/15}{118/495}=\frac{33}{118}$$

▶ **Step (III) :** Similarly :

$$P\left(\frac{E_2}{E}\right)=\frac{P(E_2)\cdot P\left(\frac{E}{E_2}\right)}{\sum_i P(E_i)\cdot P\left(\frac{E}{E_i}\right)}=\frac{\frac{1}{9}}{\frac{118}{495}}=\frac{55}{118}$$

And $$P\left(\frac{E_3}{E}\right)=\frac{P(E_3)\cdot P\left(\frac{E}{E_3}\right)}{\sum_i P(E_i)\cdot P\left(\frac{E}{E_i}\right)}=\frac{\frac{2}{33}}{\frac{118}{495}}=\frac{30}{118}$$

4.12 DECISION TREES

UQ. What is decision tree? Explain how decision tree is constructed using ID3 algorithm **(SPPU – Q. 4(a), Dec. 18, 8 Marks)**

UQ. Write an Apriori Algorithm. **(SPPU – Q. 4(a), May 19, 5 Marks)**

UQ. What is decision tree? Explain various terms used in Decision Tree. **(SPPU – Q. 5(b), Dec. 19, 8 Marks)**

▶ **Introduction :** The book 'Port-Royal Logic (Arnaula, 1662)' states :

"To judge what one must do to obtain a good or avoid an evil, it is necessary to consider not only the good and the evil in itself, but also the probability that it happens or does not happen; and to view geometrically the proportion that all these things have together."

- Today, we talk of utility rather than good or evil. To get expected utility, we have to multiply utility by probability and maximize it over all (good as well as evil) outcomes.
- To generalise 'utility' concept, we solve classification problems. 'Decision Trees' are good at solving classification problems.
- A decision tree is a map of the reasoning process. It describes the available data by a tree-like structure.
- These are other methods, like 'Rote learning method', 'Learning from observation', 'Learning from Agents', but these methods depend more or less on numerical data. But the theory of 'decision tree' does not only rely on numerical data but it draws a tree-type structure such that at each node, one can make a decision. In 'decision trees' practical methods or short-cuts are used in order to produce a solution that fits the problem.

4.12.1 Decision Tree

- Fig. 4.12.1 shows a decision tree for farmhouse plots in the vicinity of sindhudurg in Konkan area in Maharashtra. This will promote a new consumer product, local people nearby may get some jobs, opening of restaurants etc.

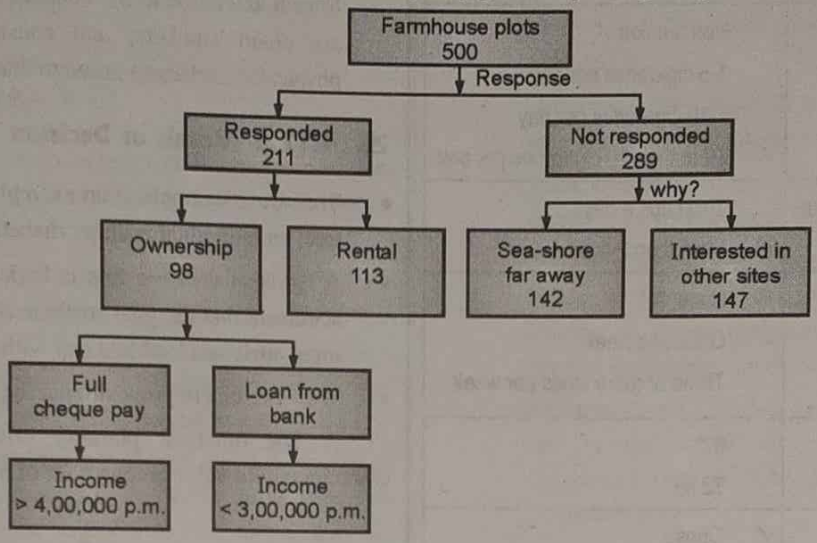


Fig. 4.12.1 : Example of 'Decision Tree'

- A decision tree consist of nodes, branches and leaves.
- A node is represented by each box.
- The top node is called the root node. The tree always starts from the root node and the root node contains the whole data. It is also called as 'Parent node'.
- The tree starts from 'root node' and grows down into new nodes. The child nodes hold the particular information and are subsets of the 'root-node'.
- Nodes at the end of the branches are called 'terminal nodes' and all nodes are connected by branches.

Remarks :

(I) The 'parent node' is called as dependent variable and it determines the aim of the study. Here 'Farmhouse' is dependent variable and it can have two values either responded or not responded. It is a Boolean Classification.

(2) Child nodes just below root-node are called 'Predictor' for the dependent variable, and it separates or splits the data.

4.12.2 'The Decision Tree' Representation

- A decision tree can be regarded as a function that takes as input values and gives out a decision – a single output value. The input and output values can be continues or discrete.
- Generally we come across the problems where the inputs have discrete values and the output has exactly two values. These two values are classified as true (i.e positive) and false. (i.e. negative). This is Boolean classification.
- Let us consider the above example. Each child node corresponds to the value of the input data, say A_i and branches from this node A_i correspond to values of the $A_i = V_{ik}$
- Let us consider an example of Community health survey : 'Diabetes Study (Pune)'
- Note that Decision trees are as good as the data they represent. Decision trees do not tolerate polluted data. Hence data must be cleaned before we begin with data base:

Table 4.12.1 : Community health survey study : Diabetes (Pune Maharashtra)

1.	Gender	✓ Male Female
2.	Age	20-35 years 36-50 years ✓ 51-65 years Above 65

Unit IV End Sem

3.	Alcohol Consumption	Occasionally Regular (one or two drinks) ✓ Heavy
4.	Smoking	Non smoker 1-5 cigarettes per day 5-10 cigarettes per day ✓ More than 10 cigarettes per day.
5.	Coffee or Tea intake	Two cups a day ✓ More than 2 cups a day
6.	Physical exercise	None ✓ Once in a week Three or more times per week
7.	Height weight	67" 72 kg
8.	Obesity	✓ Obese Not obese
9.	Blood pressure	Normal ✓ High
10.	Race	Hindus ✓ Muslims Christians Jains
11	Household income	Less than 50,000/- p.m. 50,000/-75,000/- p.m ✓ 75,000/-1,00,000/- p.m Above 1,00,000 /- p.m
12	Marital status	✓ Married Divorced Bachelor Separated

- As we continue growing the tree node by node. We find that Muslims have a much higher risk of diabetes than the other groups and smoking and heavy drinking increase this risk further.
- Note that here every variable indicates the possibility of 'diabetic patient'.
- Here, we have ignored the variable height and weight, since that is a polluted data.
- We begin with the root tree and follow the appropriate branch and come to the conclusion that the people who are chain smokers, and consume alcohol and no physical exercise are prone to diabetes.

4.12.3 Result of Decision Trees

- The above example is an example of Boolean decision tree; an individual is either diabetic or not.
- A Boolean decision tree is logically equivalent to the statement that the goal attribute is true if and only if the input attributes lead to a leaf with value true.
- We write this in propositional logic as :

∴ The function 'plurality value' selects the most common output value among a set of nodes.

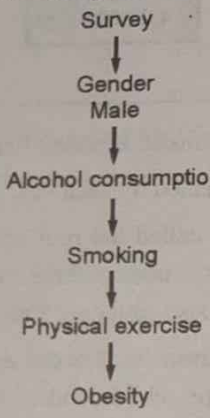


Fig. 4.12.2

4.12.4 Drawbacks of 'Decision Tree'

1. The decision tree gives concise results for a wide variety of problems. But some functions cannot be represented concisely. For example, the functions in which more than half of the inputs are true require a large number of nodes and hence a large tree. Thus there is no representation that is efficient for all kinds of functions.
2. The decision tree-learning algorithm test the most important attribute first. By "most important attribute" is meant the classification which makes the most difference of an example. That implies that the tree will be short and shallow.

3. There is a danger of over-interpreting the tree that the algorithm chooses. If there are several variables of similar importance, then the choice of choosing the node becomes difficult.

4.12.5 Extending the Applicability of Decision Trees

To extend the concept of applicability of decision trees to a wider variety of problems, it must handle below mentioned problems :

1. **Missing data :** In many problems, some of the attributes may be missing. Then it becomes difficult to classify the data when a particular attribute is missing. Also if some of attributes have unknown values, then the information-gain formula cannot be obtained.
2. **Multivalued attributes :** If the attribute has many possible values, then the usefulness of the attribute is at stake. If one of the values is chosen, then it may not yield the best tree.

   Here one of the possible value of the attribute is chosen, and is tested and later the remaining values can be tested.
3. **Continues input attributes :** Continues attributes have many possible values, e.g. height and weight. Instead of generating infinitely many branches for each possible value, we choose 'split-point' that gives the best information gain. Treating continues inputs is very important because financial and physical processes provide such numerical data.

4.12.6 Advantages of Decision Tree

1. The main advantage of decision tree approach is that it visualizes the solution of the problem.
2. It is easy to follow any path through the tree.
3. Relationships discovered by a decision tree can be expressed as a set of rules; and the set of rules can be used in developing expert system.

4.12.7 Decision-Tree Learning

- It is one of the predictive modeling approaches used in statistics, data mining and machine learning.
- It uses a decision tree to go from observations about an item to conclusions about the item's target value.
- Decision tree used in data mining are of two main types :
  - **Classification data** : It is when the predicted outcomes is the class (discrete) to which data belongs :
  - **Regression tree analysis** is when the predicted outcome can be considered a real number.
  - Some techniques, often called Ensemble methods, construct more than one decision tree.
  - **Bagged decision tree :** is an ensemble method, builds multiple decision trees by repeatedly training data with replacement, and voting the trees for a consensus prediction.
  - **Rotation forest :** In this method, every decision tree is trained by first applying Principal Component Analysis (P(A)) on a random subset of the input feature.

Notable decision tree algorithms are :

1. **ID3** (Iterative Dichotomiser 3)
2. **CA5** (Successor of ID3)
3. **CART** (Classification and regression tree)
4. **Chi-square** automatic iteration detection (CHAID)

   It performs multi-level splits when computing classification trees.
5. **MARS :** Extends decision – trees to handle numerical data better.

4.12.8 Conditional Interferance Trees

- It is statistics –based approach that uses non-parametric tests as splitting criterion, connected for multiple testing to avoid overfitting.
- This approach results in unbiased predictor selection and does not require pruning.
- ID3 and CART follow similar approach for learning 'decision tree' from training tuples.

4.12.9 Random Forest

UQ. Explain following term : Random forest (SPPU – Q. 6(a), May 19, Q. 6(a), Dec. 19, 9 Marks)

Unit IV End Sem

- **Introduction :** A random forest or random discrete function is a machine learning technique that is used to solve regression and classification problems by constructing a multitude of decision trees and training time.
- For classification tasks, the output of 'random forest' is the class selected by most trees.
- For regression tasks, the mean or average prediction of the individual trees is returned. Random discrete functions correct for decision trees habit of overfitting to their training set.
- Random forest generally outperforms decision trees, but their accuracy is lower than gradient boosted trees.
- Random forest are frequently used as 'blackbox' models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

4.12.10 Random Forest Model : Random

- Forest ensemble model made up of many decision trees using bootstrapping, random subsets of features and average voting to make predictions. This is an example of 'bagging ensemble'.
- A decision tree combines some decisions, whereas random forest combines several decision trees. It is a long process, but slow whereas 'decision tree' is fast and operates on large data set, especially the linear one. The random forest model needs rigorous training.
- The random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction is more accurate than that of any individual tree.

4.12.11 Advantages of Random Forest

1. It can perform both regression and classification tasks.
2. A random forest produces good predictions that can be understood easily.
3. It can handle large datasets efficiently.
4. The random forest algorithm produces a higher level of accuracy in predicting outcomes over the decision tree algorithm.

When the number of features is large, it is preferable to use a higher number of regression trees.

4.12.12 Random Forest Model

Random Forest Models are data-sets, the size of the trees can take up a lot of memory.

- The random forest are so called because each tree in the forest is built by randomly selecting a sample of data.

Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best **solution by means of voting.**

4.13 DECISION TREE LEARNING

- Decision-tree learning or induction of decision-tree is one of the predictive modeling approaches used in statistics, data-mining and machine learning. It uses a decision tree to go from observations about on item to conclusions about the item's target value.
- Decision Trees (DTs) are non-parametric supervised learning method used for classification and regression.
- Decision trees learn from data to approximate a sine curve with a set of 'if-then-else' decision rule. The deeper the tree, the more complex the decision rules and the fitter the model.
- A decision tree is a simple representation for classifying examples. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature.
- In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making.

4.13.1 Decision Tree Terminologies

- **Step 1 :** Start with the root node, say T, which contains the complete data set.
- **Step 2 :** Choose the best attribute in the dataset using Attribute Selection Measure (ASM).
- **Step 3 :** Divide the T into subsets that contains possible values for the best attributes.

Note that a decision tree is a map of the possible outcomes of a series of related choices. It allowed an individual or organization to weigh possible actions against one another based on their costs probabilities and benefits.

4.13.2 K-Means

- K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k-clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
- K-means clustering is one of the simplest and popular unsupervised machine learning algorithm. It is a centroid-based algorithm where we calculate the distances to assign a point to a cluster.
- The k-means clustering algorithm is used to find groups which have not been explicitly labeled in the data.
- This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

4.13.3 Calculation of K-Means

We calculate K-mean on follows :

- **Step 1 :** Choose the number of clusters k.
- **Step 2 :** Select K random points from the data as cemtroids.
- **Step 3 :** Assign all the points to the closet cluster centroid.
- **Step 4 :** Recompute the centroids of newly formed clusters.
- **Step 5 :** Repeat steps (3) and (4) and we can calculate K-mean.
- In data-mining, K-means clustering intends to partition n objects into K clusters in which each object belongs to the with the nearest mean. This method produces exactly K-different clusters of greatest possible distinction.

4.13.4 Advantages of K-means

1. K means clustering is unsupervised learning method. This is because it is one of the best ways to explore and find out more about data visually.

4.13.5 Disadvantages of K-means

1. Difficult to predict K-value.
2. With global cluster, it didn't work well.
3. Different initial partitions can result in different final clusters.

4.13.6 K-nearest neighbor (KNN)

- K-means is an unsupervised learning algorithm used for clustering problems whereas KNN is a supervised learning algorithm used for classification and regression problems.
- This is the basic difference between K-means and KNN algorithm.

