TABLE OF CONTENTS

Unit I

Unit II

Unit III

UNIT I

1 Introduction : Data Science and Big Data

Syllabus

Introduction to Data science and Big Data, *Defining Data science and Big Data, Big Data examples,* ***Data Explosion :*** *Data Volume, Data Variety, Data Velocity and Veracity. Big data infrastructure and challenges* ***Big Data Processing Architectures :*** *Data Warehouse, Re-Engineering the Data Warehouse, shared everything and shared nothing architecture, Big data learning approaches.* ***Data Science*** *- The Big Picture : Relation between AI, Statistical Learning, Machine Learning, Data Mining and Big Data Analytics.*

Contents

1.1 Introduction to Data Science

SPPU : April-20

- Data is a collection of facts and figures which relay something specific, but which are not organized in any way. It can be numbers, words, measurements, observations or even just descriptions of things. We can say, data is raw material in the production of information.
- Types of data are record data, data matrix, document data, transaction data, graph data and ordered data.
- Data science is an interdisciplinary field that seeks to extract knowledge or insights from various forms of data. At its core, data science aims to discover and extract actionable knowledge from data that can be used to make sound business decisions and predictions.
- Data science uses advanced analytical theory and various methods such as time series analysis for predicting the future. From historical data, instead of knowing how many products sold in the previous quarter, data science helps in forecasting future product sales and revenue more accurately.
- Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information and make business decisions. Data science uses complex machine learning algorithms to build predictive models.
- Data science enables businesses to process huge amounts of structured and unstructured big data to detect patterns.

1.1.1 Applications of Data Science

- Asking a personal assistant like Alexa or Siri for a recommendation demands data science. So does operating a self-driving car, using a search engine that provides useful results or talking to a chatbot for customer service. These are all real-life applications for data science.
- Following are some main reasons for using data science technology :
  - With the help of data science technology, we can convert the massive amount of raw and unstructured data into meaningful insights.
  - Data science technology is opting by various companies, whether it is a big brand or a startup. Google, Amazon, Netflix, which handle the huge amount of data, are using data science algorithms for better customer experience.
  - Data science is working for automating transportation such as creating a self-driving car, which is the future of transportation.
- Data science can help in different predictions such as various surveys, elections, flight ticket confirmation, etc.

1. Healthcare : Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.
2. Gaming : Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.
3. Image recognition : Identifying patterns in images and detecting objects in an image is one of the most popular data science applications.
4. Logistics : Data science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.
5. Predict future market trends : Collecting and analyzing data on a larger scale can enable you to identify emerging trends in your market. Tracking purchase data, celebrities and influencers and search engine queries can reveal what products people are interested in.
6. Recommendation systems : Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase or browse on their platforms.
7. Streamline manufacturing : Another way you can use data science in business is to identify inefficiencies in manufacturing processes. Manufacturing machines gather data from production processes at high volumes. In cases where the volume of data collected is too high for a human to manually analyze it, an algorithm can be written to clean, sort and interpret it quickly and accurately to gather insights.

1.1.2 Relationship between Data Science and Information Science

- Data science, as the interdisciplinary field, employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science and computer science. Data science and information science are twin disciplines by nature. The mission, task and nature of data science are consistent with those of information science.
- Data science is heavy on computer science and mathematics. Information science is used in areas such as knowledge management, data management and interaction design.
- 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.

1.1.3 Business Intelligence versus Data Science

Business Intelligent (BI)	Data Science
BI tends to provide reports, dashboards, and queries on business questions for the current period or in the past.	Data science tends to use disaggregated data in a more forward-looking, exploratory way, focusing on analyzing the present and enabling informed decisions about the future.
BI systems make it easy to answer questions related to quarter-to-date revenue, progress toward quarterly targets, and understand how much of a given product was sold in a prior quarter or year.	Data science tends to be more exploratory in nature and may use scenario optimization to deal with more open-ended questions.
BI helps monitor the current state of business data to understand the historical performance of a business.	Data science, as used in business, is basically data-driven, where many interdisciplinary sciences are applied together to extract meaning.
BI is designed to handle static and highly structured data.	Data science can handle high-speed, high-volume and complex, multi-structured data from a wide variety of data sources.

1.1.4 Data Science Life Cycle

- A data science life cycle is an iterative set of data science steps you take to deliver a project or analysis. Fig 1.1.1 shows data science life cycle.

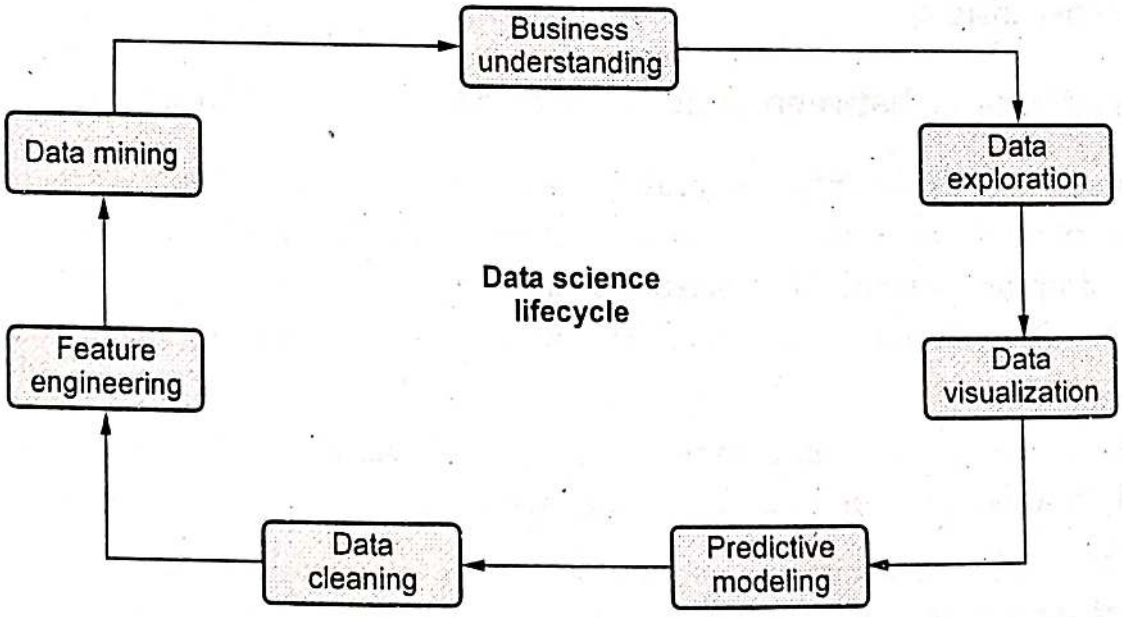


Fig. 1.1.1 Data science life cycle

a) Business understanding : Understand the basic problem you are trying to solve.

b) Data exploration : Understand the pattern and bias in your data.

c) Data visualization : Create and study of the visual representation of data.

d) Predictive modeling : It is the stage where the machine learning finally comes into your data.

e) Data cleaning : Detecting and correcting corrupt or inaccurate records.

f) Feature engineering : It is the process of cutting down the features.

g) Data mining : Gathering your data from different source.

Review Question

1. *Explain data science and its various applications.* **SPPU : April-20 (In Sem), Marks 4**

1.2 Defining Big Data

SPPU : April-18

- Big data can be defined as very large volumes of data available at various sources, in varying degrees of complexity, generated at different speed i.e., velocities and varying degrees of ambiguity, which cannot be processed using traditional technologies, processing methods, algorithms or any commercial off-the-shelf solutions.
- 'Big data' is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. In short, such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.
- The processing of big data begins with the raw data that isn't aggregated or organized and is most often impossible to store in the memory of a single computer.
- Big data processing is a set of techniques or programming models to access large-scale data to extract useful information for supporting and providing decisions. Hadoop is the open-source implementation of MapReduce and is widely used for big data processing.

1.2.1 Difference between Data Science and Big Data

Data science	Big data
It is a field of scientific analysis of data in order to solve analytically complex problems and the significant and necessary activity of cleansing, preparing of data.	Big data is storing and processing large volume of structured and unstructured data that can not be possible with traditional applications.
It is used in Biotech, energy, gaming and insurance.	Used in retail, education, healthcare and social media.
Goals : Data classification, anomaly detection, prediction, scoring and ranking.	Goals : To provide better customer service, identifying new revenue opportunities, effective marketing etc.

1.2.2 Benefits of Big Data Processing

Benefits of big data processing :

1. Improved customer service.
2. Business can utilize outside intelligence while taking decisions.
3. Reducing maintenance costs.
4. Re-develop your products : Big data can also help you understand how others perceive your products so that you can adapt them or your marketing, if need be.
5. Early identification of risk to the product / services, if any
6. Better operational efficiency.

1.2.3 Big Data Challenges

- Collecting, storing and processing big data comes with its own set of challenges :
  1. Big data is growing exponentially and existing data management solutions have to be constantly updated to cope with the three Vs.
  2. Organizations do not have enough skilled data professionals who can understand and work with big data and big data tools.

Review Question

1. *Justify your answer with example "Data science and big data are same or different".*

SPPU : April-18 (In Sem), Marks 5

1.3 Data Explosion

SPPU : April-18, 19, Dec.-18, 19

- The essence of computer applications is to store things in the real world into computer systems in the form of data, i.e., it is a process of producing data. Some data are the records related to culture and society and others are the descriptions of phenomena of the universe and life. The large scale of data is rapidly generated and stored in computer systems, which is called data explosion.
- Data is generated automatically by mobile devices and computers, think facebook, search queries, directions and GPS locations and image capture.
- Sensors also generate volumes of data, including medical data and commerce location-based sensors. Experts expect 55 billion IP - enabled sensors by 2021. Even storage of all this data is expensive. Analysis gets more important and more expensive every year.
- Fig. 1.3.1 shows the big data explosion by the current data boom and how critical it is for us to be able to extract meaning from all of this data.

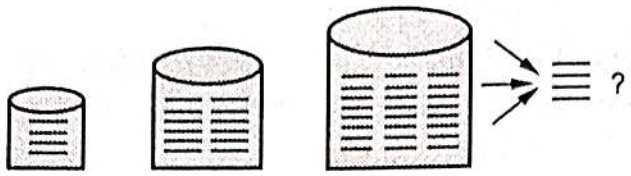


Fig. 1.3.1 Data explosion

- The phenomena of exponential multiplication of data that gets stored is termed as "Data Explosion". Continuous inflow of real-time data from various processes, machinery and manual inputs keeps flooding the storage servers every second.
- Sending emails, making phone calls, collecting information for campaigns; each day we create a massive amount of data just by going about our normal business and this data explosion does not seem to be slowing down. In fact, 90 % of the data that currently exists was created in just the last two years.
- Reason for this data explosion is **Innovation.**
  1. Business model transformation : Innovation changed the way in which we do business, provide services. The data world is governed by three fundamental trends are business model transformation, globalization and personalization of services.
     - Organizations have traditionally treated data as a legal or compliance requirement, supporting limited management reporting requirements. Consequently, organizations have treated data as a cost to be minimized.
     - The businesses are required to produce more data related to product and provide services to cater each sector and channel of customer.
  2. Globalization : Globalization is an emerging trend in business where organizations start operating on an international scale. From manufacturing to customer service, globalization has changed the commerce of the world. Variety and different formats of data are generated due to globalization.
  3. Personalization of services : To enhance customer service, the form of one-to-one marketing in the form of personalization of service is opted by the customer. Customers expect communication through various channels increases the speed of data generation.
  4. New sources of data : The shift to online advertising supported by the likes of Google, Yahoo and others is a key driver in the data boom. Social media, mobile devices, sensor networks and new media are on the fingertips of customers or users. The data generated through this is used by corporations for decision support systems like business intelligence and analytics. The growth of technology helped to emerge new business models over the last decade or more. Integration of all the data across the enterprise is used to create business decision support platform.

1.3.1 V's of Big Data

- We differentiate big data characteristics from traditional data by one or more of the five V's : *Volume, velocity, variety, veracity and value.*

1. **Volume :** Volumes of data are larger than that conventional relational database infrastructure can cope with. It consisting of terabytes or petabytes of data.
   - Fig. 1.3.2 shows big data volume.

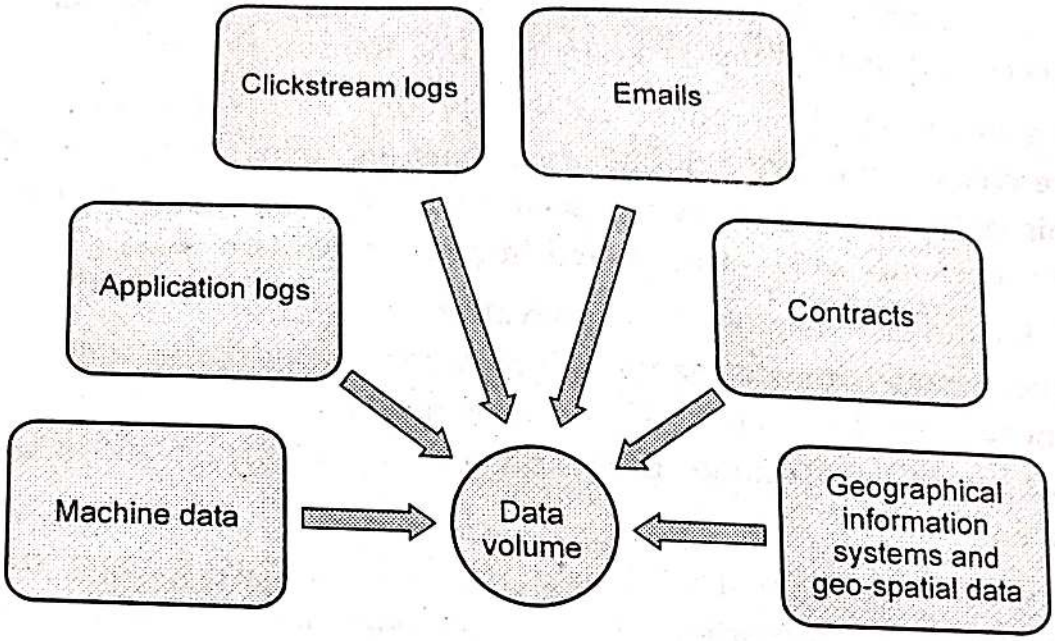


Fig. 1.3.2 Big data volume

2. **Velocity :** The term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. It is being created in or near real-time.
3. **Variety :** It refers to heterogeneous sources and the nature of data, both structured and unstructured.
   - Fig. 1.3.3 (a) and Fig. 1.3.3 (b) shows big data velocity and data variety.

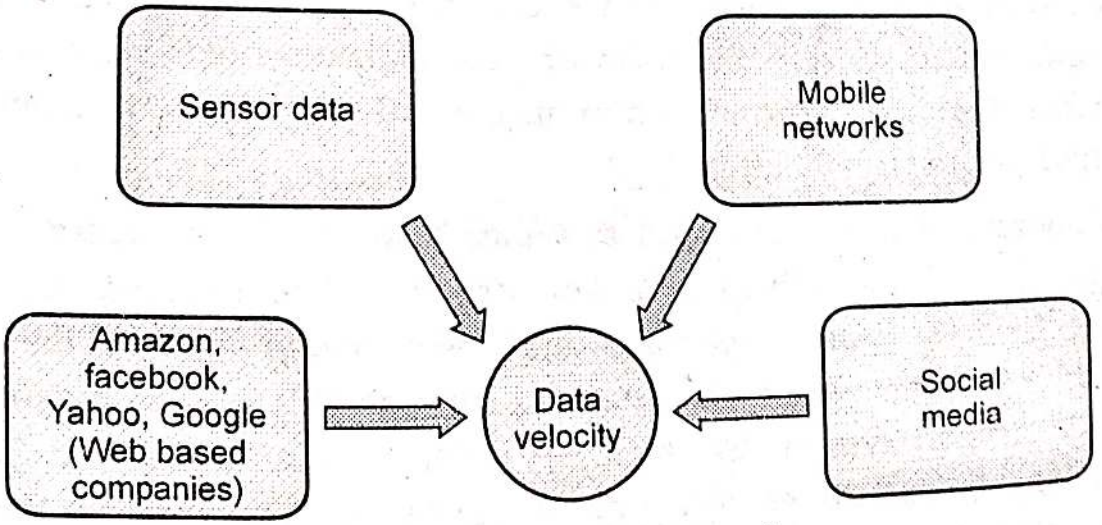


Fig. 1.3.3 (a) Data velocity

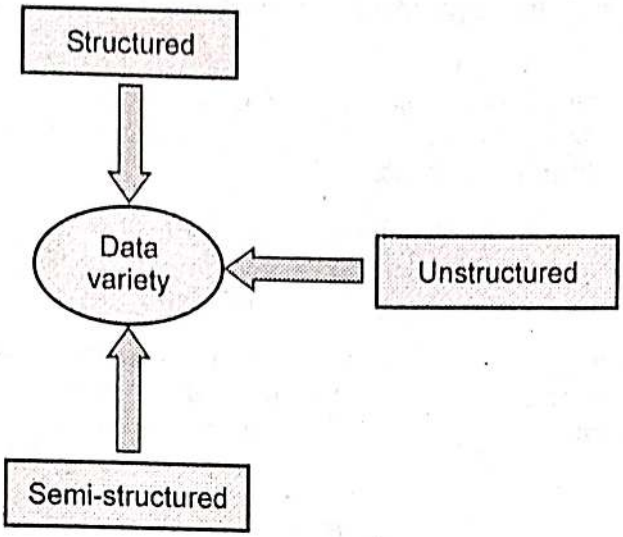


Fig. 1.3.3 (b) Data variety

4. **Value :** It represents the business value to be derived from big data.
   - The ultimate objective of any big data project should be to generate some sort of value for the company doing all the analysis. Otherwise, you're just performing some technological task for technology's sake.
   - For real-time spatial big data, decisions can be enhanced through visualization of dynamic change in such spatial phenomena as climate, traffic, social-media-based attitudes and massive inventory locations.
   - Exploration of data trends can include spatial proximities and relationships. Once spatial big data are structured, formal spatial analytics can be applied, such as spatial autocorrelation, overlays, buffering, spatial cluster techniques and location quotients.
5. **Veracity :** Big data must be fed with relevant and true data. We will not be able to perform useful analytics if much of the incoming data comes from false sources or has errors. Veracity refers to the level of trustiness or messiness of data and if higher the trustiness of the data, then lower the messiness and vice versa. It relates to the assurance of the data's quality, integrity, credibility and accuracy. We must evaluate the data for accuracy before using it for business insights because it is obtained from multiple sources.

1.3.2 Compare Cloud Computing and Big Data

Cloud computing	Big data
It provides resources on demand.	It provides a way to handle huge volumes of data and generate insights.
It refers to internet services from SaaS, PaaS to IaaS.	It refers to data, which can be structured, semi-structured or unstructured.

Cloud is used to store data and information on remote servers.	It is used to describe a huge volume of data and information.
Cloud computing is economical as it has low maintenance costs centralized platform no upfront cost and disaster safe implementation.	Big data is a highly scalable, robust ecosystem and cost-effective.
Vendors and solution providers of cloud computing are Google, Amazon web service, Dell, Microsoft, Apple and IBM.	Vendors and solution providers of big data are Cloudera, Hortonworks, Apache and MapR.
The main focus of cloud computing is to provide computer resources and services with the help of network connection.	Main focus of big data is about solving problems when a huge amount of data generating and processing.

Review Questions

1. *State one example of big data and explain how all V's are applied for big data example.*

SPPU : Dec.-18 (End Sem), Marks 4

2. *Explain 5 V's for defining big data along with the factors responsible for data explosion.*

SPPU : April-19 (In Sem), Marks 5

3. *Explain big data along with 5 V's.*

SPPU : April-18 (In Sem), Marks 5, Dec.-19 (End Sem), Marks 6

1.4 Big Data Examples

- Machine data consists of information generated from industrial equipment, real-time data from sensors that track parts and monitor machinery and even web logs that track user behavior online.
- At arcplan client CERN, the largest particle physics research center in the world, the Large Hadron Collider (LHC) generates 40 terabytes of data every second during experiments.
- Regarding transactional data, large retailers and even B2B companies can generate multitudes of data on a regular basis considering that their transactions consist of one or many items, product IDs, prices, payment information, manufacturer and distributor data and much more.

Factors responsible for data volume in big data are as follows :

1. Machine data : Machine data contains a definitive record of all activity and behavior of your customers, users, transactions, applications, servers, networks, factory machinery and so on. It's configuration data, data from APIs and message queues, change events, the output of diagnostic commands and call detail records, sensor data from remote equipment and more.

2. Application log : Most homegrown and packaged applications write local logfiles, logging services built into application servers like WebLogic, WebSphere and JBoss. These files are critical for day-to-day debugging of production applications by developers and application support. When developers put timing information into their log events, they can also be used to monitor and report on application performance.
3. Business process logs : Complex events processing and business process management system logs are treasure troves of business and IT relevant data. These logs will generally include definitive records of customer activity across multiple channels such as the web, IVR / contact center or retail.
4. Clickstream data : User activity on the Internet is captured in clickstream data. This provides insight into a user's website and web page activity. This information is valuable for usability analysis, marketing and general research.
5. Third party data : The sensitive data that's not in databases is on file systems. In some industries such as healthcare, the biggest data leakage risk is consumer records on shared file systems. Different OS, third-party tools and storage technologies provide different options for auditing read access to sensitive data at the file system level. This audit data is a vital data source for monitoring and investigating access to sensitive data.
6. Electronic mails : Every company have large collection of emails generated by customers, employees and executives on daily basis. These email communication are an important asset to an organization, which are audited case-by-case basis and entire life cycle management of emails is done.

- Some of the examples of big data are :
  1. **Social media :** Social media is one of the biggest contributors to the flood of data we have today. Facebook generates around 500 + terabytes of data everyday in the form of content generated by the users like status messages, photos and video uploads, messages, comments etc.
  2. **Stock exchange :** Data generated by stock exchanges is also in terabytes per day. Most of this data is the trade data of users and companies.
  3. **Aviation industry :** A single jet engine can generate around 10 terabytes of data during a 30 minute flight.
  4. **Survey data :** Online or offline surveys conducted on various topics which typically has hundreds and thousands of responses and need to be processed for analysis and visualization by creating a cluster of population and their associated responses.
  5. **Compliance data :** Many organizations like healthcare, hospitals, life sciences, finance etc, has to file compliance reports.

1.5 Data Processing Infrastructure Challenges

SPPU : May-18

- Data processing infrastructure challenges are storage, transportation, processing and throughput.

1. Storage : The increase in the volume of data, increases the need for storing the data and processing of data. Big data technology has changed the way we gather and store data, including data storage device, data storage architecture and data access techniques. It requires more sophisticated storage medium with higher I/O speed to meet the challenges of big data issues. Direct-Attached Storage (DAS), Network-Attached Storage (NAS) and Storage Area Network (SAN) are the enterprise storage architecture that are commonly are in use.
2. Transportation : Data is transfer from one place to other place and process the data then load into memory for manipulation. The data is transported between computer and storage layers. Increase in bandwidth is not solution to this problem.
3. Processing : Data processing needs to combine the logic and mathematical computation in one cycle. This processing can be accomplished by CPU or processor, memory and software. With each generation CPU processing speed is increased, have improved processing capabilities. Memory is required for compute and processing. Memory has become cheaper and faster with evolution of processor capability. Software are used to write the programs for transforming and processing of data.
4. Speed and throughout : This is the major challenges for data processing. Various architecture layers like hardware, software's networking and storage are responsible for storage and are added. Each layer has its own limitation, causes limitation in the overall throughput in the data processing.

Review Question

1. *List and explain data processing infrastructure challenges in big data.*

SPPU : May-18 (End Sem), Marks 6

1.6 Big Data Processing Architectures

SPPU : April-18,19,20, May-18

1.6.1 Data Warehouse

- A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process. A data warehouse stores historical data for purposes of decision support.

- A database is an application-oriented collection of data that is organized, structured, coherent, with minimum and controlled redundancy, which may be accessed by several users in due time.
- Data warehousing provides architectures and tools for business executives to systematically organize, understand and use their data to make strategic decisions.
- A data warehouse is a subject-oriented collection of data that is integrated, time-variant, non-volatile, which may be used to support the decision-making process.
- Data warehouses are databases that store and maintain analytical data separately from transaction-oriented databases for the purpose of decision support.
- The main objective of a data warehouse is to support the decision-making system, focusing on the subjects of the organization. The objective of a database is to support the operational system and information is organized on applications and processes.

Multitier architecture of data warehouse :

- Data warehouse architecture is a data storage framework's design of an organization. A data warehouse architecture takes information from raw sets of data and stores it in a structured and easily digestible format.
- Data warehouse system is constructed in three ways. These approaches are classified the number of tiers in the architecture.
  a) Single-tier architecture.
  b) Two-tier architecture.
  c) Three-tier architecture (Multi-tier architecture).
- *Single tier* warehouse architecture focuses on creating a compact data set and minimizing the amount of data stored. While it is useful for removing redundancies. It is not effective for organizations with large data needs and multiple streams.
- *Two-tier* warehouse structures separate the resources physically available from the warehouse itself. This is most commonly used in small organizations where a server is used as a data mart. While it is more effective at storing and sorting data. Two-tier is not scalable and it supports a minimal number of end-users.

Three tier (Multi-tier) architecture :

- Three tier architecture creates a more structured flow for data from raw sets to actionable insights. It is the most widely used architecture for data warehouse systems.

- Fig. 1.6.1 shows three tier architecture. Three tier architecture sometimes called multi-tier architecture.

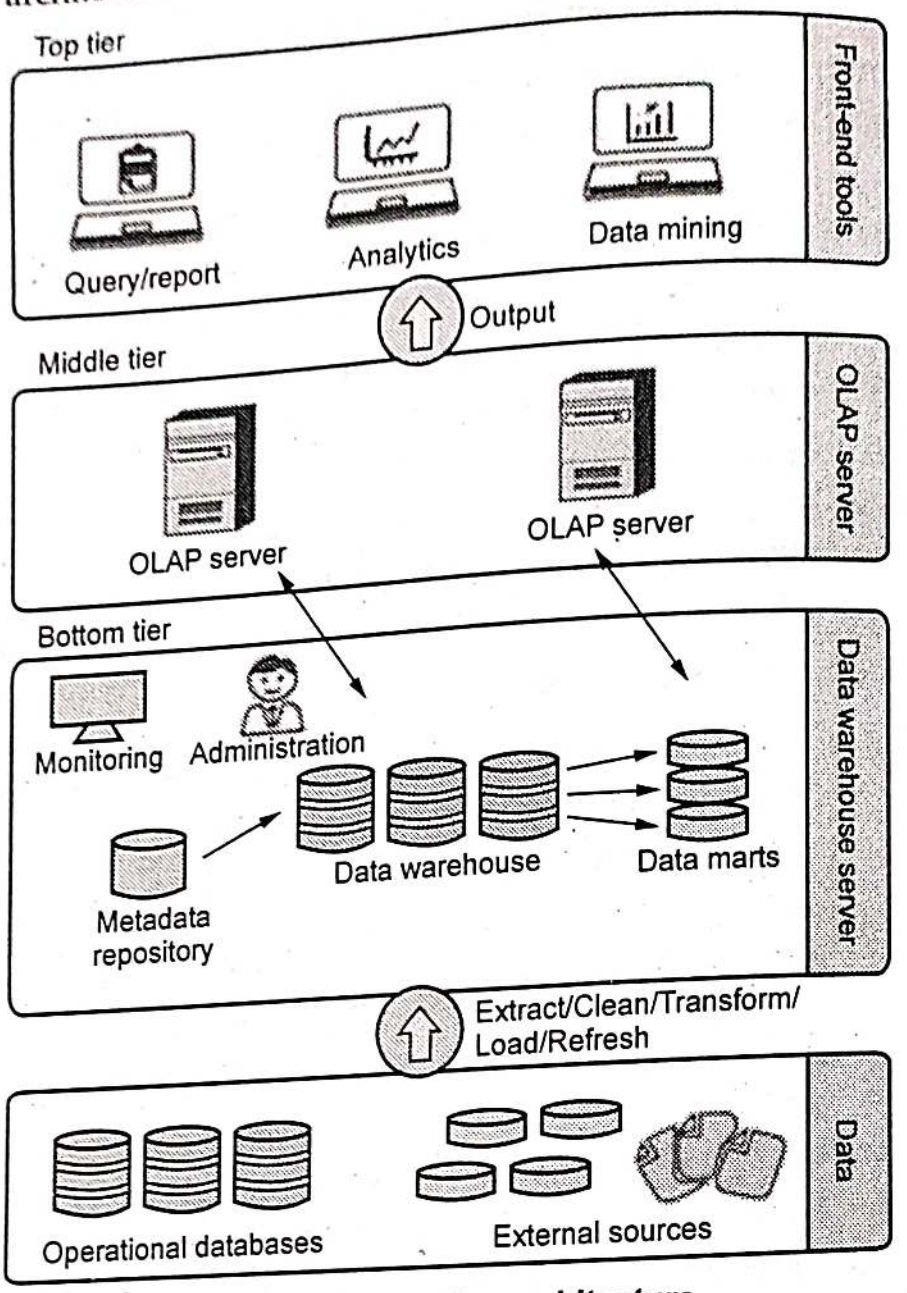


Fig. 1.6.1 Three tier architecture

- The bottom tier is the database of the warehouse, where the cleansed and transformed data is loaded. The bottom tier is a warehouse database server.
- The middle tier is the application layer giving an abstracted view of the database. It arranges the data to make it more suitable for analysis. This is done with an OLAP server, implemented using the ROLAP or MOLAP model.
- OLAPS can interact with both relational databases and multidimensional databases, which lets them collect data better based on broader parameters.
- The top tier is the front-end of an organization's overall business intelligence suite. The top-tier is where the user accesses and interacts with data via queries, data visualizations and data analytics tools.

- The top tier represents the front-end client layer. The client level which includes the tools and Application Programming Interface (API) used for high-level data analysis, inquiring and reporting. User can use reporting tools, query, analysis or data mining tools.

1.6.2 Shared Everything Architecture

- This architectural model consists of nodes that share all resources within the system. Each node has access to the same computing resources and shared storage. Shared-everything architecture refers to system architecture where all resources are shared including storage, memory and the processor.
- Fig. 1.6.2 shows shared everything architecture.

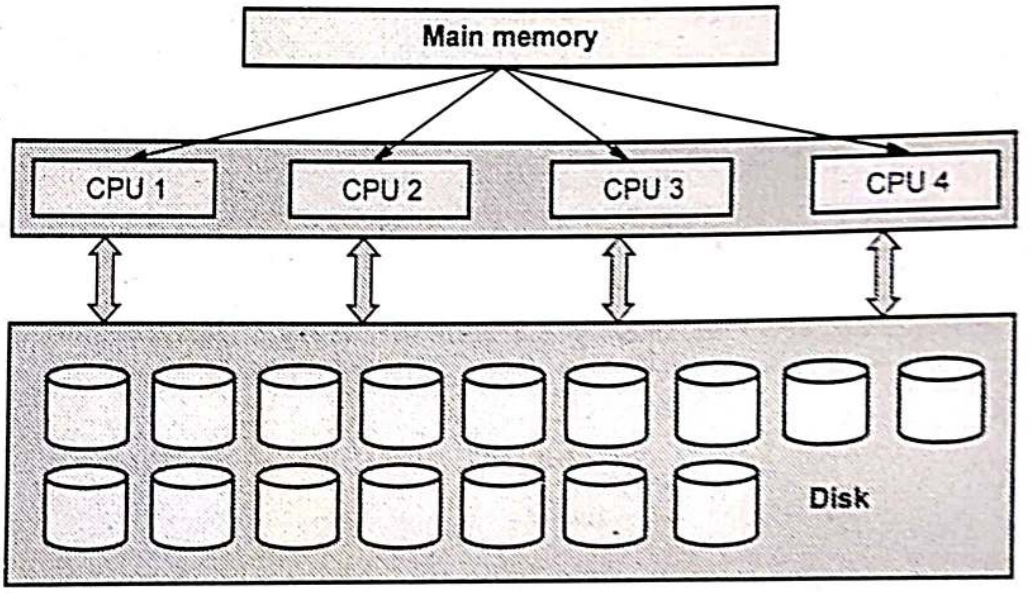


Fig. 1.6.2 Shared everything architecture

- The main idea behind such a system is maximizing resource utilization. The disadvantage is that shared resources also lead to reduced performance due to contention.
- Scalability is the main problem. Oracle RAC uses this architecture.
- Symmetric multiprocessing (SMP) and Distributed Shared Memory (DSM) are the types of shared everything architecture.
- In the SMP architecture, all the CPUs share a single pool of memory for read-write access concurrently and uniformly without latency. Sometimes this is referred to as Uniform Memory Access (UMA) architecture.

- The DSM architecture addresses the scalability problem by providing multiple pools of memory for processors to use. In the DSM architecture, the latency to access memory depends on the relative distances of the processors and their dedicated memory pools. This architecture is also referred to as Nonuniform Memory Access (NUMA) architecture.

1.6.3 Shared Nothing Architecture

- Shared nothing architecture is a distributed computing architecture that consists of multiple separated nodes that don't share resources. The nodes are independent and self-sufficient as they have their own disk space and memory.
- Each node has its own private memory (M), processor (CPUs) and storage devices independent of any other node in the configuration. This means that every node stores its own lock table and buffer pool.
- Fig. 1.6.3 shows shared nothing architecture.

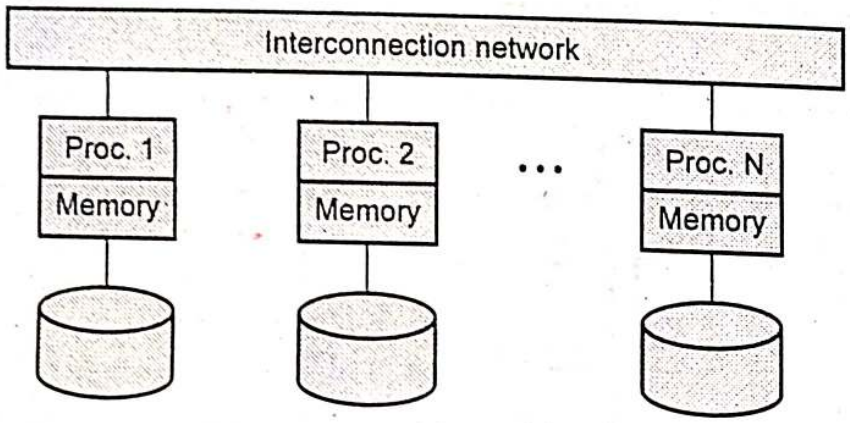


Fig. 1.6.3 Shared nothing architecture

- The key feature of shared-nothing architecture is that the operating system, not the application server, owns responsibility for controlling and sharing hardware resources. Each node is under the control of its own copy of the operating system and thus can be viewed as a local site.
- Shared nothing is also known as Massively Parallel Processing (MPP) solutions typically employed by large data warehouse systems.
- Data is horizontally partitioned across nodes, such that each node has a subset of the rows from each table that was distributed and all the replicated tables.
- Shared nothing can be made to scale to hundreds or even thousands of machines. Because of this, it is generally regarded as the best-scaling architecture. Shared-nothing architecture scales better and is well suited for a cloud data warehouse considering very low-cost commodity PCs and networking hardware.

1.6.4 Re - engineering the Data Warehouse

- Re-engineering the data warehouse means building a next generation data warehouse. Fig 1.6.4 shows re-engineering the data warehouse.

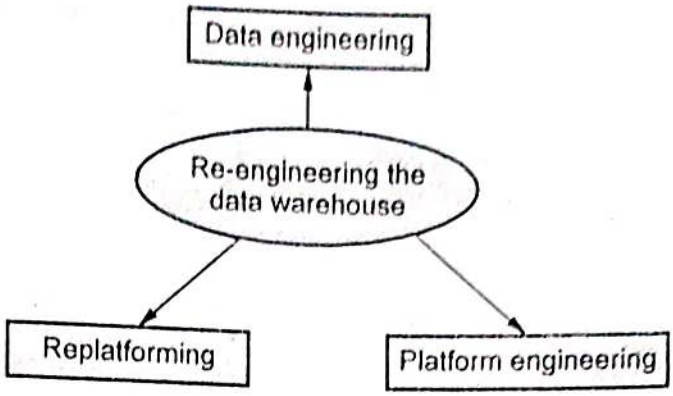


Fig. 1.6.4 Re-engineering the data warehouse

- Various methods of re-engineerings are replatforming, platform engineering and data engineering.
  1. Replatforming : Replatforming means, new infrastructure and hardware. Depending on organizations requirement, new technologies such as warehouse appliances, tiered storage, private cloud can be deployed.
- Advantages : Scalability, reliability, security, lower maintenance, code optimization.
- Disadvantages : It is time consuming and leads to disruption of business activities.
  2. Data engineering : It is re-engineering of data structures for creating better performance. There is a change in initial data model of data warehouse to make new data model. It includes partitioning the tables in vertical or horizontal partition, colocation of related tables in same storage region, distribution of data, adding new data types and adding new database functions for performance boost.
  3. Platform engineering : It is related to modifying some parts of the infrastructure, which helps to gain better scalability and improved performance.

     Platform engineering is popular concept in automotive industry as product parts are crafted to offer improved quality, service and cost.

Review Questions

1. *Explain the role of shared everything and shared nothing architecture in big data.*

   SPPU : April-18 (In Sem), Marks 5

2. *What is data warehouse ? Explain design and architecture of data warehouse.*

   SPPU : May-18 (End Sem), Marks 6

3. *Explain shared-everything and shared-nothing architectures in detail with respect to big data.*

   SPPU : April-19 (In Sem), Marks 5

4. List and explain choices for re-engineering the data warehouse.

SPPU : April-19 (In Sem), Marks 3

5. Discuss the processing complexities associated with the big data.

SPPU : April-19 (In Sem), Marks 3

6. What are the pitfalls of data warehouse ? Why companies are shifting to big data using hadoop.

SPPU : April-20 (In Sem), Marks 4

7. Draw and explain big data processing architecture with technologies used at each of the stage of big data processing.

SPPU : April-20 (In Sem), Marks 6

1.7 Big Data Learning Approaches

SPPU : April-18, 2

- Data is a boon for machine learning systems. The more data a system receives, the more it learns to function better for businesses. Hence, using machine learning for big data analytics happens to be a logical step for companies to maximize the potential of big data adoption.
- Big data refers to extremely large sets of structured and unstructured data that cannot be handled with traditional methods. Big data analytics can make sense of the data by uncovering trends and patterns. Machine learning can accelerate this process with the help of decision-making algorithms. It can categorize the incoming data, recognize patterns and translate the data into insights helpful to business operations.
- Machine learning algorithms are useful for collecting, analyzing and integrating data for large organizations. They can be implemented in all elements of big data operation, including data labeling and segmentation, data analytics and scenario simulation.
- Machine Learning (ML) is considered as a very fundamental and vital component of data analytics. In fact ML is predicted to be the main drivers of the big data revolution for obvious reasons for its ability to learn from data and provide with data driven insights, decisions and predictions.
- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.
- Machine learning provides business insight and intelligence. Decision makers are provided with greater insights into their organizations. This adaptive technology being used by global enterprises to gain a competitive edge.
- Supervised and unsupervised learning are the different types of machine learning methods.