4.13.7 K-NN Algorithm

K-NN algorithm at the raining phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example

- Suppose we have an image of a creature that looks similar to crow and Kingfisher, but we want to know either it is a Crow or Kingfisher.
- So for this identification, we can use KNN Algorithm as it works on a similarity measure. Our KNN-model will find the similar features of the new data set to the Crow and Kingfisher images and based on the most similar features it will put it in either Crow or Kingfisher category.
- KNN is one of the simplest machine learning algorithm based on supervised learning technique.
- KNN algorithm assumes the similarity between the new case or data and available cases and put the new case into the category that is most similar to the available category.
- KNN algorithm stares all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well-defined category by using KNN algorithm.
- KNN-algorithm can be used for regression as well as for classification but mostly it is used for the classification problems.
- KNN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

- It is also called as lazy learner algorithm because it does not lean from the training set immediately, instead it stores the dataset and at the time of classification, it performs an action on the dataset.

4.13.8 Need for KNN-Algorithm

- Suppose we have two categories, i.e. Crow A and Kingfisher B and we have a new datapoint x_1, so this data point will lie in which of these categories.
- To solve this type of problem, we need a KNN algorithm and with this we can easily identify the category or class of a particular dataset.
- The K-NN working can be explained on the basis of the below algorithm.
  - **Step 1 :** Select the number K of the neighbors.
  - **Step 2 :** Calculate the euclidean distance of K number of neighbor.
  - **Step 3 :** Take the K nearest neighbors as per the calculate Euclidean distance.
  - **Step 4 :** Among these K-neighbors, count the number of data points in each category.
  - **Step 5 :** Assign the new data points to that category for which the number of the neighbor is maximum.

  Now the model is ready to implement.

4.13.9 Selection of Value of K

- There is no particular method to determine the best value for 'K' in KNN-algorithm.
- So we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value of K such as K = 1 or K = 2 can be noisy and lead to effects of outliers in the model.
- Large values of K are good but it may find some difficulties.

4.13.10 Advantages of KNN Algorithm

1. It is simple to implement,
2. It is robust to the noisy training data.
3. It can be more effective if the training data is large.

4.13.11 Disadvantage of KNN Algorithm

1. Always needs to determine the value of K which may be complex some time.
2. The computation cost is high because of calculating the distance between the data points for all the training samples.

4.13.12 KNN Classification and Regression

- In KNN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k-nearest neighbor (K is a positive integer, typically small). If K = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In K-NN regression, the output is the property value for the object. This value is the average of the values of a K nearest neighbors.
- K-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evolution. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalising the training data can improve its accuracy dramatically.
- Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones, e.g. a common weighting scheme consists in giving each neighbor a weight of $\frac{1}{d}$, where d is the distance to the neighbor.
- The neighbors are taken from a set of objects for which the class (for K-NN classification) is known. This can be thought of as the training set for the algorithm, though no explicit training is required.
- A peculiarity of K-NN algorithm is that it is sensitive to the local structure of the data.

4.13.13 Parameter Selection

- The best choice of K depends upon the data, generally larger values of K reduces effect of the noise on the classification, but make boundaries between classes less distinct. A good value of K can be selected by various trial techniques. The special case where the class is predicted to be the class of the closest training sample (i.e. when K = 1) is called the nearest neighbor algorithm.
- The accuracy of the KNN algorithm is severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. A particularly popular approach is the use of evolutionary algorithm to optimise feature scaling. Another popular approach is to scale features by the mutual information of the training data with the training classes.
- In binary classification problem, it is helpful to choose K to be an odd number as this avoids tied votes. One popular way of choosing the empirically optimal K in this setting is via bootstrap method.

4.13.14 The 1-Nearest Neighbour Classifier

- The most intuitive nearest neighbor type classifier is the one nearest neighbor classifier that assigns a point x to the class of its closest neighbor in the feature space, i.e. $C^{1nn}_{n}(x) = y_{(1)}$.
- As the size of training data set approaches ∞, the one nearest neighbor classifier guarantees an error rate of no worse than twice the Bayer error rate (the minimum achievable error rate given the distribution of the data).

4.13.15 The Weighted Nearest Neighbour Classifier

- The K-nearest neighbor classifier can be viewed as assigning the K nearest neighbors a weight $\frac{1}{k}$, and all others '0' weight. This can be generalized to weighted nearest neighbor classifier. That is, where the i^{th} nearest neighbor is assigned a weight W_{n_i} with

$$\sum_{i=1}^{n} W_{ni} = 1$$

4.13.16 Properties

- K-NN is a special case of a variable bandwidth, kernel density with a uniform kernel.
- The naive version of the algorithm is easy to implement by computing the distances from the test example to all stored examples, but it is computationally intensive for large training sets using an **approximate** nearest neighbor search algorithm makes K-NN computationally tractable even for large data sets. Many nearest neighbor search algorithms have been proposed over the years, those generally seek to reduce the number of distance evaluation actually performed.
- K-NN has some strong consistency results. As the amount of data approaches infinity, the two-class K-algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate. Various improvements to the K-NN are possible by using proximity graphs :

For multi-class KNN classification, an upper bound error rate of

$$R^* \le R_{K\text{-}NN} \le R^*\left[2 - \frac{MR^*}{(M-1)}\right]$$

- Where R^* is the Bayes error rate (which is the minimal error rate possible), $R_{K\text{-}NN}$ is K-NN error rate, and M is the number of classes in the problem. For M = 2 and as the Byesian error rate R^* approaches zero, the limit reduces the "not more than twice the Bayesian error rate."

4.13.17 Feature Extraction

- When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then the input data will be transformed into a reduced representation set of features (also named feature vector). Transforming the input data into the set of features is called 'feature extraction'.
- If the features extracted are carefully chosen, it is expected that the feature set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction is performed on raw data prior to applying K-NN algorithm on the transformed data in the feature space.

Unit IV End Sem

4.13.18 Distance Functions

Euclidean : $\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$

Manhatten : $\sum_{i=1}^{k} |x_i - y_i|$

Minkoswski : $\left[\sum (|x_i - y_i|)^q\right]^{1/q}$

- The above three distance measures are only valid for continuous variables. But in case of categorical variables, we use 'Hamming distance' which is a measure of number of instances.

  Hamming distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

4.13.19 Data Points

- The number of data points that are taken into consideration is determined by the k-values. Thus, the k value is the 'core of the algorithm'. KNN classifier determines the class of a data point by the majority voting principle. If K is set to 5, the classes of 5 closest points are checked.
- KNN algorithm can also be used for regression problems.

4.13.20 Classification Accuracy

Classification accuracy of the KNN algorithm is found to be adversely affected by the presence of outliers, in the experimental datasets. An outlier score based on rank difference can be assigned to the points in these datasets by taking into consideration the distance and the density of their local and neighborhood points.

Advantages

(i) Nonparametric architecture,

(ii) Simple and powerful,

(iii) Requires no training time

(iv) KNN can be used for recommendation systems. Although in the real world, more sophisticated algorithms are used for the recommendation system, KNN is not suitable for high dimensional data, but KNN is an excellent baseline approach for the system.

(iv) KNN is not limited to merely predicting groups or values of data points. It can also be used in detecting anomalies. Identifying anomalies can be the end goal in itself, such as in fraud detection.

Disadvantages

(i) Memory intensive

(ii) Classification and estimation are slow.

(iii) The value of K in the KNN-algorithm is related to the error rate of the model. Overfitting imply that the model is well on the training data but has poor performing when the new is coming.

4.13.21 Applications of KNN

(I) Applications of KNN in finance

(i) Forecasting stock market

(ii) Predict the price of a stock on the basis of company performance measures and economic data.

(iii) Currency exchange rate.

(iv) Bank bankruptcies

(v) Understanding and managing financial risk

(vi) Trading futures

(vii) Credit rating

(viii) Loan management

(ix) Bank customer profiling

(x) Money laundering analysis.

(II) Medicine

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.
- Estimate the amount of glucose in the blood of a diabetic person, from the infrared absorption spectrum of that person's blood.
- Identify the risk factors for prostate cancer, based on clinical and demographic variables.

- The KNN algorithm has been also applied for analyzing micro-array gene expression data, where the KNN algorithm has been coupled with genetic algorithms, which are used as a search tool.
- Other applications include the prediction of solvent accessibility in protein molecules, the detection of intrusions in computer systems, and the management of databases of moving objects such as computer with wireless connections.
- KNN is not limited to merely predicting groups or values of data points, It can also be used in detecting anomalies. Identifying anomalies can be the main aim, such as in fraud detection.

4.13.22 Applications of Classification and Regression Algorithms

- Regression analysis is used for 'prediction and forecasting applications'. In short, when the intension is to assign objects to difference categories then we use classification algorithms and when we want to predict future values then we use regression algorithms.
- Applications of classification algorithms are :
  (i) Sentiment analysis ;
  (ii) E-mail spam classification,
  (iii) Document classification,
  (iv) Image classification
- The main difference between regression and classification algorithms that regression algorithms are used to predict the continuous values such as price, salary, age etc. and classification algorithms are used to predict/classify the discrete values such as male or female, True of false, spam or not spam etc.
- Fundamentally, classification is about predicting a label and regression is about predicting a quantity. That classification is the problem of predicting a discrete class label output for an example. Regression is the problem of predicting a continuous quantity output for an example.

4.13.23 Examples on Decision Tree Learning

Ex. 4.13.1 : Draw a decision tree for the following data using Information gain :
Training set : 3 features and 2 classes.

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

Soln. :

Step I :

Here, we have 3 features and 2 output classes.

We consider each feature and calculate the information for each feature

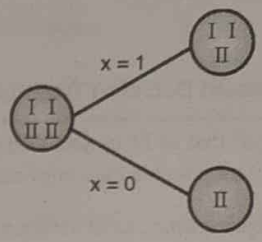


Fig. P. 4.13.1

$$E_{parent} = 1$$

$$Gain = 1 - \left(\frac{3}{4}\right)(0.9184) - \frac{1}{4}(0)$$

$$= 0.3112$$

$$E_{child} = \frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$$

$$= 0.5284 + 0.39 = 0.9184$$

$$E_{child} = 0$$

Now, we consider split on **feature X**

Step II : Split on feature X

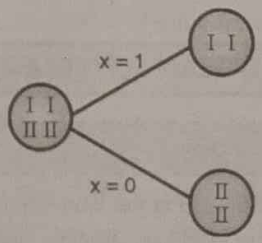


Fig. P. 4.12.1(a)

Unit IV End Sem.

$E_{parent} = 1$

Gain = 1 – 0 = 1

$E_{child} = 0$

▶ **Step III :** Split on feature Y

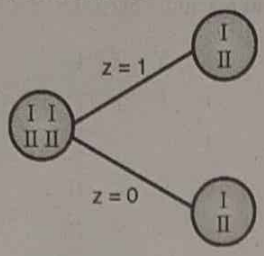


Fig. P. 4.13.1(b)

$E_{split} = 1$

Gain = 1

$E_{child} = 1$

▶ **Step IV :** Split on feature Z

From the above images we can see that the information gain is maximum when we make a split on feature Y

So for the root node best suited feature is feature Y

We see that while splitting the dataset by feature Y, the child contains pure subset of the target variable. Hence we need not proceed further to split the dataset.

The final tree for the above dataset would be look like this.

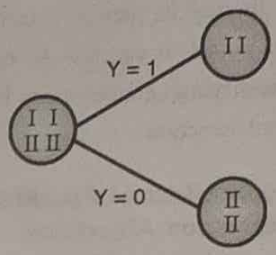


Fig. P. 4.13.1(c)

Project Management Decision Tree Example

Ex. 4.13.2 : Imagine that an IT project manager needs to decide whether to start a particular project or not. He needs to take into account important possible outcomes and consequences.

☑ **Soln. :** Project management decision helper

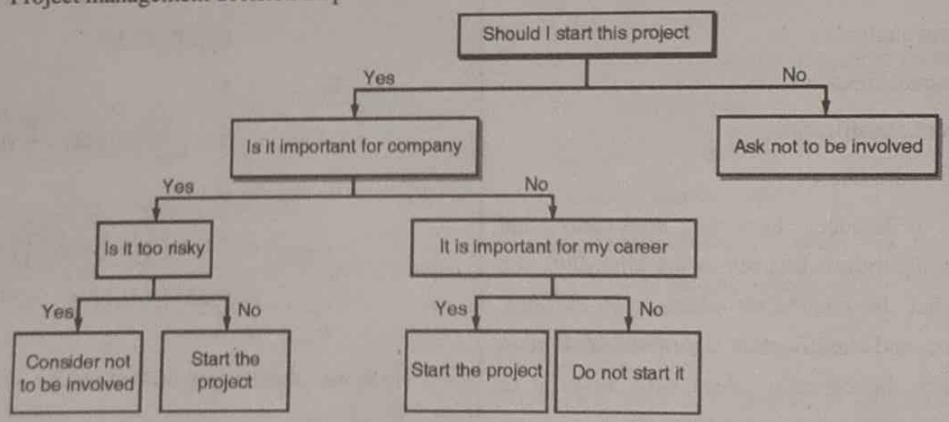


Fig. P. 4.13.2

4.14 SCIKIT-LEARN (SKLEARN)

GQ. Explain Scikit Learn ? **(4 Marks)**

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon **NumPy, SciPy** and **Matplotlib**.

Origin of Scikit-Learn

- It was originally called ***scikits.learn*** and was initially developed by David Cournapeau as a Google summer of code project in 2007.
- Later, in 2010, Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, and Vincent Michel, from FIRCA (French Institute for Research in Computer Science and Automation), took this project at another level and made the first public release (v0.1 beta) on 1st Feb. 2010.
- Let's have a look at its version history :
  - May 2019: scikit-learn 0.21.0
  - March 2019: scikit-learn 0.20.3
  - December 2018: scikit-learn 0.20.2
  - November 2018: scikit-learn 0.20.1
  - September 2018: scikit-learn 0.20.0
  - July 2018: scikit-learn 0.19.2
  - July 2017: scikit-learn 0.19.0
  - September 2016. scikit-learn 0.18.0
  - November 2015. scikit-learn 0.17.0
  - March 2015. scikit-learn 0.16.0
  - July 2014. scikit-learn 0.15.0
  - August 2013. scikit-learn 0.14

Community & contributors

- Scikit-learn is a community effort and anyone can contribute to it. This project is hosted on https://github.com/scikit-learn/scikit-learn.
- Following people are currently the core contributors to Sklearn's development and maintenance –
  - Joris Van den Bossche (Data Scientist)
  - Thomas J Fan (Software Developer)
  - Alexandre Gramfort (Machine Learning Researcher)
  - Olivier Grisel (Machine Learning Expert)
  - Nicolas Hug (Associate Research Scientist)
  - Andreas Mueller (Machine Learning Scientist)
  - Hanmin Qin (Software Engineer)
  - AdrinJalali (Open Source Developer)
  - Nelle Varoquaux (Data Science Researcher)
  - Roman Yurchak (Data Scientist)
- Various organisations like Booking.com, JP Morgan, Evernote, Inria, AWeber, Spotify and many more are using Sklearn.

Prerequisites

Before we start using scikit-learn latest release, we require the following :

- Python (>=3.5)
- NumPy (>= 1.11.0)
- Scipy (>= 0.17.0)li
- Joblib (>= 0.11)
- Matplotlib (>= 1.5.1) is required for Sklearn plotting capabilities.
- Pandas (>= 0.18.0) is required for some of the scikit-learn examples using data structure and analysis.

4.14.1 Installation

If you already installed NumPy and Scipy, following are the two easiest ways to install scikit-learn :

Using pip

Following command can be used to install scikit-learn via pip :

```
pip install -U scikit-learn
```

Using conda

Following command can be used to install scikit-learn via conda :

```
conda install scikit-learn
```

- On the other hand, if NumPy and Scipy is not yet installed on your Python workstation then, you can install them by using either **pip** or **conda**.
- Another option to use scikit-learn is to use Python distributions like ***Canopy*** and ***Anaconda*** because they both ship the latest version of scikit-learn.

Features

Rather than focusing on loading, manipulating and summarising data, Scikit-learn library is focused on modeling the data. Some of the most popular groups of models provided by Sklearn are as follows :

Unit IV End Sem.

1. **Supervised Learning algorithms :** Almost all the popular supervised learning algorithms, like Linear Regression, Support Vector Machine (SVM), Decision Tree etc., are the part of scikit-learn.
2. **Unsupervised Learning algorithms :** On the other hand, it also has all the popular unsupervised learning algorithms from clustering, factor analysis, PCA (Principal Component Analysis) to unsupervised neural networks.
3. **Clustering :** This model is used for grouping unlabeled data.
4. **Cross Validation :** It is used to check the accuracy of supervised models on unseen data.
5. **Dimensionality Reduction** : It is used for reducing the number of attributes in data which can be further used for summarisation, visualisation and feature selection.
6. **Ensemble methods :** As name suggest, it is used for combining the predictions of multiple supervised models.
7. **Feature extraction :** It is used to extract the features from data to define the attributes in image and text data.
8. **Feature selection :** It is used to identify useful attributes to create supervised models.
9. **Open Source :** It is open source library and also commercially usable under BSD license.

4.14.2 Dataset Loading

A collection of data is called dataset. It is having the following two components :

- **Features** : The variables of data are called its features. They are also known as predictors, inputs or attributes.
  - **Feature matrix** : It is the collection of features, in case there are more than one.
  - **Feature Names** : It is the list of all the names of the features.
- **Response** : It is the output variable that basically depends upon the feature variables. They are also known as target, label or output.
  - **Response Vector** : It is used to represent response column. Generally, we have just one response column.
  - **Target Names :** It represent the possible values taken by a response vector.
- Scikit-learn have few example datasets like **iris** and **digits** for classification and the **Boston house prices** for regression.

Example

Following is an example to load **iris** dataset :

```
fromsklearn.datasetsimportload_iris
iris =load_iris()
X =iris.data
y =iris.target
feature_names=iris.feature_names
target_names=iris.target_names
print("Feature names:",feature_names)
print("Target names:",target_names)
print("\nFirst 10 rows of X:\n", X[:10])
```

Output

```
Feature names: ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)']
Target names: ['setosa' 'versicolor' 'virginica']
First 10 rows of X:
[
  [5.1 3.5 1.4 0.2]
  [4.9 3. 1.4 0.2]
  [4,7 3.2 1.3 0.2]
  [4.6 3.1 1.5 0.2]
  [5. 3.6 1.4 0.2]
  [5.4 3.9 1.7 0.4]
  [4.6 3.4 1.4 0.3]
  [5. 3.4 1.5 0.2]
  [4.4 2.9 1.4 0.2]
  [4.9 3.1 1.5 0.1]
]
```

Splitting the dataset

To check the accuracy of our model, we can split the dataset into two pieces-**a training set** and **a testing set**. Use the training set to train the model and testing set to test the model. After that, we can evaluate how well our model did.

Example

The following example will split the data into 70:30 ratio, i.e. 70% data will be used as training data and 30% will be used as testing data. The dataset is iris dataset as in above example.

```
fromsklearn.datasetsimportload_iris
iris =load_iris()

X =iris.data
y =iris.target

fromsklearn.model_selectionimporttrain_test_split

X_train,X_test,y_train,y_test=train_test_split(
  X, y,test_size=0.3,random_state=1
)

print(X_train.shape)
print(X_test.shape)

print(y_train.shape)
print(y_test.shape)
```

Output

```
(105, 4)
(45, 4)
(105,)
(45,)
```

As seen in the example above, it uses **train_test_split()** function of scikit-learn to split the dataset. This function has the following arguments :

- **X, y :** Here, **X** is the **feature matrix** and y is the **response vector**, which need to be split.
- **test_size :** This represents the ratio of test data to the total given data. As in the above example, we are setting **test_data = 0.3** for 150 rows of X. It will produce test data of 150*0.3 = 45 rows.
- **random_size :** It is used to guarantee that the split will always be the same. This is useful in the situations where you want reproducible results.

☞ Train the Model

Next, we can use our dataset to train some prediction-model. As discussed, scikit-learn has wide range of **Machine Learning (ML) algorithms** which have a consistent interface for fitting, predicting accuracy, recall etc.

GQ. Explain Machine Learning Algorithm with Example.

☞ Example

In the example below, we are going to use KNN (K nearest neighbors) classifier. Don't go into the details of KNN algorithms, as there will be a separate chapter for that. This example is used to make you understand the implementation part only.

```
fromsklearn.datasetsimportload_iris
iris =load_iris()
X =iris.data
y =iris.target
fromsklearn.model_selectionimporttrain_test_split
X_train,X_test,y_train,y_test=train_test_split(
  X, y,test_size=0.4,random_state=1
)
fromsklearn.neighborsimportKNeighborsClassifier
fromsklearnimport metrics
classifier_knn=KNeighborsClassifier(n_neighbors=3)
classifier_knn.fit(X_train,y_train)
y_pred=classifier_knn.predict(X_test)
# Finding accuracy by comparing actual response values(y_test)with predicted response value(y_pred)
print("Accuracy:",metrics.accuracy_score(y_test,y_pred))
# Providing sample data and the model will make prediction out of that data
sample =[[5,5,3,2],[2,4,3,5]]
preds=classifier_knn.predict(sample)
pred_species=[iris.target_names[p]forp
inpreds]print("Predictions:",pred_species)
```

Output

```
Accuracy: 0.9833333333333333
Predictions: ['versicolor', 'virginica']
```

☞ Model Persistence

- Once you train the model, it is desirable that the model should be persist for future use so that we do not need to retrain it again and again. It can be done with the help of **dump** and **load** features of ***joblib*** package.
- Consider the example below in which we will be saving the above trained model (classifier_knn) for future use :

```
from sklearn.externals import joblib
joblib.dump(classifier_knn, 'iris_classifier_knn.joblib')
```

Unit IV End Sem.

- The above code will save the model into file named iris_classifier_knn.joblib. Now, the object can be reloaded from the file with the help of following code –

```
joblib.load('iris_classifier_knn.joblib')
```

☞ Preprocessing the Data

As we are dealing with lots of data and that data is in raw form, before inputting that data to machine learning algorithms, we need to convert it into meaningful data. This process is called preprocessing the data. Scikit-learn has package named **preprocessing** for this purpose. The **preprocessing** package has the following techniques

⏭ 4.15 MATPLOTLIB

GQ. Write a short note on MATPLOTLIB **(2 Marks)**

- Matplotlib is a plottinglibrary for the Python programming language and its numerical mathematics extension NumPy. It provides an object-orientedAPI for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.
- There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.[3]SciPy makes use of Matplotlib.

☞ Installation of Matplotlib

- If you have Python and PIP already installed on a system, then installation of Matplotlib is very easy.
- Install it using this command:

```
C:\Users\Your Name>pip install matplotlib
```

- If this command fails, then use a python distribution that already has Matplotlib installed, like Anaconda, Spyder etc.

☞ Import Matplotlib

- Once Matplotlib is installed, import it in your applications by adding the import module statement:

```
import matplotlib
```

- Now Matplotlib is imported and ready to use:

☞ Checking Matplotlib Version

The version string is stored under __version__ attribute.

☞ Example

```
import matplotlib
print(matplotlib.__version__)
```

☞ fill missing values

- The SimpleImputer is a Python class from Scikit-Learn that is used to fill missing values in structured datasets containing None or NaN data types.
- As the name suggests, the class performs simple imputations, that is, it replaces missing data with substitute values based on a given strategy. Let's have a look at the syntax for SimpleImputer initialization to understand this better:

```
SimpleImputer(*, missing_values=nan, strategy='mean', fill_value=None, verbose=0, copy=True, add_indicator=False)
```

- The parameters/arguments in the SimpleImputer class are as follows:
  - **missing_values :** This is a placeholder for the missing values to fill and it is set to np.nan by default. All occurences of this parameter's value will be imputed.
  - **Strategy :** This parameter defines the imputation strategy and you can either set it to 'mean', 'median', 'most_frequent', or 'constant'.
  - **fill_value :** This parameter is used when the strategy=constant and a constant value that is to be filled is needed to be supplied. By default, the fill_value is set as 0.
  - **Verbose :** This parameter is used to control the verbosity of the SimpleImputer and is 0 by default.
  - **Copy :** If True, a copy of the input dataset will be created. If False, imputation will be done in-place whenever possible.
  - **add_indicator :** If True, a MissingIndicator transform will stack onto output of the imputer's transform. This allows a predictive estimator to account for missingness despite imputation.

Getting started with the SimpleImputer

- To start using the SimpleImputer class, you must install the Scikit-Learn library in your machine alongside Python.

- You can run the following command from your command line/terminal to install scikit-learn using Python's Package Manager (pip):

```
pip install scikit-learn
```

- Once you've installed the library, you can import it in Python by running the following line of code in your Python IDE or Python Shell.

```
import sklearn
```

- If running this line of code doesn't give you an error, you've successfully installed Scikit-Learn and imported it in Python. Now, you can use the SimpleImputer to fill missing values.
- **Scikit-learn** is the most popular Python library for performing **classification, regression, and clustering algorithms.** It is an essential part of other Python data science libraries like `matplotlib`, `NumPy` (for graphs and visualization), and `SciPy` (for mathematics).

4.16 CLASSIFICATION & REGRESSION

GQ. Explain Classification & Regression in SCIKIT-Learn? **(4 Marks)**

- Machine Learning is a fast-growing technology in today's world. Machine learning is already integrated into our daily lives with tools like face recognition, home assistants, resume scanners, and self-driving cars.
- **Scikit-learn** is the most popular Python library for performing **classification, regression, and clustering algorithms.** It is an essential part of other Python data science libraries like `matplotlib`, `NumPy` (for graphs and visualization), and `SciPy` (for mathematics).

Refresher on Machine Learning

- Machine Learning is teaching the computer to perform and learn tasks without being explicitly coded. This means that the system possesses a certain degree of **decision-making capabilities**. Machine Learning can be divided into three major categories:

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

Types of Machine Learning

1. **Supervised Learning :** In this ML model, our system learns under the **supervision** of a teacher. The model has both a known input and output used for training. The teacher knows the output during the training process and trains the model to reduce the error in prediction. The two major types of supervised learning methods are **Classification** and **Regression**.
2. **Unsupervised Learning :** Unsupervised Learning refers to models where there is **no supervisor** for the learning process. The model uses just input for training. The output is learned from the inputs only. The major type of unsupervised learning is Clustering, in which we cluster similar things together to find patterns in unlabeled datasets.
3. **Reinforcement Learning :** Reinforcement Learning refers to models that learn to make decisions based on **rewards or punishments** and tries to maximize the rewards with correct answers. Reinforcement learning is commonly used for gaming algorithms or robotics, where the robot learns by performing tasks and receiving feedback. In this post I am going to explain the two major methods of **Supervised Learning**:

- **Classification :** In Classification, the output is discrete data. In simpler words, this means that we are going to categorize data based on certain features. For example, differentiating between Apples and Oranges based on their shapes, color, texture, etc. In this example shape, color and texture are known as features, and the output is "Apple" or "Orange", which are known as Classes. Since the output is known as classes, the method is called Classification.
- **Regression :** In Regression, the output is continuous data. In this method, we predict the trends of training data based on the features. The result does not belong to a certain category or class, but it gives a numeric output that is a real number. For example, predicting House Prices is based on certain features like size of the house, location of the house, and no. of floors, etc.

How to implement classification and regression

- Python provides a lot of tools for implementing Classification and Regression. The most popular open-

Unit IV End Se

source Python data science library is scikit-learn. Let's learn how to use scikit-learn to perform Classification and Regression in simple terms.

- The basic steps of supervised machine learning include:

1. Load the necessary libraries
2. Load the dataset
3. Split the dataset into training and test set
4. Train the model
5. Evaluate the model

1. Loading the Libraries

```
#Numpy deals with large arrays and linear algebra
import numpy as np
# Library for data manipulation and analysis
import pandas as pd
# Metrics for Evaluation of model Accuracy and F1-score
from sklearn.metrics import f1_score,accuracy_score
#Importing the Decision Tree from scikit-learn library
From sklearn.tree import DecisionTreeClassifier
# For splitting of data into train and test set
from sklearn.model_selection import train_test_split
```

2. Loading the Dataset

```
train=pd.read_csv("/input/hcirs-ctf/train.csv")
# read_csv function of pandas reads the data in CSV format
# from path given and stores in the variable named train
# the data type of train is DataFrame
```

3. Splitting into Train & Test set

```
#first we split our data into input and output
# y is the output and is stored in "Class" column of dataframe
# X contains the other columns and are features or input
y = train.Class
train.drop(['Class'], axis=1, inplace=True)
X = train
# Now we split the dataset in train and test part
# here the train set is 75% and test set is 25%
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.25, random_state=2)
```

4. Training the model

```
# Training the model is as simple as this
# Use the function imported above and apply fit() on it
DT= DecisionTreeClassifier()
DT.fit(X_train,y_train)
```

5. Evaluating the model

```
# We use the predict() on the model to predict the output
pred=DT.predict(X_test)
# for classification we use accuracy and F1 score
print(accuracy_score(y_test,pred))
print(f1_score(y_test,pred))
# for regression we use R2 score and MAE(mean absolute
error)
# all other steps will be same as classification as shown above
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score
print(mean_absolute_error(y_test,pred))
print(mean_absolute_error(y_test,pred))
```

- Now that we know the basic steps for Classification and Regression, let's learn about the top methods for Classification and Regression that you can use in your ML systems. These methods will simplify your ML programming.