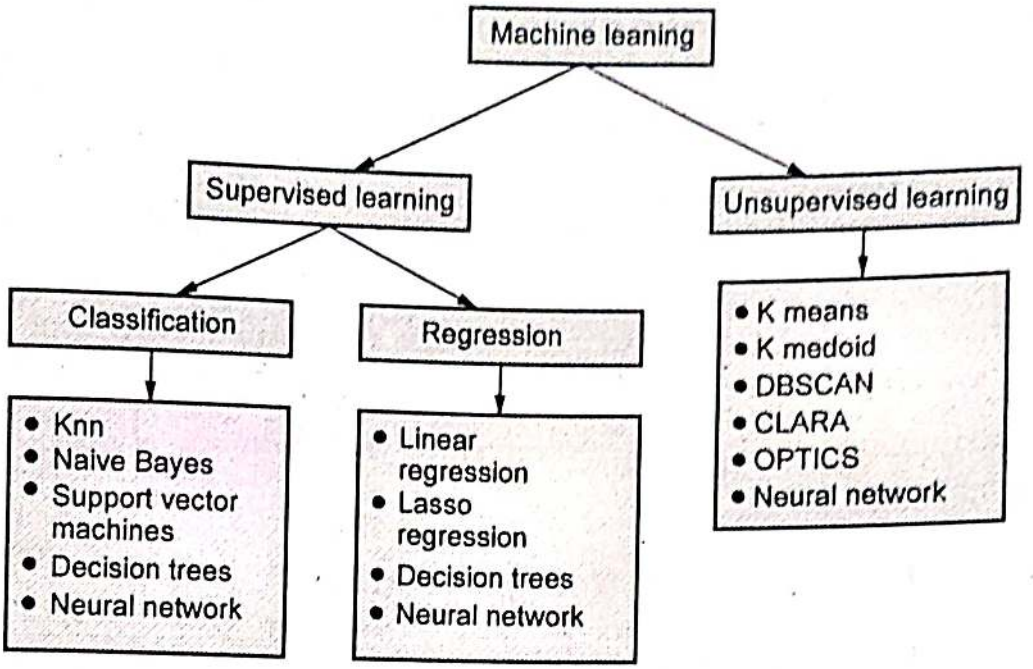


Fig. 1.7.1

- **Supervised learning** is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. The task of the supervised learner is to predict the output behavior of a system for any set of input values, after an initial training phase. Supervised learning is also called classification.
- **Unsupervised learning** algorithms aim to learn rapidly and can be used in real-time. Unsupervised learning is frequently employed for data clustering, feature extraction etc. Unsupervised learning is also called clustering.

Review Questions

1. *Enlist the impact of learning approaches in big data ? Explain different kinds of learning approaches.* **SPPU : April-18 (In Sem), Marks 5**
2. *Explain different learning approaches in big data. Explain with example.* **SPPU : April-20 (In Sem), Marks 6**

1.8 Data Science : The Big Picture

SPPU : Dec.-19

- The field of data science is fundamentally about organizing and using data to provide insights for human decision-making. Developing the capability to glean insights from data has become crucial for start-ups and fortune 500 companies alike.
- Many organizations have been collecting unprecedented amounts of data from both physical sensors and the online activity of millions of people but a pile of unorganized data will not yield insights on its own.

- This is where data scientists come into play by cleaning up and organizing data to make it suitable for analysis. They also build the statistical models necessary for analyzing data to reveal notable patterns or trends.
- Several key types of data analysis :
  a) Descriptive analytics aims to gain insight into either current or historical data trends.
  b) Predictive analytics looks to gain insight into future unknowns by using the best available data to make predictions.
  c) Prescriptive analytics can recommend what humans should do given the available data insights.

1.8.1 Relation between AI and Machine Learning

- The emergence of modern AI based on machine learning has significantly boosted predictive and prescriptive analytics.
- AI refers to a set of tools that can automate machine actions in ways that mimic intelligent behaviors. A small sample of such applications might include :
  a) Visually identifying cats and dogs in social media photos.
  b) Translating between different languages in the text found on websites.
  c) Detecting possible signs of cancer in patients' X-ray images.

Machine learning in modern AI systems :

- Most modern AI is based on machine learning : A category of computer algorithms that can automatically learn from data. Instead of relying on humans to program each step, ML models train on large datasets to identify notable patterns within the data and make their own predictions based on that information.
- They can then apply the lessons learned from their training datasets to analyzing completely new and unfamiliar datasets in the real world.
- The importance of the training data means that ML performance depends greatly upon having access to large and diverse datasets of high quality.
- For example, a machine learning model that trains to recognize dogs by only looking at 100 images of Siberian Huskies is unlikely to perform well when suddenly tasked with identifying tens of thousands of images from a diverse array of dog breeds.
- ML models can follow several different approaches :
  a) Supervised learning relies heavily upon hand-labeled training datasets and is the most common type of machine learning.

b) Unsupervised learning sifts through unlabelled data to try and find unusual patterns that might escape the human eye.

c) Reinforcement learning uses trial and error to learn from mistakes and get closer to achieving a specific goal.

- Here is just one example of how data science can intersect with AI based on machine learning. Let us assume that an Internet search engine company wants to provide and monetize the most relevant online searches in response to the query : "Allergy medicine for kids."
- Data scientists help collect and organize large datasets containing millions of user search results related to allergy medication for kids. Then, they work with software developers and engineers to build machine learning models that learn from these datasets.
- Through training, machine learning models can identify user preferences for various search results, such as information about what allergy medications come in the form of syrups and chewable tablets. This helps to continuously update the search engine, so that it delivers more relevant results and ranks them higher.
- The search and click trends identified by the machine learning models also provide information about people's medical needs and shopping habits, such as certain allergy medicine brands being more popular among families in a specific geographic area at a certain time of year.
- Data scientists analyze these trends to find business insights that they can share with corporate leaders and online advertisers.

1.8.2 Data Mining and Big Data Analytics

- Data mining refers to extracting or mining knowledge from large amounts of data. It is a process of discovering interesting patterns or knowledge from a large amount of data stored either in databases, data warehouses or other information repositories.
- It is the computational process of discovering patterns in huge data sets involving methods at the intersection of AI, machine learning, statistics and database systems.
- To make predictions, predictive mining tasks perform inference on the current data. Predictive analysis provides answers of the future queries that move across using historical data as the chief principle for decisions.
- It involves the supervised learning functions used for the prediction of the target value. The methods fall under this mining category are the classification, time-series analysis and regression.

- Descriptive analytics is the conventional form of business intelligence and data analysis, seeks to provide a depiction or "summary view" of facts and figures in an understandable format, to either inform or prepare data for further analysis.
- Descriptive analytics helps organizations to understand what happened in the past. It helps to understand the relationship between product and customers.
- Big data analytics is the often complex process of examining big data to uncover information such as hidden patterns, correlations, market trends and customer preferences that can help organizations make informed business decisions.

Review Question

1. *Explain machine learning approaches in big data.* **SPPU : Dec.-19 (End Sem), Marks 6**

1.9 Multiple Choice Questions

Q.1 Three characteristics of big data are _____ _____ _____.

a volume, velocity, variety

b value, variable, variance

c volume, vanish, various

d velocity, volume, vault

Q.2 Data is collection of data objects and their _____.

a information

b attributes

c characteristics

d none

Q.3 In big data, _____ refer to heterogeneous sources and the nature of data, both structured and unstructured.

a volume

b variety

c velocity

d all of these

Q.4 Various types of data analytics are _____.

a descriptive model

b predictive model

c prescriptive model

d all of these

Q.5 Machine learning is inherently a _____ field.

a interdisciplinary

b multidisciplinary

c single

d none

UNIT II

2 Mathematical Foundation of Big Data

Syllabus

Probability : *Random Variables and Joint Probability, Conditional Probability and concept of Markov chains, Tail bounds, Markov chains and random walks, Pair-wise independence and universal hashing Approximate counting, Approximate median.* ***Data Streaming Models and Statistical Methods :*** *Flajole Martin algorithm, Distance Sampling and Random Projections, Bloom filters, Mode, Variance, standard deviation, Correlation analysis and Analysis of Variance.*

Contents

2.1 Probability

SPPU : May-18

- Probability theory is concerned with the study of random phenomena. Such phenomena are characterized by the fact that their future behaviour is not predictable in a deterministic fashion. The role or probability theory is to analyze the behavior of a system or algorithm assuming the given probability assignments and distributions.
- Probability was developed to analyze the games of chance. It is mathematical modeling of the phenomenon of chance or randomness. The measure of chance is called the probability of the statement.
- The probability of an even is defined as the number of favorable outcomes divided by the total number of possible outcomes.

2.1.1 Classical Definition of Probability

1. Computing Probability using the Classical Method

- If an experiment has n equally likely simple events and if the number of ways that an event E can occur is m, then the probability of E, P(E), is

$$P(E) = \frac{\text{Number of way that E can occur}}{\text{Number of possible outcomes}} = \frac{m}{n}$$

So, if S is the sample space of this experiment, then

$$P(E) = \frac{N(E)}{N(S)}$$

- If $\overline{E}$ denotes the events of non-occurrence of E, then the number of elementary events in $\overline{E}$ is n – m and hence the probability of $\overline{E}$ is :

$$P(\overline{E}) = \frac{n-m}{n} = 1-\frac{m}{n}$$

$$= 1 - P(E) \Rightarrow P(E) + P(\overline{E}) = 1$$

(Here m is a non-negative integer and n is a positive integer and $m \leq n$).

It is also called mathematical or priori probability.

2.1.2 Random Experiment

- In some experiments, we are not able to control the value of certain variables so that the results will vary from one performance of the experiment to the next, even though most of the conditions are the same. These experiments are called as **random experiment.**

- Random experiment is defined as an experiment whose outcomes are known before the experiment is performed but which outcome is going to happen in a particular trial is not known.
- Example : If we toss a die, the result of the experiment is that it will come up with one of the numbers in the set {1, 2, 3, 4, 5, 6}.
- A **random variable** is simply an expression whose value is the outcome of a particular experiment.

2.1.3 Sample Space

- The totality of the possible outcomes of a random experiment is called the sample space of the experiment and it will be denoted by letter 'S'.
- There will be more than one sample space that can describe outcomes of an experiment, but there is usually only one that will provide the most information.
- The sample space is not determined completely by the experiment. It is partially determined by the purpose for which the experiment is carried out.
- Example 1 : If the experiment consists of flipping two coins, then the sample space consists of the following points :

$$S = \{(T, T), (T, H), (H, T), (H, H)\}$$

The outcome will be (T, T) if both coins are tails, (T, H) if the first coin is tails and the second heads, (H, T) if the first is heads and the second tails, and (H, H) if both coins are heads.

- Example 2 : A die is rolled once. We let X denote the outcome of this experiment. Then the sample space for this experiment is the 6-element set,

$$S = \{1, 2, 3, 4, 5, 6\},$$

- It is convenient to classify sample spaces according to the number of elements they contains. If a sample space has a finite number of points, it is called a ***finite sample space.*** If it has as many points as there are natural numbers 1, 2, 3, ... , it is called a ***countably infinite sample space.*** If it has as many points as there are in some interval on the x axis, such as $0 \leq x \leq 1$, it is called a ***non-countably infinite sample space.***
- A sample space that is finite or countably finite is often called a *discrete sample space*, while one that is non-countably infinite is called a *non-discrete sample space.*
- The result of a trail in a random experiment is called an **out come.**

2.1.4 Event

- An event is simply a collection of certain sample points.

- An *event* is a subset A of the sample space S, i.e., it is a set of possible outcome If the outcome of an experiment is an element of A, we say that the event *A h occurred*. An event consisting of a single point of S is called a *simple* or *elementa event*.
- A *simple event is any single outcome from a probability experiment. Each simple event* denoted e_i.
- A *single* performance of the experiment is known as a **trial**.
- As particular events, we have S itself, which is the *sure* or *certain event* since element of S must occur, and the empty set ϕ, which is called the *impossible eve* because an element of ϕ cannot occur.
- An **unusual event** is an event that has a low probability of occurring.
- **Independent events :** Let E_1, E_2 be two events. Then E_1, E_2 are said to b independent events if,

  $$P(E_1 \cap E_2) = P(E_1) P(E_2)$$

- **Mutually exclusive events :** $E_1, E_2, \ldots E_n$ are said to be mutually exclusive $E_i \cap E_j = \phi$, for $i \neq j$. If E_1 and E_2 are independent and mutually exclusive event then either $P(E_1) = 0$ or $P(E_2) = 0$. If E_1 and E_2 are independent events, then E and E_2^c are also independent.
- Mutually exclusive events are sometimes called **as Disjoint event.** If A and B ar mutually exclusive, then it is not possible for both events to occur on the sam trial. If two events are Mutually Exclusive Events then they do not share commo outcomes.
- **Example :**
  1. In throwing a die all six possible cases are mutually exclusive.
  2. In tossing a coin the event head turning up and tail turning up are mutuall exclusive.
- Two events of a sample space whose intersection is ϕ and whose union is th entire sample space are called **complementary events.** If E is an event of a sampl space S, its complement is denoted by E' or $\overline{E}$.
- **Equally likely events :** Two or more events are said to be equally likely events, i the chances of their happening in a trial is equal. For example, when a card i drawn from a pack, any card may be obtained. In the trial, all the 52 elementary event are equally likely.
- If the happening of an event in the first trial influences the happening of an even in the consecutive trial, then the events are said to be **dependent events.**

- **Exhaustive events :** The total number of all possible events of a random experiment is called exhaustive events. For example, in tossing a coin, there are two exhaustive elementary events, i.e. head and tail.

2.1.5 Algebra of Events

- Following Fig. 2.1.1 shows the relation between two sets.

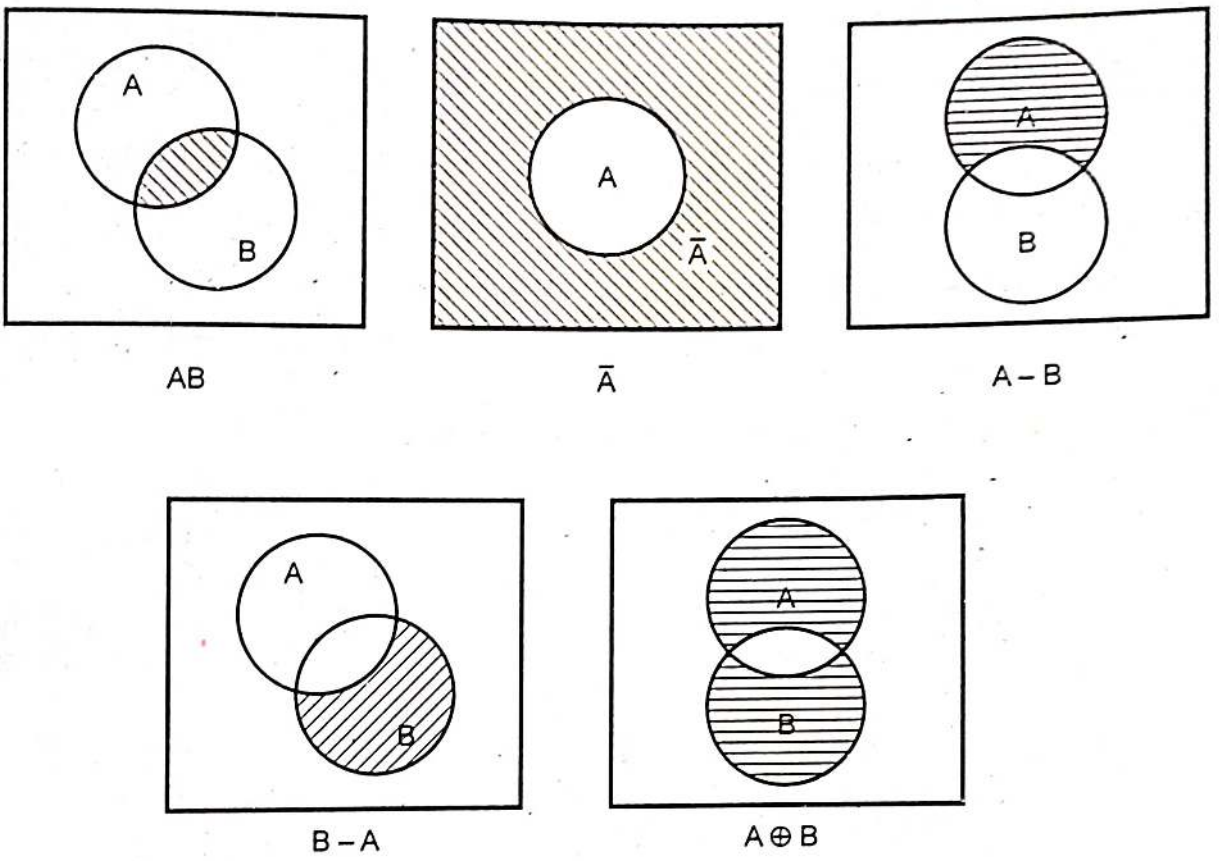


Fig. 2.1.1 Relation between two sets

1. Commutative law :

$$A \cap B = B \cap A$$

$$A \cup B = B \cup A$$

2. Distributive law :

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

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

3. Associative law :

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

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

De Morgan's law :

- The useful relationship between the three basic operations of forming unions, intersection and complements are known as De Morgrans laws.
- The complement of a union (intersection) of two sets A and B equals the intersection (union) of the complements $\overline{A}$ and $\overline{B}$. Thus

$$\overline{(A \cup B)} = \overline{A} \cap \overline{B}$$

$$\overline{(A \cap B)} = \overline{A} \cup \overline{B}$$

Example 2.1.1 *Prove that :* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Solution : Let A and B be any two events. To write $A \cup B$ as the union of three mutually exclusive events : $A \cap B^c$, $A \cap B$ and $A^c \cap B$.

$$A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B) \quad \ldots (1)$$

By axioms 3 :

$$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) \quad \ldots (2)$$

Now,

$$A = (A \cap B^c) \cup (A \cap B)$$

$$B = (A^c \cap B) \cup (A \cap B)$$

Therefore

$$P(A) = P(A \cap B^c) + P(A \cap B) \quad \ldots (3)$$

and

$$P(B) = P(A^c \cap B) + P(A \cap B) \quad \ldots (4)$$

By equations (3) and (4), we get

$$P(A) + P(B) = P(A \cap B^c) + P(A^c \cap B) + 2P(A \cap B) \quad \ldots (5)$$

Compare the equation (2) and (5),

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

So,

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

Example 2.1.2 *If* $P(A) = \frac{1}{5}$, $P(B) = \frac{2}{3}$, $P(A \cap B) = \frac{1}{15}$

Find a) $P(A \cup B)$ *b)* $P(A^c \cap B)$ *c)* $P(A \cap B^c)$.