☞ Case Study : Use IRIS dataset from Scikit and apply data preprocessing methods

Following is an example to load iris dataset :

```
fromsklearn.datasetsimportload_iris
iris =load_iris()
X =iris.data
y =iris.target
feature_names=iris.feature_names
target_names=iris.target_names
print("Feature names:",feature_names)
print("Target names:",target_names)
print("\nFirst 10 rows of X:\n", X[:10])
```

☞ Output

```
Feature names: ['sepal length (cm)', 'sepal width (cm)', 'petal
length (cm)', 'petal width (cm)']
Target names: ['setosa' 'versicolor' 'virginica']
First 10 rows of X:
[
 [5.1 3.5 1.4 0.2]
 [4.9 3. 1.4 0.2]
 [4.7 3.2 1.3 0.2]
 [4.6 3.1 1.5 0.2]
 [5. 3.6 1.4 0.2]
 [5.4 3.9 1.7 0.4]
 [4.6 3.4 1.4 0.3]
 [5. 3.4 1.5 0.2]
 [4.4 2.9 1.4 0.2]
 [4.9 3.1 1.5 0.1]
]
```

☞ Splitting the dataset

To check the accuracy of our model, we can split the dataset into two pieces-**a training set** and **a testing set**. Use the training set to train the model and testing set to test the model. After that, we can evaluate how well our model did.

☞ Example

- The following example will split the data into 70:30 ratio, i.e. 70% data will be used as training data and 30% will be used as testing data. The dataset is iris dataset as in above example.

```
fromsklearn.datasetsimportload_iris
iris =load_iris()

X =iris.data
y =iris.target

fromsklearn.model_selectionimporttrain_test_split

X_train,X_test,y_train,y_test=train_test_split(
  X, y,test_size=0.3,random_state=1
)

print(X_train.shape)
print(X_test.shape)

print(y_train.shape)
print(y_test.shape)
```

☞ Output

```
(105, 4)
(45, 4)
(105,)
(45,)
```

As seen in the example above, it uses **train_test_split()** function of scikit-learn to split the dataset. This function has the following arguments :

- **X, y :** Here, **X** is the **feature matrix** and y is the **response vector**, which need to be split.
- **test_size :** This represents the ratio of test data to the total given data. As in the above example, we are setting
- **test_data = 0.3** for 150 rows of X. It will produce test data of 150*0.3 = 45 rows.
- **random_size :** It is used to guarantee that the split will always be the same. This is useful in the situations where you want reproducible results.

☞ Train the Model

Next, we can use our dataset to train some prediction-model. As discussed, scikit-learn has wide range of **Machine Learning (ML) algorithms** which have a consistent interface for fitting, predicting accuracy, recall etc.

☞ Example

- In the example below, we are going to use KNN (K nearest neighbors) classifier.
- Don't go into the details of KNN algorithms, as there will be a separate chapter for that. This example is used to make you understand the implementation part only.

```
fromsklearn.datasetsimportload_iris
iris =load_iris()
X =iris.data
y =iris.target
fromsklearn.model_selectionimporttrain_test_split
X_train,X_test,y_train,y_test=train_test_split(
  X, y,test_size=0.4,random_state=1
)
fromsklearn.neighborsimportKNeighborsClassifier
fromsklearnimport metrics
classifier_knn=KNeighborsClassifier(n_neighbors=3)
classifier_knn.fit(X_train,y_train)
y_pred=classifier_knn.predict(X_test)
# Finding accuracy by comparing actual response values(y_test)with predicted response value(y_pred)
print("Accuracy:",metrics.accuracy_score(y_test,y_pred))
# Providing sample data and the model will make prediction out of that data

sample =[[5,5,3,2],[2,4,3,5]]
preds=classifier_knn.predict(sample)
pred_species=[iris.target_names[p]for
inpreds]print("Predictions:",pred_species)
```

☞ Output

```
Accuracy: 0.9833333333333333
Predictions: ['versicolor', 'virginica']
```

☞ Model Persistence

- Once you train the model, it is desirable that the model should be persist for future use so that we do not need

Unit IV End Sem.

to retrain it again and again. It can be done with the help of **dump** and **load** features of ***joblib*** package.

- Consider the example below in which we will be saving the above trained model (classifier_knn) for future use –

```
from sklearn.externals import joblib
joblib.dump(classifier_knn, 'iris_classifier_knn.joblib')
```

- The above code will save the model into file named iris_classifier_knn.joblib. Now, the object can be reloaded from the file with the help of following code –

```
joblib.load('iris_classifier_knn.joblib')
```

☞ Preprocessing the Data

- As we are dealing with lots of data and that data is in raw form, before inputting that data to machine learning algorithms, we need to convert it into meaningful data.
- This process is called preprocessing the data. Scikit-learn has package named **preprocessing** for this purpose. The **preprocessing** package has the following techniques.

☞ Binarisation

This preprocessing technique is used when we need to convert our numerical values into Boolean values.

☞ Example

```
importnumpyas np
fromsklearnimport preprocessing
Input_data=np.array(
[2.1,-1.9,5.5],
[-1.5,2.4,3.5],
[0.5,-7.9,5.6],
[5.9,2.3,-5.8]]
)
data_binarized=preprocessing.Binarizer(threshold=0.5).trans
form(input_data)
print("\nBinarized data:\n",data_binarized)
```

In the above example, we used **threshold value** = 0.5 and that is why, all the values above 0.5 would be converted to 1, and all the values below 0.5 would be converted to 0.

☞ Output

```
Binarized data:
[
 [ 1. 0. 1.]
 [ 0. 1. 1.]
 [ 0. 0. 1.]
 [ 1. 1. 0.]
]
```

☞ Mean Removal

This technique is used to eliminate the mean from feature vector so that every feature centered on zero.

☞ Example

```
importnumpyas np
fromsklearnimport preprocessing
Input_data=np.array(
[2.1,-1.9,5.5],
[-1.5,2.4,3.5],
[0.5,-7.9,5.6],
[5.9,2.3,-5.8]]
)

#displaying the mean and the standard deviation of the input
data
print("Mean =",input_data.mean(axis=0))
print("Stddeviation = ",input_data.std(axis=0))
#Removing the mean and the standard deviation of the input
data

data_scaled=preprocessing.scale(input_data)
print("Mean_removed =",data_scaled.mean(axis=0))
print("Stddeviation_removed =",data_scaled.std(axis=0))
```

☞ Output

```
Mean = [ 1.75 -1.275 2.2 ]
Stddeviation = [ 2.71431391 4.20022321 4.69414529]
Mean_removed = [ 1.11022302e-16 0.00000000e+00
0.00000000e+00]
Stddeviation_removed = [ 1. 1. 1.]
```

☞ Scaling

We use this preprocessing technique for scaling the feature vectors. Scaling of feature vectors is important, because the features should not be synthetically large or small.

☞ Example

```
importnumpyas np
fromsklearnimport preprocessing
Input_data=np.array(
[
[2.1,-1.9,5.5],
[-1.5,2.4,3.5],
[0.5,-7.9,5.6],
[5.9,2.3,-5.8]
]
)
```

```
data_scaler_minmax=preprocessing.MinMaxScaler(feature_range=(0,1))
data_scaled_minmax=data_scaler_minmax.fit_transform(input_data)
print("\nMin max scaled data:\n",data_scaled_minmax)
```

Output

```
Min max scaled data:
[
 [ 0.48648649 0.58252427 0.99122807]
 [ 0. 1. 0.81578947]
 [ 0.27027027 0. 1. ]
 [ 1. 0.99029126 0. ]
]
```

Normalisation

We use this preprocessing technique for modifying the feature vectors. Normalisation of feature vectors is necessary so that the feature vectors can be measured at common scale. There are two types of normalisation as follows :

L1 Normalisation

It is also called Least Absolute Deviations. It modifies the value in such a manner that the sum of the absolute values remains always up to 1 in each row. Following example shows the implementation of L1 normalisation on input data.

Example

```
importnumpyas np
fromsklearnimport preprocessing
Input_data=np.array(
[
[2.1,-1.9,5.5],
[-1.5,2.4,3.5],
[0.5,-7.9,5.6],
[5.9,2.3,-5.8]
]
)
data_normalized_l1    =preprocessing.normalize(input_data, norm='l1')
print("\nL1 normalized data:\n", data_normalized_l1)
```

Output

```
L1 normalized data:
[
 [ 0.22105263 -0.2 0.57894737]
 [-0.2027027 0.32432432 0.47297297]
 [ 0.03571429 -0.56428571 0.4 ]
 [ 0.42142857 0.16428571 -0.41428571]
]
```

L2 Normalisation

Also called Least Squares. It modifies the value in such a manner that the sum of the squares remains always up to 1 in each row. Following example shows the implementation of L2 normalisation on input data.

Example

```
importnumpyas np
fromsklearnimport preprocessing
Input_data=np.array(
[
[2.1,-1.9,5.5],
[-1.5,2.4,3.5],
[0.5,-7.9,5.6],
[5.9,2.3,-5.8]
]
)
data_normalized_l2    =preprocessing.normalize(input_data, norm='l2')
print("\nL1 normalized data:\n", data_normalized_l2)
```

Output

```
L2 normalized data:
[
 [ 0.33946114 -0.30713151 0.88906489]
 [-0.33325106 0.53320169 0.7775858 ]
 [ 0.05156558 -0.81473612 0.57753446]
 [ 0.68706914 0.26784051 -0.6754239 ]
```


Chapter Ends....

❑❑❑

UNIT V

CHAPTER 5

Big Data Analytics and Model Evaluation

Syllabus Topics

Clustering Algorithms : K-Means, Hierarchical Clustering, Time-series analysis. Introduction to Text Analysis: Text-preprocessing, Bag of words, TF-IDF and topics. Need and Introduction to social network analysis, Introduction to business analysis. Model Evaluation and Selection: Metrics for Evaluating Classifier Performance, Holdout Method and Random Sub sampling, Parameter Tuning and Optimization, Result Interpretation, Clustering and Time-series analysis using Scikit-learn, sklearn. metrics, Confusion matrix, AUC-ROC Curves, Elbow plot.

Unit V End Sem.

5.1 CLUSTERING

UQ. What is clustering ? (SPPU – Q. 3(a), Aug. 18, 2 Marks)

- From a basic standpoint, k-means finds observations that share important characteristics and classifies them together into clusters.
- Clustering is an example of unsupervised learning. Cluster algorithm can be used to segment data as in classification algorithm. Classification models are used to segment data based on previously defined classes mentioned in the target, whereas 'Clustering Models' do not use any target.
- We use clustering when we want to explore data. 'Clustering algorithm' are mainly used for natural groupings. There are different categories of clustering. Clustering categories are :

1. **Hierarchical clustering :** It identifies the cluster within the cluster; e.g. inside sport news, there could be news as 'Base-ball sport', news on 'Basket Ball', news on 'Lawn-Tennis' etc.
2. **Partitional Clustering :** P.C. creates a fixed number of clusters. The K-means clustering algorithm belongs to this category.

5.2 K-MEANS

UQ. Explain k-means clustering algorithm with use cases. (SPPU - Q. 3(a), Aug. 18, Q. 2(b), Dec. 18, Q. 4(b), Oct. 19, 5 Marks)

K-means clustering is a method of vector quantization. It originates from signal processing. And it aims to partition n observations into K clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

5.2.1 Steps in K-means Algorithm

1. First we choose the final number (required) of clusters.
2. Examine each element in the population and assign it to one of the clusters depending on the minimum distance.
3. Each time a new element is added to the cluster, the centroids position is recalculated. This process is performed until all the elements are grouped into the required number of clusters.

Centroid : Centroid is a point which represents the mean of the parameter values of all the points in the cluster.

Remark : A cluster refers to a collection of data points aggregated together because of certain similarities.

5.2.2 Objective of K-means

- The objective of K-means is simple : Group similar data points together and discover underlying patterns.
- To achieve this objective, K-means looks for a fixed number (K) of cluster in a dataset.
- We define a target number K, which refers to the number of centroids we need in the dataset. A centroid is the imaginary or real location representing the centre of the cluster.
- Each data point is allocated to each of the cluster by reducing the in-cluster sum of squares.
- In other words, the K-means algorithm identifies 'K' number of centroids and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.
- The 'means' in the K-means refers to averaging of the data, that is finding the centroid.

5.2.3 Working of K-means Algorithm

UQ. How k-means algorithm works? (SPPU - Q. 3(a), Dec. 19, 5 Marks)

- To process the learning data, the K-means algorithm in data-mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (i.e. repetitive) calculations to optimise the position of the centroids.
- It halts creating and optimising clusters when either :
  (i) the centroids have stabilised there is no change in their values because the clustering is successful,
  (ii) the defined number of iterations has been achieved.

Remark : K- means clustering algorithm computes the centroids and iterates until we find optimal centroid. In

this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centriod would be minimum.

5.2.4 Use of K-means

- The K-means cluster algorithm is used to find groups which have not been explicitly labelled in the data. This can be used to confirm business assumptions about what type of groups exist or to identify unknown groups in complex data.
- K-means clustering is an extensively used technique for data cluster analysis.
- But its performance is usually not as competitive as those of other sophisticated clustering techniques because slight variations in the data could lead to high variance.
- Furthermore, clusters are assumed to be spherical and evenly sized, something which may reduce the accuracy of the K-means clustering python-results.

5.2.5 Advantages of K-means

(1) K-means quarantees convergence.

(2) It can warm-start the position of centroids.

(3) It adapts easily to new examples.

(4) It can generalise to clusters of different shapes, sizes, such as elliptical clusters.

Remark : K-means performance is measured by convergence rate and by the Sum of Squared Error (SSE).

5.2.6 Disadvantages of K-means

UQ. What are its drawbacks?
(SPPU - Q. 2(b), Dec. 18, 2 Marks)

(1) It is difficult to predict K-value.

(2) It does not work better with global cluster.

(3) Different initial patterns can result in different final clusters.

5.2.7 Does K-means Deep Learning ?

- K-means clustering is the unsupervised machine learning algorithm that is part of a much deep pool of data techniques and operations in the realm of Data Science.
- It is the fastest and most efficient algorithm to categorise data points into groups even when very little information is available about data.

5.2.8 K-means Clustering in Image Processing

- K-means clustering analysis is an unsupervised algorithm and it is used to segment the interest area from the background. So subtracting cluster is used to generate the initial centres and these centres are used in K-means algorithm for the segmentation of image.
- **Remark :** K-means clustering analysis can be significantly improved by using a better initialisation technique and by repeating (i.e. restarting) the algorithm when the data has overlapping clusters, K-means can be improved by the results of the initialisation technique.

5.2.9 Better Models than K-means

- Gaussian Mixture Models (GMMs) give us more flexibility than K-means With GMMs we assume that, the data points are Gaussian distributed; this is a less restrictive than saying that they are circular by using the mean.
- Another model is : 'Bisector K-means'. It is more efficient when K is large. For K-means algorithm, the computation involves every data point of the data set and K-centroids. But 'Bisecting K-means' produces clusters of similar sizes, while K-means is known to produce clusters of widely different sizes.