Solution : a) $P(A \cup B)$

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

$$= \frac{1}{5} + \frac{2}{3} - \frac{1}{15} = \frac{3 + 2 \times 5}{15} - \frac{1}{15} = \frac{13 - 1}{15} = \frac{12}{15}$$

$$P(A \cup B) = \frac{4}{5}$$

b) $P(A^c \cap B)$

$$P(A^c \cap B) = P(B) - P(A \cap B)$$

$$= \frac{2}{3} - \frac{1}{15} = \frac{2 \times 5 - 1}{15} = \frac{10 - 1}{15}$$

$$P(A^c \cap B) = \frac{9}{15} = \frac{3}{5}$$

c) $P(A \cap B^c)$

$$P(A \cap B^c) = P(A) - P(A \cap B)$$

$$= \frac{1}{5} - \frac{1}{15} = \frac{3 - 1}{15}$$

$$P(A \cap B^c) = \frac{2}{15}$$

Example 2.1.3 *If A and B are two events such that* $P(A) = \frac{1}{3}$, $P(B) = \frac{3}{4}$ *and* $P(A \cup B) = \frac{11}{12}$, *find* $P(A|B)$ *and* $P(B|A)$.

Solution :

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

$$= \frac{1}{3} + \frac{3}{4} - \frac{11}{12} = \frac{4 + 3 \times 3}{12} - \frac{11}{12}$$

$$= \frac{13}{12} - \frac{11}{12} = \frac{13 - 11}{12} = \frac{2}{12}$$

$$P(A \cap B) = \frac{1}{6}$$

So, $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{\frac{1}{6}}{\frac{3}{4}} = \frac{1}{6} \times \frac{4}{3}$$

$$P(A|B) = \frac{4}{18} = \frac{2}{9}$$

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

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

Example 2.1.4 *Determine i) $P(B|A)$ ii) $P\left(\frac{A}{B^c}\right)$ if A and B are event with $P(A) = \frac{1}{3}$,*

$P(B) = \frac{1}{4}$, $P(A \cup B) = \frac{1}{2}$.

Solution :

i) $P\left(\frac{B}{A}\right)$

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

First calculate $P(A \cap B)$ from given data.

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

$$= \frac{1}{3} + \frac{1}{4} - \frac{1}{2}$$

$$= \frac{4+3}{12} - \frac{1}{2}$$

$$= \frac{7}{12} - \frac{1}{2} = \frac{7-6}{12}$$

$$P(A \cap B) = \frac{1}{12}$$

Substitute this value into the formula.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{12}}{\frac{1}{3}} = \frac{1}{12} \times \frac{3}{1} = \frac{3}{12}$$

$$P\left(\frac{B}{A}\right) = \frac{1}{4}$$

ii) $P\left(\frac{A}{B^c}\right)$

First calculate $P(B^c)$

$$P(B^c) = 1 - P(B) = 1 - \frac{1}{4} = \frac{4-1}{4}$$

$$P(B^c) = \frac{3}{4}$$

$$P\left(\frac{A}{B^c}\right) = \frac{P(A \cap B^c)}{P(B^c)}$$

$$P(A \cap B^c) = P(A) - P(A \cap B)$$

$$= \frac{1}{3} - \frac{1}{12} = \frac{4-1}{12} = \frac{3}{12}$$

$$P(A \cap B^c) = \frac{1}{4}$$

Therefore,

$$P\left(\frac{A}{B^c}\right) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{4} \times \frac{4}{3}$$

$$P\left(\frac{A}{B^c}\right) = \frac{1}{3}$$

Example 2.1.5 *Given that a person's last purchase was pepsi, there is a 90 % chance that his next purchase will also be pepsi. If a person's last purchase was coke, there is an 80 % chance that his next purchases will also be coke. What is the probability that he will purchase pepsi three purchases from now ?* **SPPU : May-18 (End Sem), Marks 6**

Solution :

$$D = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

$$D^2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

$$D^2 = \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix}$$

$$D^3 = \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

$$D^3 = \begin{bmatrix} 0.781 & 0.219 \\ 0.438 & 0.562 \end{bmatrix}$$

2.2 Random Variables

SPPU : Dec.-19

- A random variable is a set of possible values from a random experiment.
- A random variable, usually written X, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.
- Whenever you run and experiment, flip a coin, roll a die, pick a card, you assign a number to represent the value to the outcome that you get. This assign is called **a random variable.**
- A random variable is a variable X that assigns a real number [x], for each and every outcome of a random experiment. If **S** is the sample space containing all the 'n' outcomes $\{e_1, e_2, e_3, ..., e_i, ..., e_n\}$ of random experiment and X is a random variable defined as a function X(e) on S, then for every outcome e_i (where i = 1, 2, 3, ..., n) that is in S the random variable $X(e_i)$ will assign a real value x_i.
- Advantages of random variables is that user can define certain probability functions that make it both convenient and easy to compute the probabilities of various events.
- A random variable is a numerically valued variable which takes on different values with given probabilities.

Examples :

1. The return on an investment in a one-year period.
2. The price of an equity.
3. The number of customers entering a store.
4. The sales volume of a store on a particular day.
5. The turnover rate at your organization next year.

2.2.1 Discrete Random Variables

- If we can find a way to list all possible outcomes for a random variable and assign probabilities to each one, we have a discrete random variable.
- The random variable is called a **discrete random variable** if it is defined over a sample space having a finite or a countable infinite number of sample points. In

this case, random variable takes on discrete values and it is possible to enumerate all the values it may assume.

- A discrete random variable can only have a specific (or finite) number of numerical values.
- We can have infinite discrete random variables if we think about things that we know have an estimated number. Think about the number of stars in the universe.
- We know that there are not a specific number that we have a way to count so this is an example of an infinite discrete random variable.
- The mean of any discrete random variable is an average of the possible outcomes, with each outcome weighted by its probability.
- Example :
  1. Total of roll of two dice : 2, 3, ... , 12
  2. Number of desktops sold : 0, 1, ...
  3. Customer count : 0, 1,

2.2.2 Continuous Random Variable

- In the case sample space having an uncountable infinite number of sample points, the associated random variable is called a **continuous random variable**, with its values distributed over one or more continuous intervals on the real line.
- A continuous random variable is one which takes an infinite number of possible values. Continuous random variables are usually measurements. Examples include height, weight, the amount of sugar in an orange, the time required to run a mile.
- A continuous random variable is one having continuous range of values. It cannot be produced from a discrete sample space because of our requirement that all random variables be single valued functions of all sample space points.
- A continuous random variable is not defined at specific values. Instead, it is defined over an interval of values and is represented by the area under a curve. The probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.
- Both types of random variables are important in science and engineering. Maxed random variable is one for which some of its values are discrete and some are continuous.

2.2.3 Probability Mass Function and Cumulative Distribution Function of a Discrete Random Variable

- For a discrete random variable, the probability that a random variable X taking a specific value x_i, $P(X = x_i)$, is called the probability mass function $P(x_i)$.

Probability mass function is a function that maps each outcome of a random experiment to a probability.

- Let X be a discrete random variable with range $R_X = \{x_1, x_2, x_3, ...\}$ (finite or countably infinite). The function $P_X(x_k) = P(X = x_k)$, for $k = 1, 2, 3, ...$, is called the Probability Mass Function (PMF) of X.
- Thus, the PMF is a probability measure that gives us probabilities of the possible values for a random variable.
- The bahavior of a random variable is characterized by its probability distribution, that is, by the way probabilities are distributed over the values it assumes. A probability mass function are two ways to characterize this distribution for a discrete random variable.
- They are equivalent in the sense that the knowledge of either one completely specifies the random variable. The corresponding functions for a continuous random variable are the probability distribution function, defined in the same way as it the case of discrete random variable and the probability density function.
- If X is random variable, then the function $F(x)$ is defined by,

$$F(x) = P\{X \leq x\}$$

is called the **Probability Distributed Function (PDF)** of X. All probabilities concerning X can be stated in terms of F. The argument 'x' is any real number ranging from ∞ to ∞.

- The probability distribution function is also called **Cumultative Distributon Function (CDF).**
- Distribution functions of discrete random variables grows only by jumps, whereas the distribution functions of continuous random variables are continuous functions and hence have no jumps.
- If $F_X(x)$ is a continuous function of x, then

$$F_X(x) = F_X(x^-)$$

- However, if $F_X(x)$ is discontinuous at the point x then,

$$F_X(x) - F_X(x^-) = p[x^- < X \leq x]$$

$$= \lim_{e \to 0} p[x - \epsilon < X \leq x]$$

$$= \triangleq p[X = x]$$

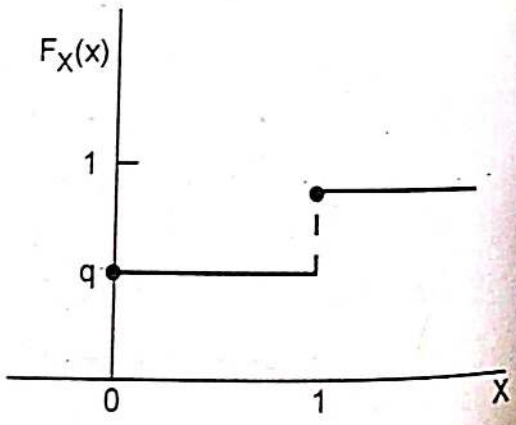


Fig. 2.2.1

Typically $p[X = x]$ is a discontinuous function of x; it is zero whenever $F_X(x)$ continuous and nonzero only at discontinuities in $F_X(x)$.

Example 2.2.1 *Probability of a function of the number of heads from tossing a coin four times. Determine the cumulative distribution function.*

Solution :

$$F(0) = f(0) = \frac{1}{16}$$

$$F(1) = f(0) + f(1)$$

$$= \frac{1}{16} + \frac{4}{16} = \frac{5}{16}$$

$$F(2) = f(0) + f(1) + f(2)$$

$$= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} = \frac{11}{16}$$

$$F(3) = f(0) + f(1) + f(2) + f(3)$$

$$= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16}$$

$$= \frac{1+4+6+4}{16} = \frac{15}{16}$$

$$F(4) = f(0) + f(1) + f(2) + f(3) + f(4)$$

$$= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16}$$

$$= \frac{1+4+6+4+1}{16} = \frac{16}{16} = 1$$

Example 2.2.2 *What is the probability distribution for the toss of one fair coin ?*

Solution :

$$P \text{ (Heads)} = \frac{1}{2}$$

$$P \text{ (Tails)} = \frac{1}{2}$$

Let heads denote the coin landing head side up.

Let tails denote the coin landing tail side up.

The possible outcomes are for the coin to land head side up or tail side up.

Using the alternative notation.

$$P(X = \text{Heads}) = \frac{1}{2}$$

$$P(X = \text{Tails}) = \frac{1}{2}$$

X	P (X)
Heads	$\frac{1}{2}$
Tails	$\frac{1}{2}$

2.2.4 Difference between Discrete and Continuous Random Variable

Sr. No.	Discrete	Continuous
1.	It uses countable set.	It uses set of interval on R.
2.	F is set of all subset of Ω.	F is made from sub-intervals of Ω with set operations.
3.	For a set $A \in F$, $P(A) = \sum_{\omega \in A} p(\omega)$	For a set $A \in F$, $p(A) = \int_{A}^{-A} f_X(x)\,dx$
4.	Distribution function (Cdf) : $F_X(x) = \sum_{\omega \le x} p\omega$	Distrinbution function (Cdf) : $F_X(x) = \int_{-\infty}^{x} f_X(x)\,dt$

2.2.5 Mean and Variance of Distribution

- The mean μ and variance σ^2 of a random variable X and of its distribution are the theoretical counterparts of the mean $\bar{x}$ and variance s^2 of a frequency distribution.
- The mean μ (mu) is defined by :

$$\mu = \sum_{j} x_j f(x_j) \qquad \text{for discrete distribution}$$

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx \qquad \text{for continuous distribution}$$

and the variance σ^2 (Sigma square) by :

$$\sigma^2 = \sum_{j} (x_j - \mu)^2 f(x_j) \qquad \text{for discrete distribution}$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx \qquad \text{for continuous distribution}$$

- The mean (μ) is also denoted by E (X) and is called the expectation of X because it gives the average value of X to be expected in many trials.
- Let us compute the variance of a normal distribution. If X has an $N(\mu, \sigma^2)$ distribution, then :

$$\text{Var}(X) = E[(X - E[X])^2]$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \sigma^2 \int_{-\infty}^{\infty} Z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2} dx$$

Here we substituted $Z = (x-\mu)/\sigma$.

Using integration,

$$\int_{-\infty}^{\infty} Z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2} dz = 1$$

Example 2.2.3 *Find the variance and standard deviation for the following set of test marks T = {75, 80, 82, 87, 96}*

Solution :

$$\text{Mean} = \frac{75+80+82+87+96}{5} = \frac{420}{5}$$

Mean = 84

$$\text{Variance} = \frac{[(x_1-\mu)^2 + (x_2-\mu)^2 + \ldots + (x_n-\mu)^2]}{n}$$

$$\sigma^2 = \frac{[(75-84)^2 + (80-84)^2 + (82-84)^2 + (87-84)^2 + (96-84)^2}{5}$$

$$= \frac{(-9)^2 + (-4)^2 + (-2)^2 + (3)^2 + (12)^2}{5}$$

$$= \frac{81+16+4+9+144}{5} = \frac{254}{5}$$

$$\sigma^2 = 50.8$$

Standard Deviation (σ)

$$\sigma = \sqrt{\sigma^2} = \sqrt{50.8} = 7.1274$$

Example 2.2.4 *If X is a normal variate with mean 30 and standard deviation 5. Find the probability that,*

a) $26 \leq x \leq 40$ *b)* $x \geq 45$.

Solution : Given data :

Mean $\mu = 30$

Standard deviation $\sigma = 5$.

i) $x_1 = 26$ and $x_2 = 40$

$$Z = \frac{x-\mu}{\sigma}$$

$$Z_1 = \frac{x_1-\mu}{\sigma} = \frac{26-30}{5} = \frac{4}{5} = -0.8$$

$$Z_2 = \frac{x_2-\mu}{\sigma} = \frac{40-30}{5} = \frac{10}{5} = 2$$

$$P(26 \leq x \leq 40) = P(-0.8 \leq z \leq 2)$$

ii) $x \geq 45$

$$Z = \frac{x-\mu}{\sigma} = \frac{45-30}{5} = \frac{15}{5} = 3$$

Example 2.2.5 *A random variable X has the following probability function :*

x	0	1	2	3	4	5	6	7
$P(x)$	0	K	$2K$	$2K$	$3K$	K^2	$2K^2$	$7K^2+K$

Determine :

i) K ii) Evaluate $P(X < 6)$, $P(X \geq 6)$, $P(0 < X < 5)$ *and* $P(0 \leq X \leq 4)$.

iii) If $P(X \leq K) > \frac{1}{2}$ *find the minimum value of K.*

iv) Determine the distribution function of X.

v) Mean vi) Variance.

Solution : i) K

$$\sum_{x=0}^{7} P(x) = 1$$

$$K + 2K + 2K + 3K + K^2 + 2K^2 + 7K^2 + K = 1$$

$$10K^2 + 9K - 1 = 0$$

$$(K+1)(10K-1) = 0$$

$$K + 1 = 0 \quad \text{and} \quad 10K - 1 = 0$$

$$K = -1 \quad \text{and} \quad K = \frac{1}{10}$$

We discard $K = -1$ value. Therefore $K = \frac{1}{10} = 0.1$.

ii) $P(X < 6) = P(X=0) + P(X=1) + P(X=2) + \ldots + P(X=5)$

$$= 0 + K + 2K + 2K + 3K + K^2$$

Put $K = 0.1$

$$= 0 + 0.1 + 2(0.1) + 2(0.1) + 3(0.1) + (0.1)^2$$

$$= 0.1 + 0.2 + 0.2 + 0.3 + 0.01$$

$$P(X < 6) = 0.81$$

$$P(X \geq 6) = 1 - P(X < 6)$$

$$= 1 - 0.81$$

$$P(X \geq 6) = 0.19$$

$$P(0 < X < 5) = P(X=1) + P(X=2) + P(X=3) + P(X=4)$$

$$= K + 2K + 2K + 3K$$

$$= 8K$$

$$= 8 \times 0.1 \qquad (K = 0.1)$$

$$P(0 \leq X < 5) = 0.8$$

$$P(0 \leq X \leq 4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)$$

$$= 0 + K + 2K + 2K + 3K$$

$$= 8K$$

$$= 8 \times 0.1$$

$$P(0 \leq X \leq 4) = 0.8$$

iii) If $P(X \leq K) > \frac{1}{2}$, minimum value of K.

$$P(X \leq 1) = P(X = 0) + P(X = 1)$$
$$= 0 + K = K = 0.1$$
$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$
$$= 0 + K + 2K$$
$$= 3K = 3 \times 0.1 = 0.3$$
$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$
$$= 0 + K + 2K + 2K$$
$$= 5K = 5 \times 0.1 = 0.5$$
$$P(X \leq 4) = 0.8 \text{ (We already calculated)}$$

But the condition is $P(X \leq K) > \frac{1}{2}$.

So K = 4 is suitable for this minimum value of K = 4.

iv) Distribution function of X.

X	F (X) = P (X ≤ x)
0	0
1	0.1
2	0.3
3	0.5
4	0.8
5	0.81
6	0.83
7	$9K + 10K^2 = 1$

v) Mean (μ)

$$\mu = \sum_{i=0}^{7} p_i x_i$$

$$= 0(0) + 1(K) + 2(2K) + 3(2K) + 4(3K) + 5(K^2) + 6(2K^2) + 7(7K^2 + K$$

$$= 0 + K + 4K + 6K + 12K + 5K^2 + 12K^2 + 49K^2 + 7K$$

$$= 30K + 66K^2$$

Substitute $K = 1/10$

$$= 30 \times \frac{1}{10} + 66 \times \left(\frac{1}{10}\right)^2$$

$$= \frac{30}{10} + \frac{66}{100} = 3 + 0.66$$

$$\mu = 3.66$$

vi) Variance (σ^2)

$$\sigma^2 = \sum_{i=0}^{7} p_i \, x_i^2 - \mu^2$$

$$= K + 8K + 18K + 48K + 25K^2 + 72K^2 + 343K^2 + 49K - (3.66)^2$$

$$= 440K^2 + 124K - 13.3956 = 440\,(0.1)^2 + 124\,(0.1) - 13.3956$$

$$= 4.4 + 12.4 - 13.3956$$

$$\sigma^2 = 3.4044$$

Example 2.2.6 *A petrol station owner is considering the effect on his business (Superpet) of a new petrol station (Global) which has opened just down the road. Currently (of the total market shared between Superpet and Global) Superpet has 80 % of the market and Global has 20 %. Analysis over the last week has indicated the following probabilities for customers switching the station they stop at each week :*

	To	
	Superpet	*Global*
From superpet	*0.75*	*0.25*
Global	*0.55*	*0.45*

What will be the expected market share for Superpet and Global after another two weeks have past ? **SPPU : Dec.-19 (End Sem), Marks 6**

Solution : We have :

state 1 = Superpet

state 2 = Global

So the initial system state s_1 is given by $s_1 = [0.80, 0.20]$ and the transition matrix P given by,

$$P = \begin{vmatrix} 0.75 & 0.25 \\ 0.55 & 0.45 \end{vmatrix}$$

- Hence after one week has elapsed the state of the system $s_2 = s_1P = [0.71, 0.29]$ and so after two weeks have elapsed the state of the system $= s_3 = s_2P$ $= [0.692, 0.308]$ and note here that the elements of s_2 and s_3 add to one (as required).
- Hence the market shares after two weeks have elapsed are 69.2 % and 30.8 % for Superpet and Global respectively.
- Assuming that in the long-run the system reaches an equilibrium $[x_1, x_2]$ where $[x_1, x_2] = [x_1, x_2]P$ and $x_1 + x_2 = 1$
- We have that

$$x_1 = 0.75\,x_1 + 0.55\,x_2 \quad (1)$$

$$x_2 = 0.25\,x_1 + 0.45\,x_2 \quad (2)$$

$$x_1 + x_2 = 1 \quad (3)$$

From equation (3) we have that $x_2 = 1 - x_1$, so substituting into equation (1) we get,

$$x_1 = 0.75\,x_1 + 0.55(1 - x_1)$$

i.e. $(1 - 0.75 + 0.55)\,x_1 = 0.55$

i.e. $$x_1 = \frac{0.55}{0.80} = \mathbf{0.6875}$$

Hence $$x_2 = 1 - x_1 = 1 - 0.6875 = \mathbf{0.3125}$$

Hence the long-run market shares are 68.75 % and 31.25 % for Superpet and Globa respectively.

2.3 Joint Probability

SPPU : Dec.-18

- A joint probability is a probability that measures the likelihood that two or mor events will happen concurrently.
- If there are two independent events A and B, the probability that A and B wi occur is found by multiplying the two probabilities. Thus for two events A and the special rule of multiplication shown symbolically is :

$$P(A \text{ and } B) = P(A)\,P(B).$$

- The general rule of multiplication is used to find the joint probability that two events will occur. Symbolically, the general rule of multiplication is,

$$P(A \text{ and } B) = P(A)\, P(B|A).$$

- The probability $P(A \cap B)$ is called the joint probability for two events A and B which intersect in the sample space. Venn diagram will readily shows that,

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

Equivalently :

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

- The probability of the union of two events never exceeds the sum of the event probabilities.
- A tree diagram is very useful for portraying conditional and joint probabilities. A tree diagram portrays outcomes that are mutually exclusive.

2.3.1 Conditional Probability

- Let A and B be two events such that $P(A) > 0$. We denote $P(B|A)$ the probability of B given that A has occurred. Since A is known to have occurred, it becomes the new sample space replacing the original S. From this, the definition is,

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

OR

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

- The notation $P(B \mid A)$ is read **"the probability of event B given event A"**. It is the probability of an event B given the occurrence of the event A.
- We say that , the probability that both A and B occur is equal to the probability that A occurs times the probability that B occurs given that A has occurred. We call **$P(B|A)$ the conditional probability of B given A**, i.e., the probability that B will occur given that A has occurred.
- Similarly, the conditional probability of an event A, given B by,

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

- The probability $P(A|B)$ simply reflects the fact that the probability of an event A may depend on a second event B. If A and B are mutually exclusive $A \cap B = \phi$ and $P(A|B) = 0$.

- Another way to look at the conditional probability formula is :

$$P(\text{Second}/\text{First}) = \frac{P(\text{First choice and second choice})}{P(\text{First choice})}$$

- Conditional probability is a defined quantity and cannot be proven.
- The key to solving conditional probability problems is to :
  1. Define the events.
  2. Express the given information and question in probability notation.
  3. Apply the formula.

Example 2.3.1 *Only 1 in 1000 people has rare disease. Given true positive = 0.9 and false positive = 0.02. If randomly tested individual is positive, what is the probability that they have a disease.* **SPPU : Dec.-18 (End Sem), Marks 6**

Solution : Infected and test indicates disease (true positive) = 1000 × 0.9 / 100 = 9

Uninfected and test indicates disease (false positive)

$$= 1000 \times (100 - 0.9) \ / 100 \times 0.02 = 19.82$$

$$\rightarrow 20 \text{ people would receive a false positive.}$$

The remaining 971 (= 1000 – (9 + 20)) tests are correctly negative.

In population, 9 out of 29 total people with a positive test result are actually infected. So, the probability of actually being infected after one is told that one is infected is only 34 % = (20 / (9 + 20)).

2.4 Concept of Markov Chains

SPPU : Apr.-18,19, May-18, Dec.-19

- Important classes of stochastic processes are Markov chains and Markov processes.
- Markov chain is a discrete time stochastic process with the Markov property.
- Markov process is the continuous time version of a Markov chain. Many queueing models are Markov processes.
- A Markov process is a random process in which the probability distribution of the current state is conditionally independent of the path of path states, a characteristic called the Markov property.
- Markov chain is completely characterised by transition probabilities $P_{i,j}$ from state 'i' to state 'j' are stored in an n×n transition matrix P.
- Markov chain is a mathematical model of a random phenomenon evolving with time in a way that the past affects the future only through the present.

- A Markov process is called a **Markov chain** if the state space is discrete, i.e. is finite or countable.

Markov process

- If for $t_1 < t_2 < t_3 < t_4 \ldots < t_n < t$,

$$P\{X(t) \le x \mid X(t_1) = x_1, X(t_2) = x_2, \ldots X(t_n) = X_n\}$$
$$= P\{X(t) \le x \mid X(t_n) = x_n\}$$

then the process $\{X(t)\}$ is called a Markov process.

Markov chain

- If for all n,

$$P\{X_n = a_n \mid X_{n-1} = a_{n-1}, X_{n-2} = a_{n-2}, \ldots X_0 = a_0\} = P\{X_n = a_n \mid X_{n-1} = a_{n-1}\}$$

then the process $\{x_n\}$, n = 0, 1, 2, ... is called a Markov chain.

- Consider a set of states, $S = \{s_1, s_2, s_3, \ldots s_r\}$. The process start with one state and moves successively to another state. Each move is called step.
- If the chain is currently in state S_i, then it moves to state S_j at the next step with a probability denoted by P_{ij}. The probability depends upon current states.
- The probabilities p_{ij} are called **transition probabilities**. The definition of the P_{ij} implies that the row sums of P are equal to 1.
- Markov chain is a random process $\{X_n, n = 0, 1, 2, \ldots\}$ where X_n belongs to the same subset of {0, 1, 2, 3, ...}, and

$P\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1} \ldots, X_1 = i_1, X_0 = i_0\} = P\{X_{n+1} = j \mid X_n = i\}$ for all states $i_0, i_1, \ldots i_{n-1}$ and all $n > 0$.

$\therefore$ $P_{ij} = P\{X_{n+1} = j \mid X_n = i\}$ then

$P_{ij} \ge 0$ for all i, j

For any i, $\sum_{j=1}^{\infty} P_{ij} = 1$

$P = [P_{ij}]$ is called one step transition probabilities.

2.4.1 The n-step Transistion Probabilities

- Given the chain is in state i at a given time, what is the probability it will be in state j after n transitions ?

$$P_{ij}^{n} = P\{X_{m+n} = j \mid X_m = i\}$$

$$= \sum_{k=0}^{\infty} P\{X_{m+n} = j \mid X_m = i, X_{m+1} = k\} P\{X_{m+1} = k \mid X_m = i\}$$

$$= \sum_{k=0}^{\infty} P\{X_{m+n} = j \mid X_{m+1} = k\} P\{X_{m+1} = k \mid X_m = i\}$$

$$= \sum_{k=0}^{\infty} P_{kj}^{n-1} P_{ik}$$

- Markov chain is characterized by three components :
  1. State space (s)
  2. Initial distribution (P^0)
  3. Markov kernel (transition)
- State space is the range of all random variables. The initial distribution quantifies the starting configuration for the chain.
- Kernel quantifies the probability of transitioning from state x_i to x_j so it establishes how the chain evolves.

2.4.2 Transition Probability Matrix of a Markov Chain

- The transition probabilities can be arranged as transition probability matrix $P = p_{i,j}$.

$$P = \begin{bmatrix} p_{0,0} & p_{0,1} & p_{0,2} & p_{0,3} \\ p_{1,0} & p_{1,1} & p_{1,2} & p_{1,3} \\ p_{2,0} & p_{2,1} & p_{2,2} & p_{2,3} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

- The row i contains the transition probabilities from state i to other state. The sum of the row probabilities is 1.
- A matrix with non-negative elements such that the sum of each row equals 1 is called a **stochastic matrix.** One can easily show that the product of two stochastic matrices is a stochastic matrix.
- The n-step transition probabilities :

$$p_{ij}^n = P[x_n = j \mid x_0 = i]$$

$i, j, n \geq 0$, p_{ij}^n is the probabilities of going from state i to state j in n step.

- The probability vector with r components is a row vector whose entries are non-negative and sum to 1. If 'u' is a probability vector which represents the initial state of a Markov chain, then the i^{th} component of 'u' as representing the probability that the chain starts in state s_i.

Example 2.4.1 *Given that a student last drink purchase was tea, there is a 90 % chance that his next drink purchase will also be tea. If a student last drink purchase was milk, there is an 80 % chance that his next drink purchase will also be milk. What is the probability that he will purchase tea two purchases from now ?*

Solution : Transition matrix

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

State diagram :

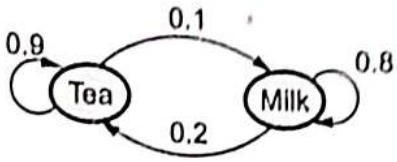


Fig. 2.4.1

$$P^2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

$$= \begin{bmatrix} 0.9\times0.9+0.1\times0.2 & 0.9\times0.1+0.1\times0.8 \\ 0.2\times0.9+0.8\times0.2 & 0.2\times0.1+0.8\times0.8 \end{bmatrix}$$

$$P^2 = \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix}$$

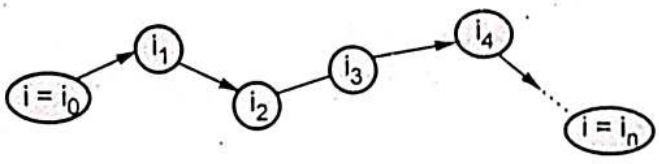
Probability that student will purchase tea two purchase from now

$= 0.2\times0.9 + 0.8\times0.2$

$= \mathbf{0.34}$

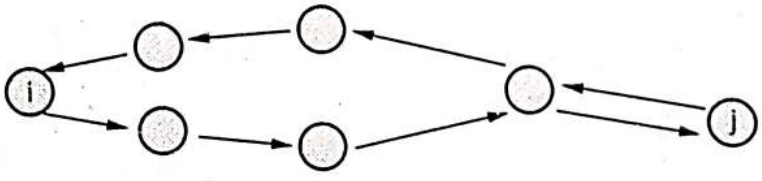
2.4.3 Classification of States of a Markov Chain

1. State i leads to state j ($i \rightarrow j$) :



2. State i and state j communicate ($i \leftrightarrow j$) :

if $i \rightarrow j$ and $j \rightarrow i$



3. A state i of a Markov chain is called a return chain if $p_{i,j}^{(n)} > 0$ for some $n \geq 1$.
4. A state i is said to be an absorbing state if and only if $p_{ij} = 1$.
5. The Markov chain is said to be 'irreducible' if there is only one class, i.e. all state communicate with each other.
6. A state with period 1 is said to be aperiodic.
7. The Markov chain does not drift away to infinity.
8. A Markov chain is **absorbing** if it has at least one absorbing state and if from every state it is possible to go to an absorbing state.
9. A Markov chain is called a **regular chain** if some power of the transition matrix has only positive elements.
10. If $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$.

- **Irreducible** means that every state is accessible from every other state. **Aperiodic** means that there exists at least one state to itself is possible.
- Positive recurrent means that for every state, the expected return time is finite. If the Markov chain is positive recurrent, there exist a stationary distribution.
- Communication is an equivalence relation : The states can be grouped into equivalent classes so that within each class all the states communicate with each other and two states from two different classes never communicate with each other. The equivalence classes defined by the relation $\leftrightarrow$ are called the **irreducible** classes of states.
- A Markov chain with a state space which is an irreducible class is called **irreducible.**
- A set of states is closed, if none of its states leads to any of the states outside the set. A single state which alone forms a closed set is called an **absorbing state.**
- Each state is either transient or recurrent,
  1. A state i is transient if the probability of returning to the state is < 1.
  2. A state i is recurrent if the probability of returning to the state is = 1.
- If the first return time of state i can only be a multiple of an integer d > 1 the state i is called **periodic.** Otherwise the state is aperiodic. An aperiodic positive recurrent state is ergodic. A Markov chain is ergodic, if and only if all its states are ergodic.

Examples of Markov chains

1. Input/output in Markov chain

Input

a. Transition probability matrix

b. Initial condition

Output

a. Probability of final/steady state

b. Probability of a specific state at a specific period.

Example 2.4.2 *Two brands of shaving cream compete for customers in the same market. The brand switching or transition probabilities at time of purchases are as follows :*

	Brand A	*Brand B*
Brand A	*0.9*	*0.1*
Brand B	*0.05*	*0.95*

1. *Which brand has the most loyal customers ?*
2. *What are the projected market shares for the two brands ?*

Solution : 1) Brand B appears to have more loyal customer because only 5 % of its customers switch to brand A, while 10 % of brand A customer switch to brand B.

2) Market share

Brand A $\pi_1 = 0.90\,\pi_1 + 0.05\,\pi_2$

Brand B $\pi_2 = 0.10\,\pi_1 + 0.95\,\pi_2$

So,

$$\pi_1 + \pi_2 = 1$$

$$\pi_2 = 1 - \pi_1$$

$$\pi_1 = 0.90\,\pi_1 + 0.05\,\pi_2$$

$$= 0.90\,\pi_1 + 0.05(1 - \pi_1)$$

$$= 0.90\,\pi_1 + 0.05 - 0.05\,\pi_1$$

$$\pi_1 = 0.85\,\pi_1 + 0.05$$

$$\pi_1 - 0.85\,\pi_1 = 0.05$$

$$\pi_1(1 - 0.85) = 0.05$$

$$0.15\,\pi_1 = 0.05$$

$$\pi_1 = \frac{0.05}{0.15}$$

$$\pi_1 = 0.3333$$

$$\pi_2 = 0.10\,\pi_1 + 0.95\,\pi_2$$

Substitute value of π_1

$$\pi_2 = 0.10\ (0.3333) + 0.95\ \pi_2$$

$$\pi_2 = 0.03333 + 0.95\ \pi_2$$

$$\pi_2\ (1 - 0.95) = 0.03333$$

$$0.05\ \pi_2 = 0.03333$$

$$\pi_2 = \frac{0.03333}{0.05}$$

$$\pi_2 = 0.6666$$

Brand B is the leader with 66.66 % of the market.

Example 2.4.3 *The transition probability matrix of a Markov chain is given by*

$\begin{bmatrix} 0.3 & 0.7 & 0 \\ 0.1 & 0.4 & 0.5 \\ 0 & 0.2 & 0.8 \end{bmatrix}$. *Is this matrix irreducible ?*

Solution : Consider the 3 states : state 0, state 1 and state 2. Here we go from state 0 to state 1 with a probability of 0.7.

We also go from state 1 to state 2 with probability of 0.5. So it is also possible that to go from state 0 to state 2.

All the states are recurrent so chain is irreducible.

Example 2.4.4 *Find the following given matrices are regular or not.*

$$P = \begin{bmatrix} 0.75 & 0.25 & 0 \\ 0 & 0.5 & 0.5 \\ 0.6 & 0.4 & 0 \end{bmatrix}$$

Solution : Calculate square of given matrices.

$$P^2 = P \cdot P$$

$$= \begin{bmatrix} 0.75 & 0.25 & 0 \\ 0 & 0.5 & 0.5 \\ 0.6 & 0.4 & 0 \end{bmatrix} \begin{bmatrix} 0.75 & 0.25 & 0 \\ 0 & 0.5 & 0.5 \\ 0.6 & 0.4 & 0 \end{bmatrix}$$

$$P^2 = \begin{bmatrix} 0.5625 & 0.3125 & 0.125 \\ 0.3 & 0.45 & 0.25 \\ 0.45 & 0.35 & 0.2 \end{bmatrix}$$

All the entries are non-negative and positive, so given matrix P is regular.

Example 2.4.5 *Three boys A, B and C are throwing a ball to each other. A always throws to B and B always throws to C, but C is as likely to throw the ball to B as to A. Find the TPM and classify the states.*

Solution :

$$P = \begin{matrix} \\ A \\ B \\ C \end{matrix} \begin{matrix} \begin{matrix} A & B & C \end{matrix} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix} \end{matrix}$$

$$P^2 = P \cdot P$$

$$= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

$$P^2 = \begin{bmatrix} 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix}$$

$$P^3 = P^2 \cdot P$$

$$= \begin{bmatrix} 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

$$P^3 = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$$

The state with period 1 is aperiodic all states are ergodic.

Example 2.4.6 *Given transition probability matrix*

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1 & 0 \end{bmatrix}$$

with Markov chain of state space $S = \{0, 1, 2\}$. *Also* $\{X_n ; n = 1, 2, 3\}$. ***Find the Markov chain is ergodic or not.***

Solution : $P^2 = P \times P$

$$= \begin{bmatrix} 0 & 1 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0\times0+1\times\frac{1}{4}+0\times0 & 0\times1+1\times\frac{1}{2}+0\times1 & 0\times0+1\times\frac{1}{4}+0\times1 \\ 1/4\times0+\frac{1}{2}\times\frac{1}{4}+\frac{1}{4}\times0 & 1/4\times1+\frac{1}{2}\times\frac{1}{2}+\frac{1}{4}\times1 & \frac{1}{4}\times0+\frac{1}{2}\times\frac{1}{4}+\frac{1}{4}\times0 \\ 0\times0+1\times\frac{1}{4}+0\times0 & 10\times1+1\times\frac{1}{2}+0\times1 & 0\times0+1\times\frac{1}{4}+0\times1 \end{bmatrix}$$

$$P^2 = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{8} & \frac{3}{4} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

$$P^3 = P^2 \times P = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{8} & \frac{3}{4} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 1 & 0 \end{bmatrix}$$

$$P^3 = \begin{bmatrix} \frac{1}{8} & \frac{3}{4} & \frac{1}{8} \\ \frac{3}{16} & \frac{5}{8} & \frac{3}{16} \\ \frac{1}{8} & \frac{3}{4} & \frac{1}{8} \end{bmatrix}$$

State is ergodic because $P_{11}^3 > 0$, $P_{21}^2 > 0$, $P_{33}^2 > 0$ etc.

Example 2.4.7 *Which of the following stochastic matrix are regular ?*

a) $\begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix}$ *b)* $\begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$

Solution :

a) $\begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix}$

For a given matrix, diagonal value is 1. So matrix is not regular.

b) $\begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$

$$B^2 = B \cdot B = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 1/4 & 1/4 & 1/2 \end{bmatrix} \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 1/2 \end{bmatrix}$$

$$= \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 3/8 & 3/8 & 1/4 \end{bmatrix}$$

$$B^3 = B^2 \cdot B = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 3/8 & 3/8 & 1/4 \end{bmatrix} \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$$

$$B^3 = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 7/16 & 7/16 & 1/8 \end{bmatrix}$$

So given matrix is not regular because of two zeros in the B^3 matrix.