Unit V End Sem.

5.2.10 Method of using K-means

- **Step 1 :** Choose the number of clusters 'K'.
- **Step 2 :** Select 'k' random points from the data as centroids.
- **Step 3 :** Assign all the points to the closest cluster centroid.

- **Step 4 :** Recompute the centroids of newly formed clusters.
- **Step 5 :** Repeat steps 3 and 4, to get K-means.

Remark : A lower within-cluster variation is an indicator of (a good clustering) compactness. The different indices for evaluating the compactness of clusters are based on distance measures such as the clusterwise distances between observations.

UEX. 5.2.1 SPPU - Q. 4(b), Aug. 18, 6 Marks

Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Use the above data and group them using k-means clustering algorithm. Show calculation of centroids.

☑ **Soln. :** K Means Clustering requires the value of K as inputs. For this example, we are considering the value of K as 2

☞ **Iteration 1**

1.0 : Initialize cluster centroid

In this example, the value of K is considered as 2. Cluster centroids are initialized with the first 2 observations.

Cluster	Initial Centroid	
	Height	Weight
K1	185	72
K2	170	56

- **Step 1 : Calculate Euclidean Distance between each of these 2 cluster centroids and each of the observations**

Euclidean is one of the distance measures used on the K Means algorithm. Euclidean distance between each of the observations and initial cluster centroids 1 and 2 is calculated.

$$\text{Euclidean Distance} = \sqrt{(X_H - H_1)^2 + (X_W - W_1)^2}$$

X_H : Observation value of variable Height

X_1 : Centroid value of Cluster 1 for variable Height

X_W : Observation value of variable weight

W_1 : Centroid value of cluster 1 for variable weight

Distance Calculations : We will use squared Euclidean Distance for the assignment.

Obs	Height	Weight	Squared Euclidean Distance-Centroid 1	Squared Euclidean Distance-Centroid 2
1	185	72	(185-185) **2+(72-72) **2	(170-185) **2+(56-72) **2
2	170	56	(185-170) **2+(72-56) **2	(170-170) **2+(56-56) **2
3	168	60	(185-168) **2+(72-60) **2	(170-168) **2+(56-60) **2
4	179	68	(185-179) **2+(72-68) **2	(170-179)**2+(56-68)**2
5	177	62	(185-177) **2+(72-62) **2	(170-177)**2+(56-62)**2
6	188	77	(185-188) **2+(72-77) **2	(170-188)**2+(56-77)**2
7	180	71	(185-180) **2+(72-71) **2	(170-180)**2+(56-71)**2

Obs	Height	Weight	Squared Euclidean Distance-Centroid 1	Squared Euclidean Distance-Centroid 2
8	180	52	(185-180) **2+(72-70) **2	(170-180)**2+(56-52)**2
9	183	84	(185-183) **2+(72-84) **2	(170-183)**2+(56-84)**2
10	180	88	(185-180) **2+(72-88) **2	(170-180)**2+(56-88)**2
11	180	67	(185-180) **2+(72-67) **2	(170-180)**2+(56-67)**2
12	177	76	(185-177) **2+(72-76) **2	(170-177)**2+(56-76)**2

Step 2 : Cluster Assignment

Obs	Height	Weight	Square Euclidean Distance-Centroid		Cluster Assignment
			1	2	
1	185	72	0.000	481	1
2	170	56	481.000	0	2
3	168	60	433.000	20	2
4	179	68	52.000	225	1
5	177	62	164.000	85	2
6	188	77	34.000	765	1
7	180	71	26.000	325	1
8	180	52	425.000	953	1
9	183	84	148.000	953	1
10	180	88	281.000	1124	1
11	180	67	50.000	221	1
12	177	76	80.000	449	1

Step 3 : Update Cluster Centroid

Cluster	Initial Centroid	
	Height	Weight
K_1	181.5	75.375
K_2	173.75	57.5

Has the cluster assignment changed?

Yes, so continue for the next iteration.

Unit V End Sem.

Iteration 2

Step 4 : Distance Calculation

Obs	Height	Weight	Square Euclidean Distance - Centroid	
			1	2
1	185	72	23.641	336.8125
2	170	56	507.641	16.3125
3	168	60	418.641	39.3125
4	176	68	60.641	137.8125
5	177	62	199.141	30.8125
6	188	77	44.891	583.3125
7	180	71	21.391	221.3125
8	180	52	548.641	69.3125
9	183	84	76.641	787.8125
10	180	88	161.641	969.3125
11	180	67	72.391	129.3125
12	177	76	20.641	352.8125

Step 5 : Cluster Assignment

Obs	Height	Weight	Square Euclidean Distance-Centroid		Cluster Assignment	Previous Assignment
			1	2		
1	185	72	23.641	336.8125	1	1
2	170	56	507.641	16.3125	2	2
3	168	60	418.641	39.3125	2	2
4	179	68	60.641	137.8128	1	1
5	177	62	199.141	30.8125	2	2
6	188	77	44.891	583.3125	1	1
7	180	71	21.391	221.3125	1	1
8	180	52	548.641	69.3125	2	2
9	183	84	76.641	787.8125	1	1
10	180	88	161.641	969.3125	1	1
11	180	67	72.391	129.3125	1	1
12	177	76	20.641	352.8125	1	1

Step 6 : Cluster Centroids

Cluster	Initial Centroid	
	Height	Weight
K_1	181.5	75.375
K_2	173.75	57.5

5.2.11 DBScan Clustering (Density Based Spatial Clustering of Applications with Noise)

DBScan is a density based cluster algorithm. The key fact of this algorithm is that the neighbourhood of each point in a cluster which is within a given Radius (R) must have a **minimum** number of points say (M). This algorithm is extremely efficient in detecting outliers and handling noise.

5.2.12 Algorithm of DBScan

1. The type of each point is to be determined. Each data point in our dataset may be either of the following :

   Core point : A data point is a core point if there are at least M points in its neighbourhood, i.e.; within the specified radius (R).

   Border point : A data point is classified as a Border point if

   (i) its neighbourhood contains less than M data points, or

   (ii) it is reachable from some core point, i.e.; it is within R-distance from a core point.

2. **Outlier point :** An outlier point is a point that is not a core point, and also , is not close enough to be reachable from a core point.
3. The outlier points are to be eliminated.
4. Core points that are neighbours are to be connected and put in the same cluster.
5. The border points are assigned to each cluster.

5.2.13 Difference Between K-means and DBScan

Sr. No.	K-means clustering	DBScan
1.	K-means clustering is sensitive to the number of clusters specified.	Number of clusters not to be specified.
2.	K-means clusters is more efficient.	DBScan cannot efficiently handle high dimensional data sets.
3.	K-means cluster does not work well with outliers and noisy database.	DBScan can efficiently handle outliers and noisy datasets.
4.	In the domain of anomaly detection, this algorithm causes problems as anomalous points will be assigned to the same cluster as 'normal' data points.	DBScan on the other hand, locates region of high density that we separated from one another by region of low density.
5.	It requires one parameter, i.e. number of clusters (K), and varying densities of the data points does not affect K-means cluster algorithm.	It requires two parameters : Radius (R) and minimum points (M). 'R' determines a chosen radius such that if it includes enough points within it, it is a dense area. 'M' determines the minimum number of data points required in a neighbourhood to be defined on a cluster.
		DBScan cluster does not work very well for sparse datasets or for data points with varying density.

Unit V End Sem.

5.3 HIERARCHICAL CLUSTERING (H.C.)

- Hierarchical clustering is also called as Hierarchical Cluster Analysis (H. C. A.)
- It is an algorithm that groups similar objects into groups called clusters. The end point is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.
- Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom, e.g., all files and folders on the hard disk are organised in a hierarchy. There are two types of hierarchical clustering :
  - **(1) Divisive (Top-down) :** Divisive H.C. is also known as DIANA (Divisive Analysis) and it works in a top-down manner.
  - **(2) Agglomerative (Bottom-up) :** (AGNES – Ag-Nestives) . It works in Bottom-up manner.
- H.C. is the most popular and widely used method to analyse social network data. In this method, nodes are compared with one another based on their similarity. Larger groups are built by joining groups (of nodes) based on their similarity.

5.3.1 Use of Hierarchical Clustering

H. C. Starts by treating each observation as a separate cluster. Then it repeatedly executes the following two steps:

(1) Identify the two clusters that are closest together and

(2) merge the two most similar clusters.

This iterative process continues until all clusters are merged together.

Remark :

(1) In K-means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While the results are responsible in H.C. K-means is found to work well when the shape of the cluster is hyper spherical, (like circle in 2D, sphere in 3D).

(2) H. C. Groups data over a variety of scales by creating a cluster tree or dendrogram.

(3) H.C. A. is another unsupervised machine learning approach for grouping unlabeled datasets into clusters.

(4) H.C. is **more flexible** and has fewer hidden assumptions about the distribution of the underlying data.

With K-means clustering, one needs to have a sense ahead-of-time what the desired number of clusters is.

5.3.2 Advantages of H.C.

(1) No apriori information about the number of clusters required.

(2) Easy to implement and gives best result in some cases.

  (i) Algorithm can never undo what was done previously.

  (ii) Time complexity of order O $[n^2 \log n]$ is required, where n is the number of data points.

(3) Hierarchical clustering methods summarise the data hierarchy, i.e. they construct a number of local data partitions that are eventually nested. The clustering outcome depends on selected linkage strategy (single, complete, average, centroid) and the similarity measure being considerd.

5.3.3 Hierarchical Methods in Classification

There are two main methods:

(1) a flat classification that refers to the standard binary or multi-class methods, and

(2) hierarchical classification where the classes are classified at each level of a defined dendrogram.

Agglomerative hierarchical classification is the most common type of hierarchical classification used to group objects in clusters based on their similarity. It is also known as AGNES.

5.3.4 Advantages and Disadvantages of H.C.

(I) Advantages of H.C

(1) There is clear chain of command among the centroids,

(2) A clear path of advancement,

(3) There is specialisation due to iterative process

(II) Disadvantages of H.C

(1) Poor flexibility,

(2) Communication barriers

(3) Organisational disunity.

5.4 TIME-SERIES ANALYSIS (T.S.A.)

GQ. Write a short note on T.S.A ? **(2 Marks)**

- In Mathematics, a time-series is a series of data points indexed in time order. Generally, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data.
- T.S.A. helps organisations understand the underlying causes of trends or systematic patterns over time. Using data visualisations, business users can see seasonal trends and dig dipper into why these trends occur. With modern analytics platforms, these visualisations can go far beyond line graphs.

5.4.1 Examples of TSA

Field	Example Topics
Epidemology →	Disease rates, mortality rates, mosquito population
Medicine →	Blood pressure tracking, weight tracking, cholesterol measurements, heart-rate monitoring.
Physical sciences →	Global temperatures, monthly sunspot, observations, pollution levels

5.4.2 Method of T.S.A.

It is given as :

- **Step 1 :** Visualise the time-series. It is essential to analyse-the trends prior to building any kind of time-series model.
- **Step 2 :** Stationarise the series
- **Step 3 :** Find optimal parameters
- **Step 4 :** Build ARIMA model.
- **Step 5 :** Make Predictions

Remark : ARIMA Model : It is Autoregressive Integrated moving average. In statistics and econometrics, and in particular in time series analysis, an ARIMA model is a generalisation of an autoregressive moving average model. Both of these models are fitted to time series data either to better understand the data or predict future points in the series.

5.4.3 Four Components of Time-series Analysis

1. **Secular trend**, which describes the movement along the term.
2. **Seasonal variations,** which represent seasonal changes,
3. **Cyclical fluctuations**, which correspond to periodical but not seasonal variations.
4. **Irregular variations**, which are other non-random sources of variations of series.

5.4.4 Methods of Time-Series

Time series is a sequence of time-based data points collected at specific intervals of a given phenomenon that undergoes changes over time. It is indexed according to time.

The four variations to time series are :

(1) Seasonal variations,

(2) Trend variations,

(3) Cyclical variations,

(4) Random variations

5.4.5 Time-series Plot

- A time-plot is basically a line plot showing the evolution of the time-series over time.
- We can use it as the starting point of the analysis to get some basic understanding of the data, e.g., in terms of trend /seasonality/ outliers etc.

Unit V End Sem.

5.4.6 Use of Time-Series

Time-series is used in statistics, Signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering, and largely in any domain of applied science and engineering which involve understanding of the data.

5.4.7 Components of Time-series

An observed time-Series can be decomposed into three components :

(i) The trend (long term direction)

(ii) The seasonal (systematic, calendar related movements)

(iii) The irregular (unsymmetric, short term fluctuations).

5.4.8 Need of Time-Series

T. S. analysis can be useful to see how a given asset, security, or economic variable changes over time. It can also be used to examine how the changes associated with the chosen data point compare to shifts in other variables over the same time period.

5.4.9 Characteristics of Time Series Analysis

First step is to plot the data on a graph. When plotted, many time series exhibit one or more of the following features :

(i) Trend,

(ii) Seasonal and Nonseasonal cycles,

(iii) Pulses and steps,

(iv) Outliers.

(i) **Trend in a Time Series**

- Trend is a pattern in data that shows the movement of a series to relatively higher or lower values over a long period of time.
- In other words, a trend is observed when there is an increasing or decreasing slope in the time series. Trend usually happens for some time and then disappears, it does not repeat.

(ii) Identifying whether there is a seasonality component in your time series problem is subjective. The simplest approach to determining if there is an aspect of seasonality is 'to plot and review the data', at different scales and with the addition of trend lines.

(iii) Outliers in time series data are values that significantly differ from the patterns and trends of the other values in the time series. For example, large numbers of online purchase around holidays or high number of accidents during heavy rainstorms may be detected as outliers in their time series.

5.4.10 Advantages of TSA

(1) The most effective form of TSA is to simply plot the data on a line chart. With this step, there will be no longer any doubts as to whether or not sales truly peak before Diwali or dip in February.

(2) Time series is an effective tool of forecasting. Time series forecasting is a technique in machine learning, which analyse data and the sequence of time to predict future events.

5.4.11 Limitations of Time Series

(1) There are some weak points in Time Series Analysis. It includes problems with generalisation from a single study, difficulty in obtaining appropriate measures and problems with accurately identifying the correct model to represent the data.

(2) T.S. data presents one of the most difficult analytical challenges : You typically have the least amount of data to work with, while needing to inform some of the most important decisions.