Example 2.4.8 *Which of the following matrices are stochastic ?*

i) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ *ii)* $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ *iii)* $\begin{bmatrix} 0 & 2 \\ 1/3 & 1/4 \end{bmatrix}$

Solution :

i) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Given matrix is square matrix and sum of the elements in each row is equal to 1. Therefore matrix is stochastic.

ii) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$

Given matrix is not square matrix. It is not stochastic.

iii) $\begin{bmatrix} 0 & 2 \\ 1/3 & 1/4 \end{bmatrix}$

Given matrix is square matrix but sum of each row is not equal to 1. So it is not stochastic matrix.

Example 2.4.9 *A computer system can operate in two different modes. Every hour, it remains in the same mode or switches to a different mode according to the transition probability matrix.*

$$P = \begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}$$

i) Compute the 2-step transition probability matrix.

ii) If the system is in mode I at 5:30 pm, what is the probability that it will be in mode I at 8:30 pm on the same day ? **SPPU : Apr.-19 (In Sem), Marks 6**

Solution :

i) The 2-step transition probability matrix : Since the Markov chain is homogeneous 2-step transition probability matrix is given by,

$$P^{(2)} = P^2 = \begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix}\begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.52 & 0.48 \\ 0.48 & 0.52 \end{pmatrix}$$

ii) If the system is in Mode I at 5:30 pm, what is the probability that it will be in Mode I at 8:30 pm on the same day ?

$$P(X_{8:30} = j | X_{5:30} = i) = P(X_3 = j | X_0 = i) = P_{i,j}^{(3)}$$

Therefore

$$P^{(3)} = P^3 = \begin{pmatrix} 0.52 & 0.48 \\ 0.48 & 0.52 \end{pmatrix}\begin{pmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.496 & 0.504 \\ 0.504 & 0.496 \end{pmatrix}$$

So $P_{i,i}^{(3)} = 0.496$ for $i = 1, 2$.

Example 2.4.10 *Assume that a man's profession can be classified as professional, skilled labourer or unskilled labourer. Assume that of the sons of professional men, 80 percent are professional, 10 percent are skilled labourers and 10 percent are unskilled labourers. In the case of sons of skilled labourers, 60 percent are skilled labourers, 20 percent are professional and 20 percent are unskilled. Finally, in the case of unskilled labourers, 50 percent of the sons are unskilled labourers and 25 percent each are in the other two categories. Assume that every man has at least one son and form a Markov chain by following the profession of a randomly chosen son of a given family through several generations. Set up the matrix of transition probabilities. Find the probability that a randomly chosen grandson of an unskilled labourer is a professional man.* **SPPU : Apr.-19 (In Sem), Marks 6**

Solution : • The Markov chain in this exercise has the following set states :

S = {Professional, Skilled, Unskilled} with the following transition probabilities

	Professional	Skilled	Unskilled
Professional	0.8	0.1	0.1
Skilled	0.2	0.6	0.2
Unskilled	0.25	0.25	0.5

So that the transition matrix for this chain is,

$$P = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

With $$P^2 = \begin{pmatrix} 0.685 & 0.165 & 0.150 \\ 0.330 & 0.430 & 0.240 \\ 0.375 & 0.300 & 0.325 \end{pmatrix}$$

and thus the probability that a randomly chosen grandson of an unskilled labourer is a professional man is 0.375.

2.4.4 Tail Bound

- In probabilistic analysis, we often need to bound the probability that a random variable deviate far from its mean. There are various formulas for this purpose. These are called **tail bounds.** The weakest of these is the Markov bound, which states that for any nonnegative random variable X with mean $\mu = \xi X$,

$$\Pr(X \geq k) \leq \mu/k$$

- A better bound is the Chebyshev bound, which states that for a random variable X with mean $\mu = EX$ and standard deviation $\sigma = \sqrt{\xi (X - \mu)^2}$, for any $\delta \geq 1$,

$$\Pr(|X - \mu| \geq \delta\sigma) \leq \delta^{-2}$$

- The tail bounds are used to satisfy the following :
  1. Measure the confident interval of a statistic within a certain distance of parameter.
  2. What is the probability that the parameter is within a certain range that includes our sample statistic.
- The Markov and Chebyshev bounds converge linearly and quadratically, respectively and are often too weak to achieve desired estimates.

- In machine learning, tail bounds help quantifying the extraction of information from large data sets by estimating the probability for a learning algorithm to be approximately correct. Typical bounds quantify the deviation of sample means from the exact expectation. They help understanding concentration of measure phenomena in high-dimensional geometry underlying unsupervised learning concepts. They are important tools in statistical estimation.

2.4.5 Random Walks

- Random walk is a special type of Markov chain.
- In probability theory, a process for determining the probable location of a point subject to random motions, given the probabilities of moving some distance in some direction. Random walks are an example of Markov processes, in which future behaviour is independent of past history.
- Random walk is the process by which randomly-moving objects wander away from where they started.
- The simplest random walk to understand is a 1-dimensional walk. Suppose that the black dot below is sitting on a number line. The black dot starts in the center.

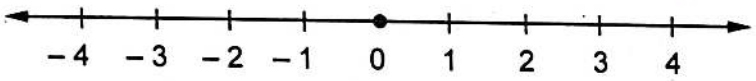


Fig. 2.4.2

- Then, it takes a step, either forward or backward, with equal probability. It keeps taking steps either forward or backward each time. Let's call the 1st step a_1, the second step a_2, the third step a_3 and so on. Each "a" is either equal to + 1 (if the step is forward) or – 1 (if the step is backward). The Fig. below shows a black dot that has taken 5 steps and ended up at – 1 on the number line.

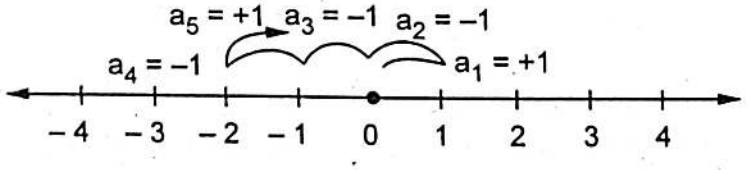


Fig. 2.4.3 Random walk

- Random walk theory suggests that changes in stock prices have the same distribution and are independent of each other.

Types of random walk :

1. **Sample random walk :** Simple random walk is a regular random walk on a regular lattice, forming a lattice path for each step the location jumps to another site according to some probability distribution. Location can only jump to neighboring sites of the lattice forming a lattice path.
2. **One dimensional random walk :** One dimensional random walk is like a square game where you start at one square and move one step ahead and find out the neighbourer square to jump left on right side, depending upon the probability distribution.

2.4.6 Pair-wise Independence

- The random variables $X_1; X_2; : : : ; X_n$ are said to be pair-wise independent if, for all $i \neq j$ and any values a ; b

$$\Pr((X_i = a) \cap (X_j = b)) = \Pr(X_i = a) \Pr(X_j = b)$$

- Pair-wise independence is a much weaker requirement than mutual independence. It is used in data dependency issues. Pair-wise independence come from the natural independence.
- A random bit Y is uniform if $\Pr(Y = 0) = \Pr(Y = 1) = 1/2$. We show a method to derive $m = 2^b - 1$ uniform and pairwise independent bits from b mutually independent uniform random bits $X_1; : : : ; X_b$.
- Enumerate the $m = 2^b - 1$ nonempty subsets of {1; 2; : : : ; b} in some order and let S_j denote the j^{th} subset.
- Define Y_j

$$Y_j = \left(\sum_{i \in S_j} X_i \right) \bmod 2$$

2.4.7 Universal Hashing

- No matter how we choose our hash function, it is always possible to devise a set of keys that will hash to the same slot, making the hash scheme perform poorly.
- To circumvent this, we randomize the choice of a hash function from a carefully designed set of functions. Let U be the set of universe keys and H be a finite collection of hash functions mapping U into {0, 1, ... , m – 1}. Then H is called universal if,

for $x, y \in U, (x \neq y)$,

$$| \{h H \in : h(x) = h(y)\} | = \frac{|H|}{m}$$

- In other words, the probability of a collision f or two different keys x and y given a hash function randomly chosen from H is 1/m.
- Creating set of hash functions :

1. Choose the table size m to be prime.
2. Decompose the key x into r + 1 "bytes" so that $x = (x_0, x_1, \ldots, x_r)$, where the maximal value of any x_i is less than m.
3. Let $a = (a_0, a_1, \ldots, a_r)$ denote a sequence of r + 1 elements choosen randomly such that $a_i \in \{0, 1, \ldots, m-1\}$. There are m^{r+1} possible such sequence.
4. Define a hash function h_a with $h_a(x) = \sum_{i=0}^{r} a_i x_i \bmod m$.
5. $H = \bigcup_a \{h_a\}$ with m^{r+1} members, one for each possible sequence a.

Example 2.4.11 *Find the first three powers of following transition matrix using Markov chain.*

$$D = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

SPPU : Apr.-18 (In Sem), Marks 6

Solution :

$$D = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

$$D^2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}\begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

$$= \begin{bmatrix} 0.9\times0.9+0.1\times0.2 & 0.9\times0.1+0.1\times0.8 \\ 0.2\times0.9+0.8\times0.2 & 0.2\times0.1+0.8\times0.8 \end{bmatrix}$$

$$D^2 = \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix}$$

$$D^3 = \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix}\begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

$$D^3 = \begin{bmatrix} 0.781 & 0.219 \\ 0.438 & 0.562 \end{bmatrix}$$

Example 2.4.12 *Solve following problem using Markov chain.*

In the dark ages, Harvard, Dartmouth and Yale admitted as per below scenario. Assume that, 80 % of the sons of Harvard men went to Harvard and rest went to Yale. 40 % of the sons of Yale men went Yale and rest split evenly between Harvard and Dartmouth. Of the sons of Dartmouth men, 70 % went to Dartmouth, 20 % to Harvard and 10 % to Yale.

i) Find the probability that the grandson of a man from Harvard man went to Harvard.

ii) Modify the above by assuming that the son of a Harvard man always went to Harvard. Again find the probability that the grandson of a man from Harvard went to Harvard.

SPPU : Apr.-18 (In Sem), Marks 6

Solution : i) The probability that the grandson of a man from Harvard went to Harvard.

- We first form a Markov chain with state space S = {H, D, Y} and the following transition probability matrix :

$$P = \begin{array}{c} \\ H \\ Y \\ D \end{array} \begin{array}{c} \begin{array}{ccc} H & Y & D \end{array} \\ \begin{pmatrix} .8 & .2 & 0 \\ .3 & .4 & .3 \\ .2 & .1 & .7 \end{pmatrix} \end{array}$$

- Note that the columns and rows are ordered : First H, then D, then Y. Recall : The ij^{th} entry of the matrix P^n gives the probability that the Markov chain starting in state i will be in state j after n steps. Thus, the probability that the grandson of a man from Harvard went to Harvard is the upper-left element of the matrix.

$$P^2 = \begin{pmatrix} .7 & .06 & .24 \\ .33 & .52 & .15 \\ .42 & .33 & .25 \end{pmatrix}$$

- It is equal to (.8) (.8) + (.2) (.3) + (0) (.2) = 0.7

ii) If all sons of men from Harvard went to Harvard, this would give the following matrix for the new Markov chain with the same set of states :

$$P = \begin{pmatrix} 1 & 0 & 0 \\ .2 & .7 & .1 \\ .3 & .3 & .4 \end{pmatrix}$$

The upper-left element of P2 is 1, which is not surprising, because the offspring of Harvard men enter this very institution only.

Review Questions

1. *What is the application of tail and bound in big data.* **SPPU : May-18 (End Sem), Marks 4**
2. *Explain following terms : i) Expectation ii) Pairwise independence* **SPPU : Apr.-19 (In Sem), Marks 4**
3. *Explain pairwise independent hashing.* **SPPU : Dec.-19 (End Sem), Marks 4**

2.5 Data Streaming Models

- Conceptually, a data stream is a sequence of data items that collectively describe one or more underlying signals. For instance, a network traffic stream describes the type and volume of data transmitted among nodes in the network; one possible signal is a mapping between pairs of source and destination IP addresses to the number of bytes transmitted from the given source to the given destination.
- Computational model is called the **data streaming model.** In this model, elements going past in a stream and there is very little space for storing things. For example, a user might be running a program on an Internet router, the elements might be IP addresses and there is limited space.
- Fig. 2.5.1 shows streaming model.

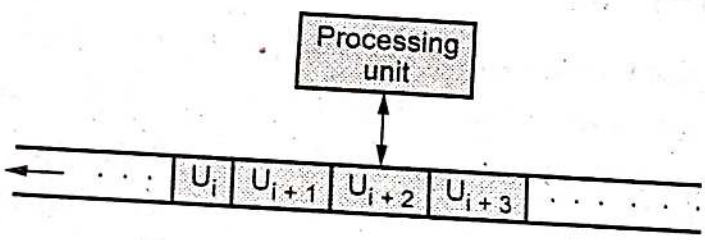


Fig. 2.5.1 Streaming model

- Data item arrive over time. The U_j processed before U_{j+1} arrives.
- You can imagine the applications of this model. An Internet router might see a lot of packets whiz by and may want to gure out which data connections are using the most space ?
- How many different connections have been initiated since midnight ? Or the median (or the 90 %) of the file sizes that have been transferred.
- Which IP connections are elephants ? Even if you are not working at "line speed", but just looking over the server logs, you may not want to spend too much time to find out the answers, you may just want to read over the file in one quick pass and come up with an answer.
- Such an algorithm might also be cache-friendly. **Two of the recurring themes** will be :

- **Approximate solutions :** In several cases, it will be impossible to compute the function exactly using small space. Hence we'll explore the trade-off between approximation and space.
- **Hashing :** This will be a very powerful technique.
- **Constraints :**
  1) Items can only be read sequentially in the order it arrives
  2) Items can be examined in only a few passes
  3) Limited working memory
  4) Limited processing time per item.

2.6 Flajole Martin Algorithm

SPPU : Apr.-18, May-18, Dec.-19

- The Flajolet-Martin (FM) algorithm can better solve the problem of estimating the number of independent elements in the data stream sequence.
- It is used for counting distinct elements in a given stream. It is an approximate algorithm to detect distinct elements.
- When counting the number of distinct elements in a stream, what we usually do is to keep a set in the memory and union the set with the new incoming element. This approach requires storing all unique elements in the main memory, which can be infeasible when the size of the stream is huge.
- The Flajolet - Martin algorithm proposes a way of estimating the number of unique elements in the stream without the need to store all elements. The basic idea is as follows : We generate the hash value of each record and count the number of trailing zeros in the hashed bits. Suppose the greatest trailing zero is R, then the number of unique elements is 2^R.
- Approximate neighborhood function algorithm uses Flajolet - Martin (FM) counters to maintain size estimates.
- Flajolet-Martin algorithm approximates the number of unique objects in a stream or a database in one pass. If the stream contains n elements with m of them unique, this algorithm runs in O(n) time and needs O(log(m)) memory.
- So the real innovation here is the memory usage, in that an exact, brute-force algorithm would need O(m) memory.
- The Flajolet - Martin algorithm :
  1. Create a bit vector of sufficient length L, such that $2^L > n$, the number of elements in the stream. Usually a 64-bit vector is sufficient since 2^{64} is quite large for most purposes.

2. The i-th bit in this vector/array represents whether we have seen a hash function value whose binary representation ends in 0^i. So initialize each bit to 0.
3. Generate a good, random hash function that maps input to natural numbers.
4. Read input. For each word, hash it and determine the number of trailing zeros. If the number of trailing zeros is k, set the k^{th} bit in the bit vector to 1.
5. Once input is exhausted, get the index of the first 0 in the bit array. By the way, this is just the number of consecutive 1s plus one.
6. Calculate the number of unique words as $2^R / \varphi$, where φ is 0.77351.
7. The standard deviation of R is a constant : $\sigma(R) = 1.12$.

To improve accuracy of this approximation algorithm, we do the following :

8. (Averaging) Use multiple hash functions and use the average R instead.
9. (Bucketing) Averages are susceptible to large fluctuations. So use multiple buckets of hash functions from the above step and use the median of the average R. This gives fairly good accuracy.
10. Overall accuracy of this algorithm can be tuned by using appropriate number of hash functions in the averaging and bucketing steps. Of course, if more accuracy is desired, more hash functions need to be used, which implies higher computation cost.

Example 2.6.1 *Determine distinct elements in below input stream of integers using Flajolet algorithm. Consider hash function $h(X) = 6X + 1$, X = 1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1.*

SPPU : Apr.-18 (In Sem), Marks 4

Solution : Hash function $h(X) = (6X + 1) \bmod 5$

After h(a) the stream will become

2, 4, 3, 2, 3, 4, 0, 4, 2, 3, 4, 2

Convert it into binary :

010, 100, 011, 010, 011, 100, 000, 100, 010, 011, 100, 010

Trailing 0's

1, 2, 0, 1, 0, 2, 0, 2, 1, 0, 2, 1

$$R = \max r(a) = 2$$

$$E = 2^R = 2^2 = 4$$

Four distinct element found in a stream = 1, 2, 4, 7.

Example 2.6.2 *Explain the Flajolet Martin distance sampling. Find the distinct element from the element stream 1, 4, 2, 1, 2, 4, 4, 4, 1, 2, 4, 1, 7. Assume suitable hash function.*

SPPU : May-18 (End Sem), Marks 6

Solution : Refer section 2.6.

Distinct element = 4

$h(a) = (3X + 1) \bmod 5$ (given hash function)

After h(a) the stream will become

4, 3, 2, 4, 2, 3, 3, 3, 4, 2, 3, 4, 2

Convert it into binary

100, 011, 010, 100, 010, 011, 011, 011, 100, 010, 011, 100, 010

Trailing 0's

$r(a) = 2, 0, 1, 2, 1, 0, 0, 0, 2, 1, 0, 2, 1$

$R = \max r(a) = 2$

$E = 2^R = 2^2 = 4$

i.e. 4 distinct element found in a stream = 1, 2 ,4, 7.

Example 2.6.3 *Explain the Flajolet Martin distance sampling. Find the distinct element from the element stream 4, 2, 5, 9, 1, 6, 3, 7. Consider Hash function $h(x) = (x+6)$ mod 32.*

SPPU : Dec.-19 (End Sem), Marks 4

Solution :

$h(x) = (x + 6) \bmod 32$

$h(4) = (4 + 6) \bmod 32 = 10 \bmod 32 = 10 = (01010)$

$h(2) = (2 + 6) \bmod 32 = 8 \bmod 32 = 8 = (01000)$

$h(5) = (5 + 6) \bmod 32 = 11 \bmod 32 = 11 = (01011)$

$h(9) = (9 + 6) \bmod 32 = 15 \bmod 32 = 15 = (01111)$

$h(1) = (1 + 6) \bmod 32 = 7 \bmod 32 = 7 = (00111)$

$h(6) = (6 + 6) \bmod 32 = 12 \bmod 32 = 12 = (01110)$

$h(3) = (3 + 6) \bmod 32 = 9 \bmod 32 = 9 = (01001)$

$h(7) = (7 + 6) \bmod 32 = 13 \bmod 32 = 13 = (01101)$

Trailing zero's {1, 3, 0, 0, 0, 1, 0, 0}

$R = \max$ [Trailing Zero] = 3

Output $= 2^R = 2^3 = 8$

Review Question

1. *Explain Flajolet Martin algorithm. List the limitations of algorithm and how will you overcome these limitations ?* **SPPU : Apr.-19 (In Sem), Marks 4**

2.7 Distance Sampling and Random Projections

- Often in data science, it can be very difficult to work with features that are very high-dimensional. This is because data in high-dimensions cannot be analyzed by computers or humans because it is simply too complex to draw conclusions from. So in data science, there are a series of techniques that can be utilized in order to reduce the dimensions of this data.
- The most popular method for reducing the dimensions of high-dimensional data is most certainly singular value decomposition. However, there are many lesser-known methods of decomposition. One example of such a technique is a technique called **random projection.**
- Random projection is another decomposition method that is used to reduce the dimensionality of high-dimensional data.
- The method of random projections is a simple yet powerful dimensionality reduction technique that uses random projection matrices to project the data into lower dimensional spaces.
- Random projections have been used in Machine Learning, VLSI layout, analysis of Latent Semantic Indexing (LSI), set intersections , finding motifs in bio-sequences, face recognition, privacy preserving distributed data mining.
- The dimensions and distribution of random projections matrices are controlled so as to preserve the pairwise distances between any two samples of the dataset. Thus, random projection is a suitable approximation technique for distance-based methods.
- In random projection, the original d-dimensional data is projected to a k-dimensional ($k << d$) subspace through the origin, using a random "k x d" matrix R whose columns have unit lengths. Using matrix notations where $X_{d \times N}$ is the original set of N d-dimensional observations,

$$X_{k \times N}^{RP} = R_{k \times d} X_{d \times N}$$

is the projection of the data onto a lower k-dimensional subspace.