5.5 TEXT DATA (PREPROCESSING)

GQ. Write a short note on Data Preprocessing? **(2 Marks)**

- Text Data originate from languages. The spoken languages are mostly converted to data as audio, or text. From technical point of view, text data may be found in various forms like 'news-articles', 'books', 'web pages', 'comments', 'software codes', 'computer system logs' and many more.
- These text-data have their own structural characteristics that are defined by the language. In the domain of science the text representation and text similarity are potential concerns of Research domain.

- Text representation are mostly used in :

(1) **Machine Translation :** Automatically translating text from one language to another.

(2) **Document Clustering :** Groping text documents based on the structural and/or semantic similarity.

(3) **Topic Detection :** Identifying the topic of a large text corpus.

(4) Text summarisation.

(5) Question-answering.

and also in the fields like information retrieval and document ranking.

5.5.1 Representation Technique

- There are several representation techniques. Some of them are 'one-hot encoding', 'n-gram model', 'Bag-of-words' and 'neural word embedding'. These techniques have their own strengths and weaknesses. And we have to choose the best option.
- We note that representation of the data is the way we shall be presenting our text data to our algorithm. So we must be careful to choose the most convenient data representation technique.

5.5.2 Bag-of-Words Model

- It is the collection of words. This bag of words does not count grammar or the positioning or structure of the words in the text. It counts only frequencies of words in the target text and put that word into a **bag**.
- The frequencies of words appearing in a text is the feature that is used on bag-of-words model.

5.5.3 Bow for a Text Corpus

We create a bag of word model for the following text :

Consider a set (P) of prepositions : our aim is to generate a numerical representation for the following prepositions :

Prep. 1 : Ashok is more intelligent than Gopal

Prep. 2 : Ramesh likes to drink milk at night.

Prep. 3 : the moon is shining in the sky.

Prep. 4 : the night sky is full of stars.

▶ **Step I :** Preprocess Data :

First, we lowercase all the words to avoid the same words as distinct words.

We remove stop-words like 'is', 'at' 'in' etc. And we stick to only lower-casing.

▶ **Step II :** After the above preprocessing step, our prepositions will be,

Prep. 1 : Ashok is more intelligent than Gopal

Prep. 2 : Ramesh likes to drink milk at night.

Prep. 3 : the moon is shining in the sky.

Prep. 4 : the night sky is full of stars.

Now, we create our 'bag' of words as follows :

Now, we create our 'bag', of words as follows :

Term	Frequency
the	3
is	3
night	2
sky	2
night	2
ashok	1
intelligent	1
than	1
gopal	1
ramesh	1
likes	1
to	1
drink	1
milk	1
at	1
moon	1
shining	1
in	1
full	1
of	1
stars	1
more	1

Bag of words for selected corpus.

Unit V End Sem.

5.5.4 Tf-Idf for Text Representation

- Tf-Idf (Term frequency –Inverse document frequency) is a bag of word model which is very powerful in capturing the most important words in the text.
- The concept of Tf-Idf can be understood by the term frequency (Tf) and inverse preposition (document) frequency (Idf). The collective representation generated from Tf and Idf representations is called Tf-Idf. Term frequency is a function of the term t and document d.
- Tf representation gives a degree or presence of the terms in a document and Inverse document frequency (Idf) measures the uniqueness of a term to a document.
- For example, consider documents (Preposition) 1 and 4. The only term 'is', is common for the two documents. Hence 'is', dimension will show the degree of **similarity in** document 1 and 4. But are the documents similar ?
- Actually no. Here Idf comes to our help.
- The idea behind Idf is that a term which appears in a majority of documents does not add special information to the target.
- Inverse document frequency is defined for each term in BOW. It has the ability to emphasize the uniqueness or the importance of a term in BOW. It a term appears in a majority of the documents, it is not unique.
- A unique or rare word will appear in less numbers of documents and its Idf value will be larger.
- We note that the terms like in, of, to do not add a specific meaning to the document. We keep these terms in BOW just for simplicity. But in practical case, we omit these 'stop words' in the prepossessing stages.

5.5.5 Meaning of Tf-Idf

- We observe that Tf-Idf is nothing but the multiplication of Tf (t, d) and Idf (t, D).
- A term which can be seen almost everywhere in a target document and cannot be found in other documents in the corpus has high degree of uniqueness on the target document.

5.5.6 Shortcomings of Tf-Idf

- Even though, Tf-Idf is such a powerful technique to represent text-data, it is still a BOW model.
- Tf-Idf does not count the positioning of words in the document. This leads Tf-Idf to give wrong interpretation on the text data.

For example,

(P(1)) : Ashok is more intelligent than Gopal.

(P(2)) : Gopal is more intelligent than Ashok.

- have same set {'ashok', 'more', 'intelligent', 'than', 'Gopal'} of word frequencies, but their meanings are different. But, Tf-Idf as a BOW text representation model gives the same representation for both documents.
- Another major shortcoming in Tf-Idf model is it consumes a lot of memory and processing power.
- Thus the main drawback is Tf-Idf does, not count the positioning of words when generating **vector representation.**

5.6 SOCIAL NETWORK ANALYSIS

- Social Network Analysis (SNA) is the process of investigating social structures through the use of networks and graph theory.
- It characterises networked structures in terms of nodes (individual actors, people or things within the network) and the ties, edges or links (relationship or interactions) that connect them.

5.6.1 Social Network Measures

- SNA is the mapping and measuring of relationships and flows between people, groups, organisations, computers, URLS, and other connected information entities.
- The nodes in the network are the people and groups while the links show relationships or flows between the nodes.

5.6.2 Importance of Social Network Analysis

- SNA can provide insights into social influences within teams, and identify cultural issues.
- SNA has been used as a strategic approach to team building, and to understand how team building can change the dynamics of an organisation's social network.

5.6.3 Steps in S.N.A.

There are main six generic steps of the SNA process :

1. Problem definition.
2. Data gathering and preparation.
3. Social network modelling.
4. Knowledge extraction.
5. Evaluation; interpretation and deployment

5.6.4 S.N.A. in Research

- S.N.A. is the study of structure, and how it influences health, and it is based on theoretical constructs of sociology and mathematical foundations of graph theory.
- Structure refers to the regularities in the patterning of relationships among individuals, groups and organisations.

5.6.5 Focus of S.N.A.

The aim of Social Network Analysis is to understand a community by mapping the relationships that connect them as a network, and then trying to draw out key individuals, groups within the network (components) associations between the individuals.

5.6.6 Collecting Data by S.N.A.

- The Most common data collection methods used in SNA are surveys and interviews.
- A survey should include questions regarding the background of the respondent and a way for them to provide information on connections.

5.6.7 Benefits of S.N.A.

The benefits or objective are :

1. Minimise production delay, interruption and conflicts,
2. Minimisation of total project cost
3. Trade-off between Time and cost of project
4. Minimisation of Total Project duration.
5. Minimisation of Idle Resources

5.6.8 Social Network Analysis in Data Mining

- S.N.A. is the study of social networks to understand their structure and behaviour.
- Data mining based techniques are useful for analysis of social network data, especially for large datasets that cannot be handled by traditional methods.

5.6.9 Examples of S.N.A.

Examples of social structures commonly visualised through social network analysis include 'social media networks, information circulation, friendship and acquaintance networks', business networks, knowledge networks, difficult working relationships, social networks, collaboration graphs, kinship, etc.

5.7 BUSINESS ANALYSIS

GQ. Write a short note Business Analysis? **(2 Marks)**

- Business analysis is a research discipline of identifying business needs and determining solutions to business problems. Solutions often include a software-systems development component, but may also consist of process improvements, organisational change or strategic planning and policy development.
- The person who carries out this task is called a business analyst or BA.
- Business analysts do not work solely on developing software systems. But work across the organisation, solving business problems in consultation with business stake holders. While most of the work that business analysts do today relate to software development, this derives from the ongoing massive changes business all over the world are experiencing in their attempts to digitise.

Unit V End Sem.

- Although there are different role definitions, depending upon the organisation, there does seem to be an area of common ground where most business analysts work.

5.7.1 The Responsibilities of B.As.

(i) To evaluate actions to improve the operation of a business system. Again, this may require an examination of organisational structure and staff development needs, to ensure that they are in line with any proposed process redesign and IT system development.

(ii) To document the business requirements for the IT system support using appropriate documentation standards.

(iii) The core business analyst role can be defined as an internal consistency role that has the responsibility for investigating business situations, identifying and evaluating options for improving business systems, defining requirements and ensuring the effective use of information systems in meeting the needs of the business.

(iv) To investigate business systems, taking a holistic view of the situation. This may include examining elements of the organisation structures and staff development issues as well as current processes and IT systems.

5.8 METRICS FOR CLASSIFIER PERFORMANCE

The key classification metrics are :

Accuracy, Recall, Precision and F1-score.

To measure the performance of a classifier we simply measure the number of correct decisions the classifier makes, and divide by the total number of test examples. The result is the accuracy of the classifier.

5.8.1 Performance of Classifier

- The evaluation measure used for evaluating the performance of the classifier is 'Area under Curve (AUC)'. It is one of the most widely used metrics for evaluation. It is used for binary classification problem.
- AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.

5.8.2 Good F1 Score

- A good F1 score means that one has low false positives and low false negatives, so one can correctly identify real threats and you are not disturbed by false alarms.
- An F1 score is perfect when it is 1 and the model is total failure when it is 0.

5.8.3 Evaluation Metrics

- An evaluation metric quantifies the performance of a predictive model.
- This involves training a model on a dataset, using the model to make predictions on a holdout dataset not used during training, and then comparing the predictions to the expected values in the holdout dataset.

5.8.4 Good Accuracy for a Classifier

- Most practitioners develop an intuition that large accuracy score (or conversely small error rate scores) are good, and values above 90% are great.
- Metrics like accuracy, precision, recall are good ways to evaluate classification models for balanced datasets, but if the data is imbalanced and there is a class disparity, then other methods like ROC/AUC, Gini coefficient perform better in evaluating the model performance.

5.9 HOLDOUT METHOD

GQ. Write a short note on HOLDOUT Method? **(2 Marks)**

- Holdout method is the simplest sort of method to evaluate a classifier.
- In this method, the dataset (a collection of data items or examples) is separated into two sets, called the Training set and Test set.
- A classifier performs function of assigning data items in a given collection to a target category or class.

5.9.1 Holdout method for Machine Learning

The holdout method for training machine learning model is the process of splitting the data in different splits and using one split for training the model and other splits for validating and testing the models.

5.9.2 Cross-validation Method

- 'Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample'.
- The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. Hence, the procedure is often called K-fold cross-validation.

5.9.3 Difference between Cross-Validation and Holdout Validation

- Cross-validation is usually the preferred method because it gives our model the opportunity to train on multiple train-test splits. This give us a better indication of how well our model will perform on unseen data.
- Hold-out method is dependent on just one train- test split.

5.9.4 Holdout dataset

Holdout data refers to a portion of historical, labelled data that is held out of the data sets used for training and validating supervised machine learning models. It can also be called test-data.

5.9.5 Holdout Evaluation

Holdout evaluation is an approach to out-of-sample evaluation whereby the available data are partitioned into a training set and a test-set. This provides an unbiased estimate of learning performance, in contrast to in-sample evaluation.

Purpose of a holdout-set

The purpose is to verify the accuracy of a forecast technique.

5.9.6 Disadvantages of Holdout Method

- The limitation of such a method is that the error found in the test dataset can highly depend on the observations included in the train and test dataset.
- Also if the train or test dataset are not able to represent the actual complete data then the results from the test sets can be skewed.

5.9.7 The Purpose of Holdout Validation

The holdout dataset is not used in the model training process and the purpose is to provide an unbiased estimate of the model performance during the training process. This set of data will only be used once the model has finished training with the Training dataset and validation dataset.

5.10 RANDOM SUB-SAMPLING

- Random sub-sampling, also known as monte-carlo cross validation, as multiple holdout or as repeated evaluation set is based on randomly splitting the data into subsets, whereby the size of the subsets is defined by the user'.
- The random partitioning of the data can be repeated arbitrarily often.

5.10.1 Random Subsampling in Data Mining

Random subsampling performs K data splits of the entire sample. For each data split , a fixed number of observations is chosen without replacement from the sample and kept aside as the test data.

5.10.2 Subsampling Techniques

- Subsampling is a method that reduces data size by selecting a subset of the original data.
- The subset is specified by choosing a parameter n, specifying that every n^{th} data point is to be extracted.

5.10.3 Subsampling in CIVIV

Sub-sampling is incorporated within CIVIV by adding a subsampling layer 'where each unit within the layer has a receptive field of a fixed size that is imposed on the input (feature maps from previous layer), where an operation is

performed on the pixels that are in the scope of the receptive field of the unit.

5.10.4 Subsampling and Pooling

- Average pooling calculates the average and processes that in output image.
- On the other hand, 'Subsampling chooses a pixel in the grid and replaces surrounding pixels of said grid' by the same pixel value in the output image.

5.10.5 Subsampling Layer

A pooling or subsampling layer follows a convolution layer in CNN. Its role is to down sample the output of a convolution layer along both the spatial dimensions of height and width.

5.10.6 Subsampling in Signal Processing

Subsampling is the 'process of sampling a signal with a frequency lower than twice the highest signal frequency, but higher than two times the signal bandwidth.

5.10.7 Subsampling Factor

The subsample algorithm allows one to reduce an image size by a factor of 2,4 or 8 times. For example, subsampling a 3D image with the x,y, and z dimensions of $256 \times 256 \times 32$ respectively by a factor of 2 produces a new image with x,y and z dimensions of $128 \times 128 \times 16$ respectively.

5.10.8 Downsampling

Downsamplilng is taking a random sample without replacement. Downsampling from the negative cases reduces the dataset to a more manageable size.

5.10.9 Subsampling in Statistics

Subsampling is collecting data in two or more stages at successive levels of observation. In collecting data on urban households, we might begin with a first stage of identifying a randomly selected group of cities and then, as a second stage, sample households randomly within those cities.

5.10.10 Downsampling an Image

The steps involved to downsample an image in Elements :

1. To open a photo in the Photo-Editor
2. Choose : Image-Resize → Image size → ...
3. In the document size area, redefine the dimensions and resolution
4. if we are okay with resampling your image to get the desired size, select the Resample Image, check box.

5.10.11 Chroma Subsampling

- Chroma subsampling involves the reduction of colour resolution in video signals in order to save bandwidth.
- The color component information (chroma) is reduced by sampling them at a lower rate than the brightness (luma).
- Continuous tone (natural) images are less impacted by subsampling than synthetic (computer) imagery.
- Chroma subsampling is a 'type of compression that reduces the colour information in a signal in favour of luminance data'. This reduces bandwidth without significantly affecting picture quality. This allows one to maintain picture clarity while effectively reducing the file size upto 50%.

5.11 PARAMETER TUNING OPTIMISATION

- In machine learning, hyper parameter optimisation or tuning is the problem of choosing a set of optimal hyper parameters for a learning algorithm.
- A hyper parameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.
- The same kind of machine learning model can require different constraints, weights or learning rates to generalise different data patterns. These measures are called hyperparameters, and have to be tuned so that the model can optimally solve the machine learning problem.

- Hyperparameter optimisation finds a tuple of hyper parameters that yields an optimal model which minimises a predefined loss function on given independent data.
- The objective function takes a tuple of hyperparameters and returns the associated loss. Cross-validation is often used to estimate this generalisation performance.

Grid Search

- The traditional way of performing hyerparameter optimisation is **grid-search** or a parameter sweep, which is simply on exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm.
- A grid search algorithm must be guided by some performance metric, typically measured by cross-validation on the training set or evaluation on a held-out validation set.
- Since the parameter space of a machine learner may include real-valued or unbounded value spaces for certain parameters, manually set bounds and discretisation may be necessary before applying **grid search**.

5.12 CONFUSION MATRIX

GQ. Write a short note on Confusion matrix ? **(2 Marks)**

- Here we present measures for assessing how "good" or how "accurate" the classifier is at predicting the class label of tuples.
- Before we discuss various measures, we get acquainted with some terminology.
- We can talk in terms of 'positive tuples' (tuples of the main class of interest) and 'negative tuples' (all other tuples).
- Given two classes, for example, the positive tuples may be buys... computer = yes, while the negative tuples are buys... computer = no. Suppose we use our classifier on a test set of labeled tuples. P is the number of positive tuples and IV is the number of negative tuples.

We consider four additional terms that are the "building blocks" used in computing many evaluation measures.

(i) **True positives (TP) :** These refer to the positive tuples that were correctly labelled by the classifier. Let TP be the number of true positives.

(ii) **True negatives (TN) :** These are the negative tuples that were correctly labelled by the classifier. Let TN be the number of true negatives.

(iii) **False Positives (FP) :** These are the negative tuples that were incorrectly labelled as the positive (e.g., tuples of class buys computer = no. for which the classifier predicted buys....Computer = yes). Let FP be the number of false positives.

(iv) **False negatives (FN) :** These are the positive tuples that were mislabelled as negative (e.g., tuples of class, buys. Computer = yes for which the classifier predicted buys. Computer = no). Let FN be the number of false negatives.

- These terms are summarised in the 'confusion matrix'.
- The confusion matrix is a useful tool for analysing how a classifier can recognise tuples of different classes.
- TP and TN tell us when the classifier is getting things right, while FP and FN tell us when the classifier is getting things wrong.

	Predicted class		
	Yes	No	Total
Yes	TP	FN	P
no	FP	TN	N
Total	P	N	P + IV

- Now we calculate the evaluation measures, starting with accuracy. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier, i.e.

$$\text{accuracy} = \frac{TP + TN}{P + N} \quad ...(5.12.1)$$

- We can also calculate the error rate or misclassification rate of a classifier and it is given as,

$$\text{error rate} = \frac{FP + FN}{P + N} \quad ...(5.12.2)$$

- Now, we define the measures : The sensitivity and specificity. Sensitivity is referred to as the 'true positive' (recognition rate), (i.e. the proportion of

Unit V End Sem.

positive tuples that are correctly identified), while specificity is the 'true negative rate' (i.e. the proportion of negative tuples that are correctly identified).

- These measures are given as :

$$\text{Sensitivity} = \frac{TP}{P} \quad ...(5.12.3)$$

$$\text{and specificity} = \frac{TN}{N} \quad ...(5.12.4)$$

- Thus, we can write,

$$\text{accuracy} = \frac{TP + TN}{P + N}$$

$$= \text{sensitivity}\left[\frac{P}{P+N}\right] + \text{specificity}\left[\frac{N}{P+N}\right] \quad ...(5.12.5)$$

- Thus accuracy is a function of sensitivity and specificity

5.13 ELBOW PLOT

GQ. Write a short note on ELOW PLOT? **(2 Marks)**

- In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set.
- The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

5.13.1 Elbow in a Plot

- WCSS is the 'sum of squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease.
- WCSS is ' within cluster sum of squares'.

5.13.2 To Plot the Elbow

- Elbow method is an empirical method to find the optimal number of clusters for a dataset.
- In this method, we pick a range of candidate values of K, then apply K-mean clustering using each of the values of K. Find the average distance of each point in a cluster to its centroid, and represent it in a plot.

5.13.3 Elbow Graph

- If the line chart looks like an arm, then the 'elbow' (the point of inflexion on the curve) is the best value of K.
- The 'arm' can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.

5.13.4 Distortion of an Elbow Plot

- The elbow method plots the value of the cost function produced by different values of K.
- We have seen, if K increases, the average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids.

5.13.5 Elbow-Rule

- In the case of the elbow-rule, one typically uses the percentage of unexplained variance.
- This number is 100% with zero cluster, and it decreases, initially sharply, then more modestly, as the number of cluster increases in the model.

5.13.6 Elbow method in Python

The Elbow method is a very popular technique and the idea is to run K-means clustering for a range of clusters K (let us say from 1 to 10) and for each value, we are calculating the sum of squared distances from each point to its assigned centre (distortions).

5.13.7 Elbow Point

The elbow point is calculated simply by instantiating the KneeLocator class with x, y and the appropriate curve and direction. Here, kneedle, knee and/or kneedle, elbow store the point of maximum curvature.

5.13.8 Silhouette Score

- The value of 2 and 3 for n-clusters looks to be the optimal one. The silhouette score for each cluster is above average silhouette scores. Also, the fluctuations in size is similar. The thickness of the silhouette plot representing each cluster also is a deciding point.

- A silhouette is the image of a person, animal, object or scene represented as a solid shape of a single colour, usually black, with its edges matching the outline of the subject.
- The interior of the silhouette is featureless, and the silhouette is usually presented on a light background, usually white or none at all.

5.13.9 Choosing an Elbow

First we calculate Within-Cluster Sum of Squared Errors (WSS) for different values of K, and choose the K for which WSS becomes first starts to diminsh; In the plot of WSS-versus-K, this is visible as an elbow.

5.13.10 Elbow Curve-Method

1. Perform K-means clustering with all these different values of K. For each of the K values, we calculate average distances to the centroid across all data points.
2. Plot these points and find the point where the average distance from the centroid falls suddenly. ("Elbow").

5.14 IRIS DATASET

Iris dataset is a multivariate dataset introduced by the British statistician and biologist Ronald Fisher in his 1936 paper : The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis.

5.14.1 Features of IRIS Dataset

- The Iris Dataset contains four features (length and width of sepals and petals) 'of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris Versicolor). These measures were used to create a linear discriminate model to classify the species.
- The Iris dataset is a built-in dataset in R that contains measurements on 4 different attributes (in centimeters) for 50 flowers from 3 different species.

5.14.2 Objective of Iris Dataset

- Iris data is a multivariate dataset. Four features of this dataset are –sepal length, sepal width, petal length, and petal width, in centimeters.
- Iris data is publicly available to use and is one of the most widely used data set, mostly by the beginners in the area of Data Science and Machine Learning.

5.14.3 Use of Iris Dataset and Iris Target

- The dataset is often used in data mining, classification and clustering examples and to test algorithms. Iris dataset is a part of sklearn library.
- Iris has 4 numerical features and a tri class target variable. This dataset can be used for classification as well as clustering.

5.14.4 Meaning of Iris and Database

- The iris commonly means wisdom, hope, trust and valor.
- An iris database is 'a collection of images that contain, at a minimum, the iris region of the eye.
- The images are typically collected by sensors that operate in the visible spectrum, 380-750 nm, or the near infrared spectrum (NIR), 700-900 nm.

5.14.5 Function of Iris

- The iris controls the amount of light that enters the eye by opening and closing the pupil.
- The iris uses muscles to change the size of the pupil. These muscles can control the amount of light entering the eye by making the pupil larger (dilated) or smaller (constricted).

5.14.6 Data Set Information

- The dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant.
- One class is linearly separable from the other 2; the latter are **not** linearly separable form each other.

5.14.7 Linearly Separable Dataset

One species, Iris Setosa, is 'linearly separable' from the other two. This means that we can draw a line (or a hyper plane in higher dimensional spaces) between Iris setosa samples and samples corresponding to the other two species.

5.14.8 Loading Iris Dataset into Scikit Learn

Loading dataset is : go to sklearn datasets and then import /get iris dataset and store it in a variable named iris. Scikit-learn comes with a few small standard datasets that do not require to download any file from some external website.

5.14.9 Iris in Machine Learning

The dataset is often used in data mining, classification and clustering examples and the test algorithms.

5.14.10 Popularity of Iris Dataset

- The data is open source, the accuracy is known, and because there are three classes, it allows for more than just binary classification.
- All these factors and more contribute to the popularity of Iris dataset. It is the dexterity of this dataset that makes it so popular.
- The dataset contains a set of 150 records under five attributes-sepal length, sepal width, petal length, petal width and species.

5.14.11 Iris in R-Programming

Iris is a data, frame, which is probably the most commonly used data structure in R. It is basically a table where each column is a variable and each row has one set of values for each of those variables.

5.15 RESULT INTERPRETATION

GQ. Write a short note on Result Interpretation ? **(2 Marks)**

- Results interpretation is statistical, specific and constrained. E.g. The 4 way analysis of variance showed significant interactions between the variables a,b,c and d. further analysis using test X showed that the locus of the significant effect was.
- The discussion section explains the statistical findings in the context of the research hypothesis presented and any other related research that it either supports or refutes.
- Hopefully you can provide an explanation for why your results turned out the way they did, why they agree or disagree with prior results, explanations or plans for dealing with any confounded variables. And finally plans for future research.
- Clustering methods, one of the most useful unsupervised ML methods, used to find similarity and relationship patterns among data samples. After that, they cluster those samples into groups having similarity based on features.
- Clustering determines the intrinsic grouping among the present unlabeled data, that's why it is important.
- The Scikit-learn library have **sklearn.cluster** to perform clustering of unlabeled data. Under this module scikit-leran have the following clustering methods –

K-Means

- This algorithm computes the centroids and iterates until it finds optimal centroid. It requires the number of clusters to be specified that's why it assumes that they are already known.
- The main logic of this algorithm is to cluster the data separating samples in n number of groups of equal variances by minimizing the criteria known as the inertia. The number of clusters identified by algorithm is represented by 'K.
- Scikit-learn have **sklearn. cluster. KMeans** module to perform K-Means clustering. While computing cluster centers and value of inertia, the parameter named **sample_weight** allows **sklearn. cluster. KMeans** module to assign more weight to some samples.

Affinity Propagation

- This algorithm is based on the concept of 'message passing' between different pairs of samples until convergence. It does not require the number of clusters to be specified before running the algorithm.
- The algorithm has a time complexity of the order $O(N2T)$, which is the biggest disadvantage of it.
- Scikit-learn have **sklearn. cluster. Affinity Propagation** module to perform Affinity Propagation clustering.

Mean Shift

- This algorithm mainly discovers ***blobs*** in a smooth density of samples. It assigns the datapoints to the clusters iteratively by shifting points towards the highest density of datapoints.
- Instead of relying on a parameter named **bandwidth** dictating the size of the region to search through, it automatically sets the number of clusters.
- Scikit-learn have **sklearn. cluster. MeanShift** module to perform Mean Shift clustering.

Spectral Clustering

- Before clustering, this algorithm basically uses the eigen values i.e. spectrum of the similarity matrix of the data to perform dimensionality reduction in fewer dimensions.
- The use of this algorithm is not advisable when there are large number of clusters.
- Scikit-learn have **sklearn. cluster. Spectral Clustering** module to perform Spectral clustering.

Hierarchical Clustering

- This algorithm builds nested clusters by merging or splitting the clusters successively. This cluster hierarchy is represented as dendrogram i.e. tree. It falls into following two categories –
  1. **Agglomerative hierarchical algorithms :** In this kind of hierarchical algorithm, every data point is treated like a single cluster. It then successively agglomerates the pairs of clusters. This uses the bottom-up approach.
  2. **Divisive hierarchical algorithms :** In this hierarchical algorithm, all data points are treated as one big cluster. In this the process of clustering involves dividing, by using top-down approach, the one big cluster into various small clusters.
- Scikit-learn have **sklearn.cluster. Agglomerative Clustering** module to perform Agglomerative Hierarchical clustering.

DBSCAN

- It stands for **"Density-based spatial clustering of applications with noise"**. This algorithm is based on the intuitive notion of "clusters" & "noise" that clusters are dense regions of the lower density in the data space, separated by lower density regions of data points.
- Scikit-learn have **sklearn. cluster. DBSCAN** module to perform DBSCAN clustering. There are two important parameters namely min_samples and eps used by this algorithm to define dense.
- Higher value of parameter **min_samples** or lower value of the parameter eps will give an indication about the higher density of data points which is necessary to form a cluster.

OPTICS

- It stands for **"Ordering points to identify the clustering structure"**. This algorithm also finds density-based clusters in spatial data. It's basic working logic is like DBSCAN.
- It addresses a major weakness of DBSCAN algorithm- the problem of detecting meaningful clusters in data of varying density-by ordering the points of the database in such a way that spatially closest points become neighbors in the ordering.
- Scikit-learn have **sklearn. cluster. OPTICS** module to perform OPTICS clustering.

BIRCH

- It stands for Balanced iterative reducing and clustering using hierarchies. It is used to perform hierarchical clustering over large data sets. It builds a tree named **CFT** i.e. **Characteristics Feature Tree**, for the given data.
- The advantage of CFT is that the data nodes called CF (Characteristics Feature) nodes holds the necessary information for clustering which further prevents the need to hold the entire input data in memory.
- Scikit-learn have **sklearn. cluster. Birch** module to perform BIRCH clustering.

☞ Comparing Clustering Algorithms

Following table will give a comparison (based on parameters, scalability and metric) of the clustering algorithms in scikit-learn :

Sr. No	Algorithm Name	Parameters	Scalability	Metric Used
1.	K-Means	No. of clusters	Very large n_samples	The distance between points.
2.	Affinity Propagation	Damping	It's not scalable with n_samples	Graph Distance
3.	Mean-Shift	Bandwidth	It's not scalable with n_samples.	The distance between points.
4.	Spectral Clustering	No. of clusters	Medium level of scalability with n_samples. Small level of scalability with n_clusters.	Graph Distance
5.	Hierarchical Clustering	Distance threshold or No. of clusters	Large n_samples Large n_clusters	The distance between points.
6.	DBSCAN	Size of neighborhood	Very large n_samples and medium n_clusters.	Nearest point distance
7.	OPTICS	Minimum cluster membership	Very large n_samples and large n_clusters.	The distance between points.
8.	BIRCH	Threshold, Branching factor	Large n_samples Large n_clusters	The Euclidean distance between points.