- The key idea of random mapping arises from the Johnson-Lindenstrauss lemma : If points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distance between the points are approximately preserved.

2.8 Bloom Filters

SPPU : April-18, 20

- Bloom filter is a space-efficient probabilistic data structure that is used to test whether an element is a member of a set.
- For example, checking availability of username is set membership problem, where the set is the list of all registered username.
- The price we pay for efficiency is that it is probabilistic in nature that means, there might be some false positive results. False positive means, it might tell that given username is already taken but actually it's not.

Interesting Properties of Bloom Filters :

1. Bloom filter of a fixed size can represent a set with an arbitrarily large number of elements.
2. Adding an element never fails.
3. Bloom filters never generate false negative result, i.e., telling you that a user name doesn't exist when it actually exists.
4. Deleting elements from filter is not possible because, if we delete a single element by clearing bits at indices generated by k hash functions, it might cause deletion of few other elements.

- An empty Bloom filter is a bit array of m bits, all set to 0. There must also be k different hash functions defined, each of which maps or hashes some set element to one of the m array positions with a uniform random distribution.

0	0	0	0	0	0	0	0	0	0	0

- To add an element, feed it to each of the k hash functions to get k array positions. Set the bits at all these positions to 1.

0	1	0	1	0	1	0	0	0	0	0

- The algorithm is based on a bit vector of size m and k independent and uniformly distributed hash functions. When a new element is handled by the filter, it's hashed against each of the functions. Their results correspond to the bit vector indexes that will be set to 1. The same operation is made to check membership. The algorithm reads the values from the computed indexes and depending on the result, it tells whether :

a) If any of the bits at these positions are 0, the element is definitely not in the set, if it were, then all the bits would have been set to 1 when it was inserted.

b) If all are 1, then either the element is in the set, or the bits have by chance been set to 1 during the insertion of other elements, resulting in a false positive.

- Fig. 2.8.1 shows Bloom filter concept.

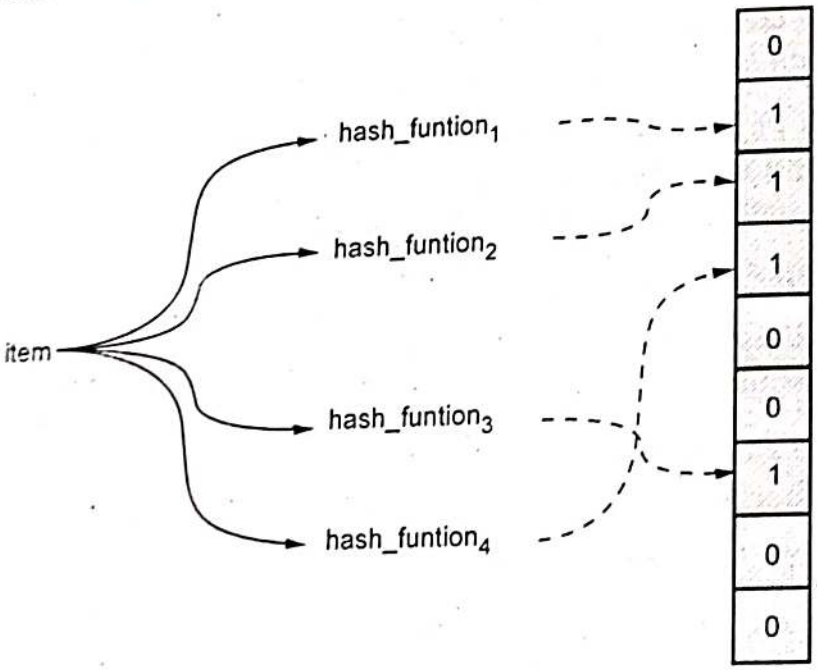


Fig. 2.8.1 Bloom filter concept

- *Bloom filter never returns false negatives. The efficiency of the Bloom filter relies on the size of the bit vector. The precision decreases as soon as the vector fulfills, since every value will be returned as present. Thus it is not a good solution for unbounded data.*
- *Bloom filter was invented in 1970 by Burton Howard Bloom. Some of its more common applications include :*
  a) *Determining whether a user ID or domain is already taken*
  b) *Reducing disk lookups for the non-existing keys in a database*
  c) *Filtering out previously shown posts on recommendation engines*
  d) *Checking words for misspellings and profanity with a spellchecker*
  e) *Identifying malicious URLs, blocked IPs and fraudulent transactions*
  f) *Counting the number of active users or total unique clicks on a website*
  g) *Determining heavy hitters.*
- **Advantages of Bloom filter :**
  a) *It uses constant space, regardless of the number of elements inserted.*
  b) *No false negatives, so you can trust the Bloom filter when it says the item does not exist.*

c) Adding an element never fails.

d) It does not store the actual elements, ensuring privacy out of the box.

- **Disadvantages of Bloom filter :**

a) It can return false positives.

b) Adding elements never fails.

c) Reducing false-positive rates requires an additional bit array or recreation of the Bloom filter.

d) It cannot retrieve the inserted elements.

e) It cannot delete the inserted elements.

Review Questions

1. *Explain Bloom filter with proper example.* **SPPU : April.-18 (In Sem), Marks 4**
2. *Explain Bloom filter with example. State the applications of Bloom filter in big data.* **SPPU : April-20 (In Sem), Marks 4**

2.9 Mode

It is the value of maximum frequency. It occurs most frequency.

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} h$$

L = Lower limit of the class containing mode

Δ_1 = Excess of modal frequency over frequency of preceding class

Δ_2 = Excess of modal frequency over following class

h = Size of modal class

2.9.1 Mean

Let $x_1, x_2, x_3, \ldots, x_n$ be the set 'n' values of the variate, then arithmetic mean or mean is given as,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

- In the frequency distribution if $x_1, x_2, \ldots, x_n$ are the midvalues of the class intervals with frequencies $f_1, f_2, \ldots, f_n$ respectively, then we have,

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum f_i x_i}{\sum f_i}$$

2.9.2 Variance

- The second central moment is called **variation**. It is given as,

$$\sigma_x^2 = \text{Var}[X] = E[(X - m_x)]^2$$

$$= \int_{-\infty}^{\infty} (x - m_x)^2 f_X(x) dx$$

or $$\sigma_x^2 = \overline{X^2} - m_x^2 = E[X^2] - m_x^2$$

- Variance can also be given as,

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

Let 'A' be assumed mean, 'h' be the magnitude of the class interval and let $d = \frac{x - A}{h}$

Then, $$\text{mean } \bar{x} = A + \frac{h \sum fd}{N} \text{ and } \sigma_x^2 = h^2 \left[\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2 \right]$$

2.9.3 Standard Deviation

- It is the measure of spread over the values of 'X' relative to mean value. It is give as,

$$\sigma_x = \sqrt{\text{Variance}} = \sqrt{E[X^2] - m_x^2}$$

$$\text{S.D. } \sigma_x = \sqrt{\frac{1}{N} \sum_i f_i (x_i - \bar{x})^2}$$

2.9.4 Difference between Standard Deviation and Variance

Sr. No.	Standard deviation	Variance
1.	Standard deviation is a measure of dispersion of the values of a data set from their mean.	It is the statistical measure of how fa the numbers are spread in a data se from their average.
2.	It is a common term in statistical theory to calculate central tendency.	Variance is primarily used for statistica probability distribution to measur volatility from the mean.
3.	It measures the absolute variability of the dispersion.	It helps determine the size of the da spread.

4.	It is calculated by taking the square root of the variance.	It is calculated by taking the average of the squared deviation of each value in the data set from the mean.
5.	The standard deviation is symbolized by the Greek letter sigma "σ" as in lower case sigma.	The notation for the variance of a variable is "σ^2" sigma squared.
6.	$\sigma = \sqrt{\frac{\sum (x - M)^2}{n}}$ where M = Mean, x = A values in a data set and n = Number of values.	$\sigma^2 = \frac{\sum (x - M)^2}{n}$ where M = Mean, x = Each value in the data set, n = Number of values in the data set.
7.	Used in the finance sector as a measure of market and security volatility.	Used in asset allocation.

Example 2.9.1 *Find sample mean and sample standard deviation for the following data set : 5, 10, 15, 20.*

Solution : Sample mean ($\bar{x}$) :

$$\bar{x} = \frac{\sum x}{n} = \frac{5 + 10 + 15 + 20}{4} = \frac{50}{4} = \mathbf{12.5}$$

Sample standard deviation :

Data	$x - \bar{x}$	$(x - \bar{x})^2$
5	5 – (12.5) = – 7.5	$(-7.5)^2 = 56.25$
10	10 – (12.5) = – 2.5	$(-2.5)^2 = 6.25$
15	15 – (12.5) = 2.5	$(2.5)^2 = 6.25$
20	20 – (12.5) = 7.5	$(7.5)^2 = 56.25$
		$\sum (x - \bar{x})^2 = 125.01$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{125.01}{4 - 1}} \approx \mathbf{6.455}$$

Example 2.9.2 *The heights at the shoulders are : 600 mm, 470 mm, 170 mm, 430 mm and 300 mm. Find out the mean and the variance.*

Solution :

Your first step is to find the mean : $\frac{600 + 470 + 170 + 430 + 300}{5} = \mathbf{394}$

Variance :

$600 - 394 = 206$

$470 - 394 = 76$

$170 - 394 = -224$

$430 - 394 = 36$

$300 - 394 = -94$

$$\text{Variance} = \frac{(206)^2 + (76)^2 + (-224)^2 + (36)^2 + (-94)^2}{5}$$

$$\text{Variance} = \frac{42436 + 5776 + 50176 + 1296 + 8836}{5} = \mathbf{21704}$$

2.10 Correlation Analysis

- When one measurement is made on each observation, uni-variate analysis is applied. If more than one measurement is made on each observation, multivariate analysis is applied. Here we focus on bivariate analysis, where exactly two measurements are made on each observation.
- The two measurements will be called X and Y. Since X and Y are obtained for each observation, the data for one observation is the pair (X, Y).
- Some examples :
  1. Height (X) and weight (Y) are measured for each individual in a sample.
  2. Stock market valuation (X) and quarterly corporate earnings (Y) are recorded for each company in a sample.
  3. A cell culture is treated with varying concentrations of a drug and the growth rate (X) and drug concentrations (Y) are recorded for each trial.
  4. Temperature (X) and precipitation (Y) are measured on a given day at a set of weather stations.
- There is difference in bivariate data and two sample data. In two sample data, the X and Y values are not paired and there are not necessarily the same number of X and Y values.
- **Correlation** refers to a relationship between two or more objects. In statistics, the word correlation refers to the relationship between two variables. Correlation exists between two variables when one of them is related to the other in some way.
- **Examples :** One variable might be the number of hunters in a region and the other variable could be the deer population. Perhaps as the number of hunters increases,

the deer population decreases. This is an example of a **negative correlation** : As one variable increases, the other decreases.

- A **positive correlation** is where the two variables react in the same way, increasing or decreasing together. Temperature in Celsius and Fahrenheit has a positive correlation.
- The term "correlation" refers to a measure of the strength of association between two variables.
- **Covariance** is the extent to which a change in one variable corresponds systematically to a change in another. Correlation can be thought of as a standardized covariance.
- The correlation coefficient r is a function of the data, so it really should be called the sample correlation coefficient. The (sample) correlation coefficient r estimates the population correlation coefficient ρ.
- If either the X_i or the Y_i values are constant (i.e. all have the same value), then one of the sample standard deviations is zero, and therefore the correlation coefficient is not defined.

2.10.1 Types of Correlation

1. Positive and negative
2. Simple and multiple
3. Partial and total
4. Linear and non-linear.

1. Positive and negative

- **Positive correlation :** Association between variables such that high scores on one variable tends to have high scores on the other variable. A direct relation between the variables.
- **Negative correlation :** Association between variables such that high scores on one variable tends to have low scores on the other variable. An inverse relation between the variables.

2. Simple and multiple

- Simple : It is about the study of only two variables, the relationship is described as simple correlation.
- Example : Quantity of money and price level, demand and price.
- Multiple : It is about the study of more than two variables simultaneously, the relationship is described as multiple correlations.
- Example : The relationship of price, demand and supply of a commodity.

3. Partial and total correlation

- Partial correlation : analysis recognizes more than two variables but consider only two variables keeping the other constant. Example : Price and demand eliminating the supply side.
- Total correlation is based on all the relevant variables, which is normally no feasible. In total correlation, all the facts are taken into account.

4. Linear and non-linear correlation

- Linear correlation : Correlation is said to be linear when the amount of change one variable tends to bear a constant ratio to the amount of change in the othe The graph of the variables having a linear relationship will form a straight line.
- Non linear correlation : The correlation would be non linear if the amount change in one variable does not bear a constant ratio to the amount of change the other variable.

Classification of correlation

- Two methods are used for finding relationship between variables.
  1. Graphic methods
  2. Mathematical methods.
- Graphic methods contain two sub methods : **Scatter diagram and simple graph.**
- Types of mathematical methods are,
  a. Karl 'Pearson's coefficient of correlation
  b. Spearman's rank coefficient correlation
  c. Coefficient of concurrent deviation
  d. Method of least squares.

2.10.2 Scatter Diagram

- When two variables x and y have an association (or relationship), we say there exists a **correlation** between them. Alternatively, we could say x and y are *correlated*. To find such an association, we usually look at a scatterplot and try to find a pattern.
- Scatterplot (or scatter diagram) is a graph in which the paired (x, y) sample data are plotted with a horizontal x axis and a vertical y axis. Each individual (x, y) pair is plotted as a single point.
- One variable is called independent (X) and the second is called dependent (Y)

- **Example :**

Weight (kg)	67	69	85	83	74	81	97	92	114	85
Blood pressure (mmHg)	120	125	140	160	130	180	150	140	200	130

- Fig. 2.10.1 shows the scatter diagram.

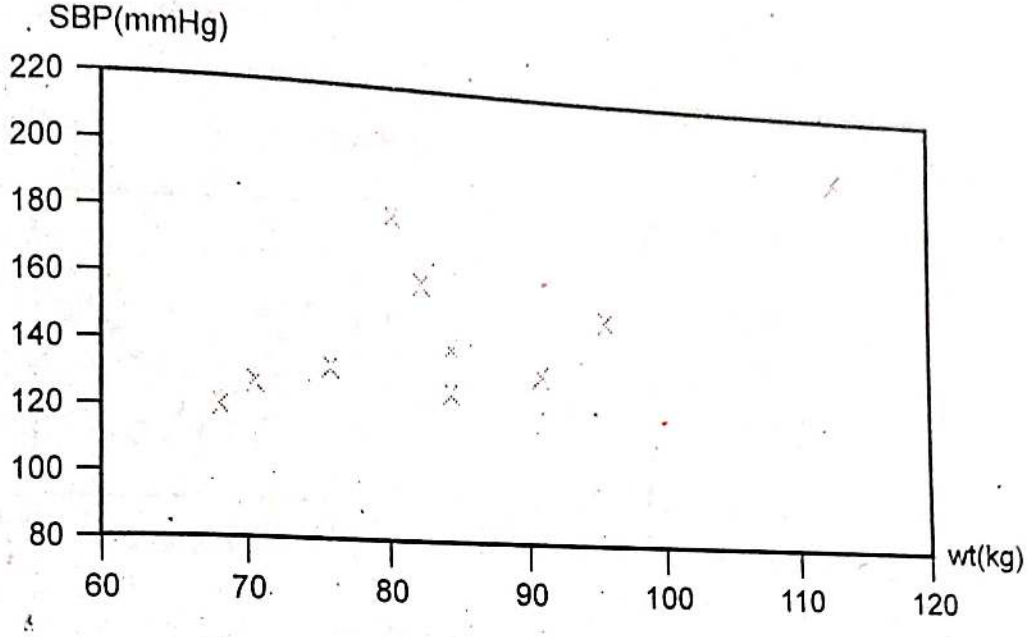


Fig. 2.10.1 (a) Scatter diagram of weight

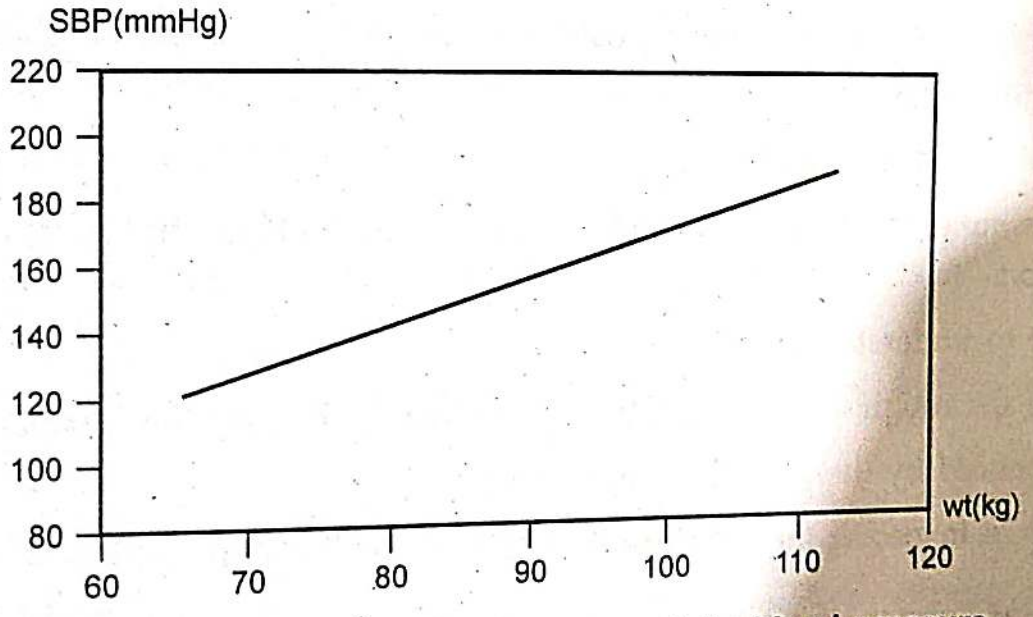


Fig. 2.10.1 (b) Scatter diagram of systolic blood pressure

- The pattern of data is indicative of the type of relationship between your two variables :
  1. Positive relationship
  2. Negative relationship
  3. No relationship.