☞ K-Means Clustering on Scikit-learn Digit dataset

In this example, we will apply K-means clustering on digits dataset. This algorithm will identify similar digits without using the original label information. Implementation is done on Jupyter notebook.

```
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()
import numpy as np
from sklearn.cluster import KMeans
from sklearn.datasets import load_digits
digits = load_digits()
digits.data.shape
```

Output

```
1797, 64)
```

This output shows that digit dataset is having 1797 samples with 64 features.

Example

```
Now, perform the K-Means clustering as follows −
kmeans = KMeans(n_clusters = 10, random_state = 0)
clusters = kmeans.fit_predict(digits.data)
kmeans.cluster_centers_.shape
```

Output

```
(10, 64)
```

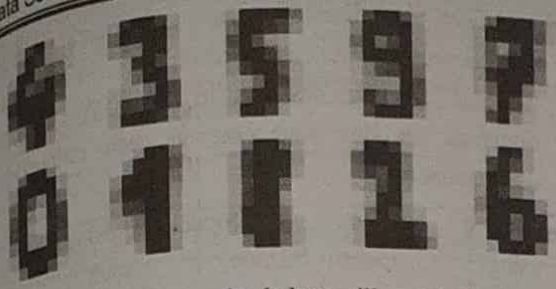
This output shows that K-means clustering created 10 clusters with 64 features.

Example

```
fig, ax = plt.subplots(2, 5, figsize = (8, 3))
centers = kmeans.cluster_centers_.reshape(10, 8, 8)
for axi, center in zip(ax.flat, centers):
axi.set(xticks = [], yticks = [])
axi.imshow(center, interpolation = 'nearest', cmap = plt.cm.binary)
```

Output

The below output has images showing clusters centers learned by K-Means Clustering.



Next, the Python script below will match the learned cluster labels (by K-Means) with the true labels found in them –

```
from scipy.stats import mode
labels = np.zeros_like(clusters)
for i in range(10):
mask = (clusters == i)
labels[mask] = mode(digits.target[mask])[0]
```

We can also check the accuracy with the help of the below mentioned command.

```
from sklearn.metrics import accuracy_score
accuracy_score(digits.target, labels)
```

Output

```
0.7935447968836951
```

☞ Complete Implementation Example

```
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()
import numpy as np

from sklearn.cluster import KMeans
from sklearn.datasets import load_digits
digits = load_digits()
digits.data.shape
kmeans = KMeans(n_clusters = 10, random_state = 0)
clusters = kmeans.fit_predict(digits.data)
kmeans.cluster_centers_.shape
fig, ax = plt.subplots(2, 5, figsize = (8, 3))
centers = kmeans.cluster_centers_.reshape(10, 8, 8)
for axi, center in zip(ax.flat, centers):
  axi.set(xticks=[], yticks = [])
  axi.imshow(center, interpolation = 'nearest', cmap =
plt.cm.binary)
from scipy.stats import mode
labels = np.zeros_like(clusters)
for i in range(10):
  mask = (clusters == i)
  labels[mask] = mode(digits.target[mask])[0]
from sklearn.metrics import accuracy_score
accuracy_score(digits.target, labels)
```

☞ What is a Time Series ?

- Time series is a sequence of observations recorded at regular time intervals.
- Depending on the frequency of observations, a time series may typically be hourly, daily, weekly, monthly, quarterly and annual.
- Sometimes, you might have seconds and minute-wise time series as well, like, number of clicks and user visits every minute etc. Why even analyze a time series? Because it is the preparatory step before you develop a forecast of the series.
- Besides, time series forecasting has enormous commercial significance because stuff that is important to a business like demand and sales, number of visitors to a website, stock price etc are essentially time series data. So what does analyzing a time series involve?
- Time series analysis involves understanding various aspects about the inherent nature of the series so that you are better informed to create meaningful and accurate forecasts.

☞ How to import time series in python?

- So how to import time series data?
- The data for a time series typically stores in .csv files or other spreadsheet formats and contains two columns: the date and the measured value.
- Let's use the read_csv() in pandas package to read the time series dataset (a csv file on Australian Drug Sales) as a pandas dataframe. Adding the parse_dates=['date'] argument will make the date column to be parsed as a date field.


```
from dateutil.parser import parse
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
```

```
plt.rcParams.update({'figure.figsize': (10, 7), 'figure.dpi': 120})

# Import as Dataframe
df =
pd.read_csv('https://raw.githubusercontent.com/selva86/datasets/master/a10.csv', parse_dates=['date'])
df.head()
```

	date	value
0	1991-07-01	3.526591
1	1991-08-01	3.180891
2	1991-09-01	3.252221
3	1991-10-01	3.611003
4	1991-11-01	3.565869

☞ **Dataframe Time Series**

Alternately, you can import it as a pandas Series with the date as index. You just need to specify the index_col argument in the pd.read_csv() to do this.

```
ser =
pd.read_csv('https://raw.githubusercontent.com/selva86/datasets/master/a10.csv', parse_dates=['date'],
index_col='date')
ser.head()
```

date	value
1991-07-01	3.526591
1991-08-01	3.180891
1991-09-01	3.252221
1991-10-01	3.611003
1991-11-01	3.565869

☞ **Series Timeseries**

Note, in the series, the 'value' column is placed higher than date to imply that it is a series.

☞ **sklearn. metrics,**

There are 3 different approaches to evaluate the quality of predictions of a model:

- Estimator score method : Estimators have a score method providing a default evaluation criterion for the problem they are designed to solve. This is not discussed on this page, but in each estimator's documentation.
- Scoring parameter : Model-evaluation tools using *cross-validation* (such as cross_validation.cross_val_score and grid_search. GridSearchCV) rely on an internal *scoring* strategy. This is discussed on section *The scoring parameter: defining model evaluation rules*.
- Metric functions : The metrics module implements functions assessing prediction errors for specific purposes. This is discussed in the section *Function for prediction-error metrics*.
- Finally, *Dummy estimators* are useful to get a baseline value of those metrics for random predictions.

See also

- For "pairwise" metrics, between *samples* and not estimators or predictions, see the *Pairwise metrics, Affinities and Kernels* section.

5.15.1 The Scoring Parameter: Defining Model Evaluation Rules

Model selection and evaluation using tools, such as grid_search.GridSearchCV and cross_validation.cross_val_score, take a scoring parameter that controls what metric they apply to estimators evaluated.

5.15.2 Common Cases : Predefined Values

For the most common use cases, you can simply provide a string as the scoring parameter. Possible values are:

Scoring	Function
Classification	
'accuracy'	sklearn.metrics.accuracy_score
'average_precision'	sklearn.metrics.average_precision_score
'f1'	sklearn.metrics.f1_score
'precision'	sklearn.metrics.precision_score

Scoring	Function
'recall'	sklearn.metrics.recall_score
'roc_auc'	sklearn.metrics.roc_auc_score
Clustering	
'adjusted_rand_score'	sklearn.metrics.adjusted_rand_score
Regression	
'mean_absolute_error'	sklearn.metrics.mean_absolute_error
'mean_squared_error'	sklearn.metrics.mean_squared_error
'r2'	sklearn.metrics.r2_score

Setting the scoring parameter to a wrong value should give you a list of acceptable values:

```
>>>
>>> from sklearn import svm, cross_validation, datasets
>>> iris = datasets.load_iris()
>>> X, y = iris.data, iris.target
>>> model = svm.SVC()
>>> cross_validation.cross_val_score(model, X, y, scoring='wrong_choice')
Traceback (most recent call last):
ValueError: 'wrong_choice' is not a valid scoring value. Valid options are ['accuracy', 'adjusted_rand_score', 'average_precision', 'f1', 'log_loss', 'mean_absolute_error', 'mean_squared_error', 'precision', 'r2', 'recall', 'roc_auc']
```

Note

- The corresponding scorer objects are stored in the dictionary sklearn.metrics.SCORERS.
- The above choices correspond to error-metric functions that can be applied to predicted values. These are detailed below, in the next sections.

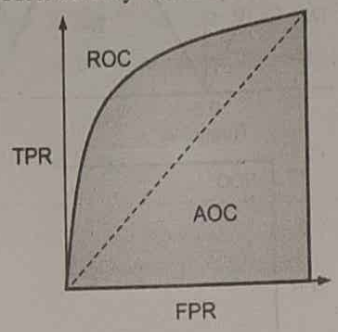
AU

- In Machine Learning, performance measurement is an essential task. So when it comes to a classification problem, we can count on an AUC - ROC Curve. When we need to check or visualize the performance of the multi-class classification problem, we use the AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve.
- It is one of the most important evaluation metrics for checking any classification model's performance. It is also written as AUROC (Area Under the Receiver Operating Characteristics)

What is the AUC - ROC Curve?

GQ. Write a short note on AUC & ROC Curve? **(2 Marks)**

- AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.
- Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.
- The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.



Defining terms used in AUC and ROC Curve.

TPR (True Positive Rate) / Recall /Sensitivity

$$\text{TPR/Recall/Sensitivity} = \frac{TP}{TP + FN}$$

Specificity

$$\text{Specificity} = \frac{TN}{TN + FP}$$

FPR

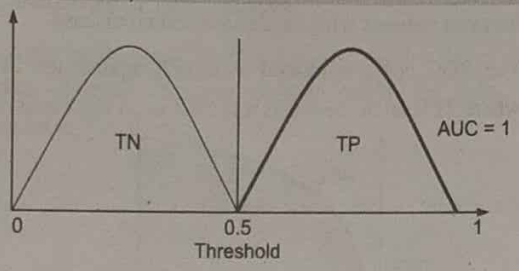
$$FPR = 1 - \text{Specificity} = \frac{FP}{TN + FP}$$

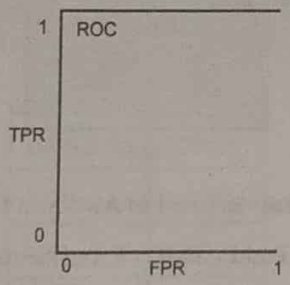
Unit V End Sem.

☞ How to speculate about the performance of the model ?

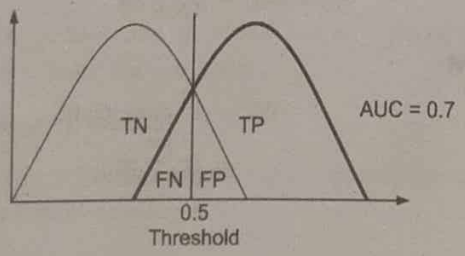
- An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability.
- In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means the model has no class separation capacity whatsoever.
- Let's interpret the above statements.
- As we know, ROC is a curve of probability. So let's plot the distributions of those probabilities:

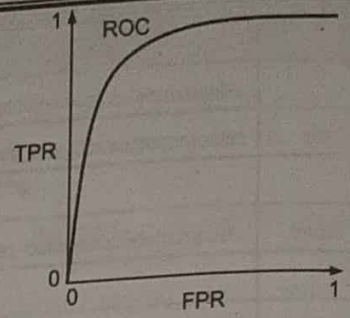
Note : Red distribution curve is of the positive class (patients with disease) and the green distribution curve is of the negative class (patients with no disease).

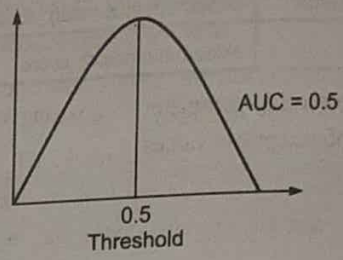




- This is an ideal situation. When two curves don't overlap at all means model has an ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.

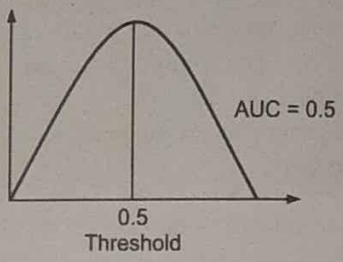


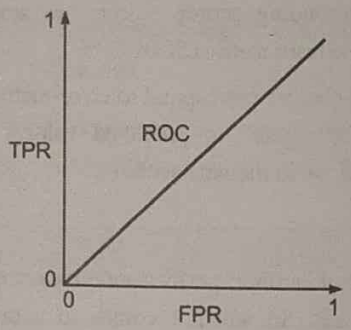




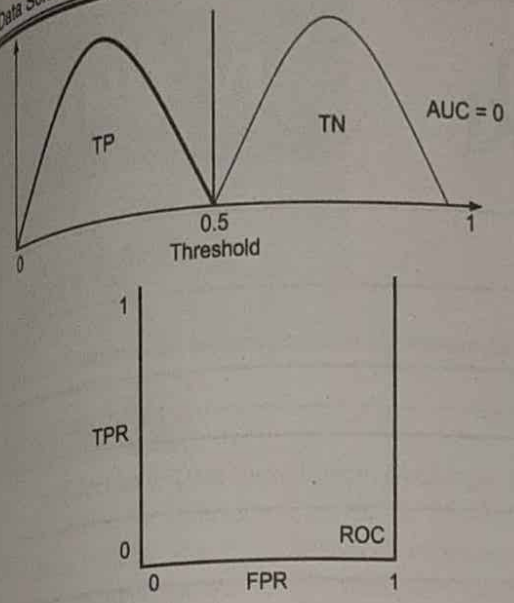
[When two distributions overlap, we introduce type 1 and type 2 errors

- Depending upon the threshold, we can minimize or maximize them. When AUC is 0.7, it means there is a 70% chance that the model will be able to distinguish between positive class and negative class.





- This is the worst situation. When AUC is approximately 0.5, the model has no discrimination capacity to distinguish between positive class and negative class.



- When AUC is approximately 0, the model is actually reciprocating the classes. It means the model is predicting a negative class as a positive class and vice versa.

☞ The relation between Sensitivity, Specificity, FPR, and Threshold.

- Sensitivity and Specificity are inversely proportional to each other. So when we increase Sensitivity, Specificity decreases, and vice versa.

  Sensitivity↑, Specificity↓ and Sensitivity↓, Specificity↑

- When we decrease the threshold, we get more positive values thus it increases the sensitivity and decreasing the specificity.
- Similarly, when we increase the threshold, we get more negative values thus we get higher specificity and lower sensitivity.
- As we know FPR is 1 - specificity. So when we increase TPR, FPR also increases and vice versa.

  TPR↑, FPR↑ and TPR↓, FPR↓

☞ How to use the AUC ROC curve for the multi-class model ?

In a multi-class model, we can plot the N number of AUC ROC Curves for N number classes using the One vs ALL methodology. So for example, If you have three classes named X, Y, and Z, you will have one ROC for X classified against Y and Z, another ROC for Y classified against X and Z, and the third one of Z classified against Y and X.

Chapter Ends...

❑❑❑