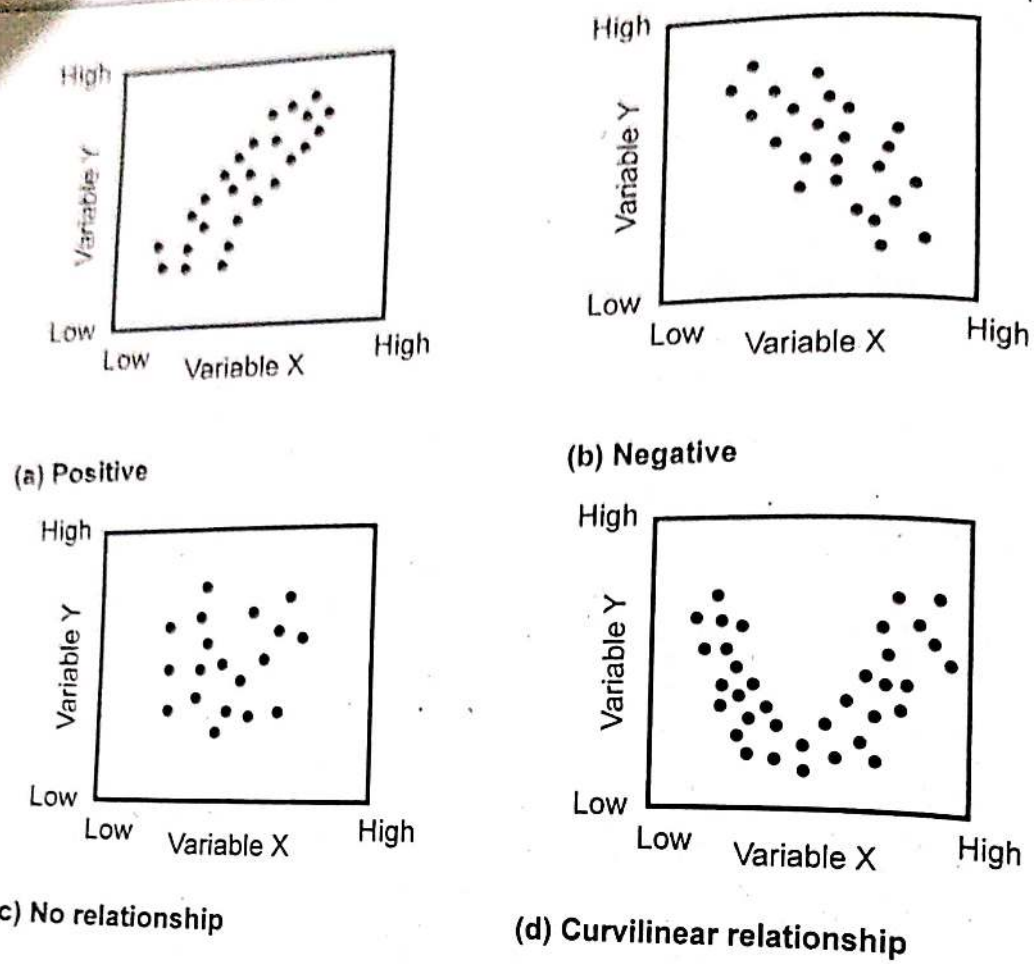


Fig. 2.10.2

- The scattergram can indicate a **positive** relationship, a **negative** relationship, or zero relationship.

Advantages of Scatter Diagram

1. It is a simple to implement and attractive method to find out the nature correlation.
2. It is easy to understand.
3. User will get rough idea about correlation (positive or negative correlation).
4. Not influenced by the size of extreme item
5. First step in investing the relationship between two variables.

Disadvantage of scatter diagram

- Can not adopt the an exact degree of correlation.

2.10.3 Coefficient of Correlation

- **Correlation :** The degree of relationship between the variables under consideration is measure through the correlation analysis.

- The measure of correlation called the **correlation coefficient**. The degree of relationship is expressed by coefficient which range from correlation $(-1 \leq r \geq +1)$. The direction of change is indicated by a sign.
- The correlation analysis enables us to have an idea about the degree and direction of the relationship between the two variables under study.
- Correlation is a statistical tool that helps to measure and analyze the degree of relationship between two variables. Correlation analysis deals with the association between two or more variables.
- Correlation denotes the interdependency among the variables for correlating two phenomenon, it is essential that the two phenomenon should have cause-effect relationship and if such relationship does not exist then the two phenomenon can not be correlated.
- If two variables vary in such a way that movement in one are accompanied by movement in other, these variables are called **cause and effect relationship.**

2.10.4 Properties of Correlation

1. Correlation requires that both variables be quantitative.
2. Positive r indicates positive association between the variables, and negative r indicates negative association.
3. The correlation coefficient (r) is always a number between -1 and $+1$.
4. The correlation coefficient (r) is a pure number without units.
5. The correlation coefficient measures clustering about a line, but only relative to the SD's.
6. The correlation can be misleading in the presence of outliers or nonlinear association.
7. Correlation measures association. But association does not necessarily show causation.

2.10.5 KARL Pearson Correlation Coefficient

- The **product moment correlation**, r, summarizes the strength of association between two metric (interval or ratio scaled) variables, say X and Y. It is an index used to determine whether a linear or straight-line relationship exists between X and Y.
- As it was originally proposed by Karl Pearson, it is also known as the *Pearson correlation coefficient*. It is also referred to as *simple correlation, bivariate correlation,* or merely the *correlation coefficient*.

- The correlation coefficient between two variables will be the same regardless of their underlying units of measurement.
- It measures the nature and strength between two variables of the quantitative type.
- The sign of r denotes the nature of association. While the value of r denotes the strength of association.
- If the sign is positive this means the relation is direct (an increase in one variable is associated with an increase in the other variable and a decrease in one variable is associated with a decrease in the other variable).
- While if the sign is negative this means an inverse or indirect relationship (which means an increase in one variable is associated with a decrease in the other).
- The value of r ranges between (– 1) and (+ 1). The value of r denotes the strength of the association as illustrated by the following diagram.

1. If r = Zero this means no association or correlation between the two variables.
2. If $0 < r < 0.25$ = Weak correlation.
3. If $0.25 \leq r < 0.75$ = Intermediate correlation.
4. If $0.75 \leq r < 1$ = Strong correlation.
5. If r = 1 = Perfect correlation

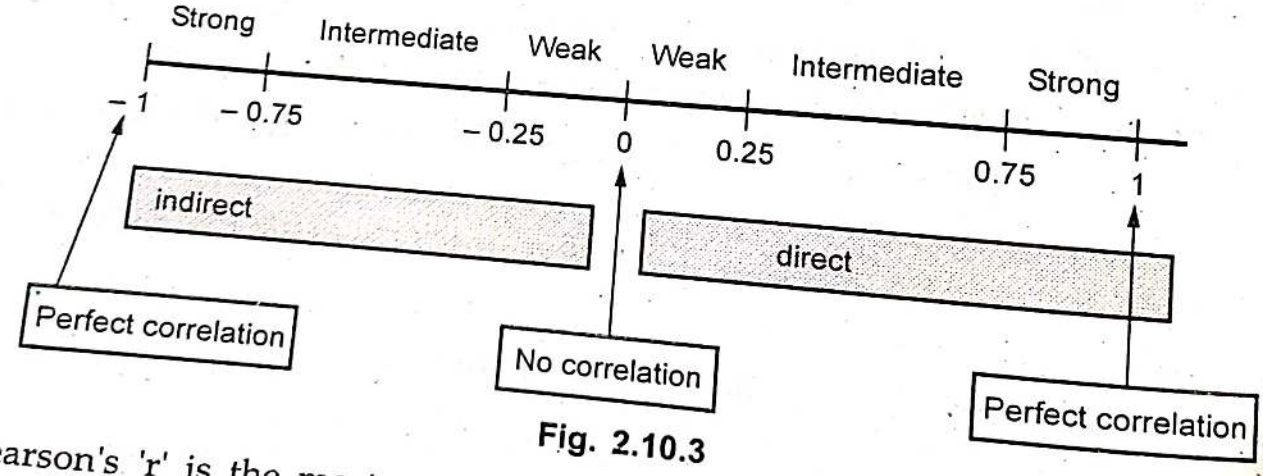


Fig. 2.10.3

- Pearson's 'r' is the most common correlation coefficient. Karl Pearson's Coefficient of Correlation denoted by - 'r' The coefficient of correlation 'r' measure the degree of linear relationship between two variables say x and y.
- Formula for calculating correlation coefficient (r) :

. When deviation taken from actual mean :

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

2. When deviation taken from an assumed mean :

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x^2\right)}{n}\right) \cdot \left(\sum y^2 - \frac{\left(\sum y^2\right)}{n}\right)}}$$

Example 2.10.1 *A sample of 6 children was selected, data about their age in years and weight in kilograms was recorded as shown in the following table. It is required to find the correlation between age and weight.*

Weight (kg)	*Age (years)*
12	*7*
8	*6*
12	*8*
10	*5*
11	*6*
13	*9*

Solution :

X = Variable age is the independent variable

Y = Variable weight is the dependent

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right) \cdot \left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)}}$$

Y = Weight (kg)	X = Age (years)	XY	X^2	Y^2
12	7	84	49	144
8	6	48	36	64
12	8	96	64	144
10	5	50	25	100
11	6	66	36	121
13	9	117	81	169
$\sum y = 66$	$\sum x = 41$	$\sum XY = 461$	$\sum X^2 = 291$	$\sum Y^2 = 742$

$$r = \frac{461 - \frac{41 \times 66}{6}}{\sqrt{\left[291 - \frac{(41)^2}{6}\right]\left[742 - \frac{(66)^2}{6}\right]}}$$

$$r = \frac{461 - 451}{\sqrt{(291 - 280.166)\,(742 - 726)}} = \frac{10}{\sqrt{(10.834)\,(16)}} = \frac{10}{13.166} = 0.7595$$

- Other formula for calculating correlation coefficient is as follows :

Interpreting the correlation coefficient $C_r = \frac{\sum (Z_X Z_Y)}{N}$

- Because the relationship between two sets of data is seldom perfect, the majority of correlation coefficients are fractions (0.92, – 0.80, and the like).
- When interpreting correlation coefficients it is sometimes difficult to determine what is high, low and average.
- The value of correlation coefficient 'r' ranges from – 1 to + 1.
- If r = + 1, then the correlation between the two variables is said to be perfect and positive.
- If r = – 1, then the correlation between the two variables is said to be perfect and negative.
- If r = 0, then there exists no correlation between the variables.

2.10.6 Coefficient of Determination

- The **coefficient of determination** is the amount of variability in one measure that is explained by the other measure.
- The **coefficient of determination** is the square of the correlation coefficient (r^2).
- For example, if the correlation coefficient between two variables is r = 0.90, the coefficient of determination is $(0.90)^2 = 0.81$.

Procedure for computing the correlation coefficient

1. Calculate the mean of the two series 'x' & 'y'.
2. Calculate the deviations 'x' & 'y' in two series from their respective mean.
3. Square each deviation of 'x' & 'y' then obtain the sum of the squared deviation i.e. $\sum x^2$ and $\sum y^2$.
4. Multiply each deviation under x with each deviation under y and obtain the product of 'xy'. Then obtain the sum of the product of x , y i.e. $\sum xy$.
5. Substitute the value in the formula.

Properties of correlation coefficient

1. The correlation coefficient lies between – 1 and + 1. It can be represented as $(-1 \leq r \geq 1)$.
2. The correlation coefficient is independent of the change of origin and scale.
3. The coefficient of correlation is the geometric mean of two regression coefficient.
4. If the values of either variable are converted to a different scale, r will be the same.
5. If the variables x and y are interchanged, r will be the same.
6. The correlation coefficient r will only measure the strength of a *linear* relationship. It says nothing about other kinds of relationships.

Example 2.10.2 *Compute Pearson's coefficient of correlation between maintains cost and sales as per the data given below :*

Maintains cost	*39*	*65*	*62*	*90*	*75*	*78*	*82*	*98*	*25*	*36*
Sales	*58*	*60*	*91*	*84*	*51*	*62*	*53*	*47*	*86*	*68*

Solution : Given data :

n = 10

x = Maintains cost

y = Sales cost

Calculate coefficient of correlation.

x	y	X^2	Y^2	xy
39	58	1521	3364	2262
65	60	4225	3600	3900
62	91	3844	8281	5642
90	84	8100	7056	7560
75	51	5625	2601	3825
78	62	6084	3844	4836
82	53	6724	2809	4346
98	47	9604	2209	4606
25	86	625	7396	2150
36	68	1296	4624	2448
$\sum x = 650$	$\sum y = 660$	$\sum X^2 = 47648$	$\sum Y^2 = 45784$	$\sum xy = 41575$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \cdot \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

$$= \frac{45604 - \frac{(650)(660)}{10}}{\sqrt{47648 - \frac{(650)^2}{10}} \sqrt{45784 - \frac{(660)^2}{10}}}$$

$$= \frac{45604 - 42900}{(73.47)(47.1)}$$

$$r = \frac{2704}{3460.437} = \mathbf{0.7814}$$

Correlation coefficient is positively correlated.

Example 2.10.3 *Two persons taste 10 red italian wines, grading them by an ordinal scale between 1 and 5. The results are as follows :*

Wine No.	Grades person 1	Grades person 2
1	1	2
2	2	3
3	4	5
4	5	4
5	2	2
6	2	2
7	4	3
8	3	4
9	1	3
10	4	2

Calculate Spearman's rank correlation coefficient.

Solution : For calculating Spearman's rank correlation coefficient, we sort the grades for each person. The resulting rank numbers are averaged for tied observations :

Wine No.	Grades Person 1	Rank	Rank Ties
1	1	1	1.5
9	1	2	1.5
6	2	3	4
5	2	4	4
2	2	5	4
8	3	6	6
10	4	7	8
7	4	8	8
3	4	9	8
4	5	10	10

For grades person 2 :

Wine No.	Grades Person 2	Rank	Rank Ties
5	2	1	2.5
1	2	2	2.5
10	2	3	2.5
6	2	4	2.5
2	3	5	6
7	3	6	6
9	3	7	6
8	4	8	8.5
4	4	9	8.5
3	5	10	10

The final table of ranks includes the differences of the ranks as well as the squared differences :

Wine No.	Person 1 (X)	Rank	Person 2 (Y)	Rank	Rank difference (X – Y)	Squared rank difference
1	1	1.5	2	2.5	– 1	1
2	2	4	3	6	– 2	4
3	4	8	5	10	– 2	4
4	5	10	4	8.5	1.5	2.25
5	2	4	2	2.5	1.5	2.25
6	2	4	2	2.5	1.5	2.25
7	4	8	3	6	2	4
8	3	6	4	8.5	– 2.5	6.25
9	1	1.5	3	6	– 4.5	20.25
10	4	8	2	2.5	5.5	30.25
		Sum			0	$\sum D^2 = 76.50$

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 76.50}{10((10)^2 - 1)} = \frac{459}{990} = 0.5364$$

Example 2.10.4 *A sample of 12 fathers and their elder sons gave the following data about their heights in inches. Calculate the coefficient of rank correlation.*

Fathers	65	63	67	64	68	62	70	66	68	67	69	71
Sons	68	66	68	65	69	66	68	65	71	67	68	70

Solution :

Fathers heights (X)	Ranks (x)	Sons heights (Y)	Ranks (y)	Rank difference D = x – y	D^2
65	7	68	4	3	9
63	9	66	6	3	9
67	5	68	4	1	1
64	8	65	7	1	1

68	4	69	3	1	1
62	10	66	6	4	16
70	2	68	4	– 2	4
66	6	65	7	– 1	1
68	4	71	1	3	9
67	5	67	5	0	0
69	3	68	4	– 1	1
71	1	70	2	– 1	1
					$\sum D^2 = 53$

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6 \times 53}{12((12)^2-1)} = 1 - \frac{318}{1716} = 1 - 0.1853 = 0.8147$$

Example 2.10.5 *A random sample of 5 college students is selected and their grades in operating system and software engineering are found to be ?*

Subject	*1*	*2*	*3*	*4*	*5*
Operating system	*85*	*60*	*73*	*40*	*90*
Software engineering	*93*	*75*	*65*	*50*	*80*

Calculate Pearsons rank correlation coefficient ?

Solution :

Operating system (X)	Ranks (x)	Software engineering (Y)	Ranks (y)	Rank difference D = x – y	D^2
85	2	93	1	1	1
60	4	75	3	1	1
73	3	65	4	– 1	1
40	5	50	5	0	0
90	1	80	2	– 1	1
					$\sum D^2 = 4$

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6 \times 4}{5((5)^2-1)} = 1 - 0.2 = 0.8$$

Example 2.10.6 *Find Karl Pearson's correlation coefficient for the following paired data.*

Wages	100	101	102	102	100	99	97	98	96	95
Cost of living	98	99	99	97	95	92	95	94	90	91

Solution : Let

x = *Wages* y = *Cost of living*

Calculate $$\bar{X} = \frac{100+101+102+102+100+99+97+98+96+95}{10} = \frac{990}{10} = 99$$

Calculate $$\bar{Y} = \frac{98+99+99+97+95+92+95+94+90+91}{10} = \frac{950}{10} = 95$$

Wages (x)	$X = x - \bar{X}$	X^2	Cost of living (y)	$Y = y - \bar{Y}$	Y^2	XY
100	1	1	98	3	9	3
101	2	4	99	4	16	8
102	3	9	99	4	16	12
102	3	9	97	2	4	6
100	1	1	95	0	0	0
99	0	0	92	−3	9	0
97	−2	4	95	0	0	0
98	−1	1	94	−1	1	1
96	−3	9	90	−5	25	15
95	−4	16	91	−4	16	16
$\sum x = 990$	$\sum X = 0$	$\sum X^2 = 54$	$\sum y = 950$	$\sum Y = 0$	$\sum Y^2 = 96$	$\sum XY =$

$$r = \frac{\sum XY}{\sqrt{(\sum X^2)(\sum Y^2)}} = \frac{61}{\sqrt{(54)(96)}} = \frac{61}{\sqrt{5184}} = \frac{61}{72} = 0.847$$

Karl Pearson's correlation coefficient r = 0.847

Example 2.10.7 *Calculate coefficient of correlation between age of cars and annual maintenance and comment.*

Age of cars (years)	2	4	6	7	8	10	12
Annual maintenance Cost (Rupees)	1600	1500	1800	1900	1700	2100	2000

Solution : Let

x = Age of cars y = Annual maintenance cost n = 7

Calculate $\overline{X} = \frac{2+4+6+7+8+10+12}{7} = \frac{49}{7} = 7$

Calculate $\overline{Y} = \frac{1600+1500+1800+1900+1700+2100+2000}{7} = \frac{12600}{7} = 1800$

x	$X = x - \overline{X}$	X^2	y	$Y = y - \overline{Y}$	Y^2	XY
2	– 5	25	1600	-200	40000	1000
4	– 3	9	1500	– 300	90000	900
6	– 1	1	1800	0	0	0
7	0	0	1900	100	10000	0
8	1	1	1700	– 100	10000	-100
10	3	9	2100	300	90000	900
12	5	25	2000	200	40000	1000
$\sum x = 49$	$\sum X = 0$	$\sum X^2 = 70$	$\sum y = 12600$	$\sum Y = 0$	$\sum Y^2 = 280000$	$\sum XY = 3700$

$$r = \frac{\sum XY}{\sqrt{(\sum X^2)(\sum Y^2)}} = \frac{3700}{\sqrt{70(280000)}} = \frac{3700}{\sqrt{19600000}} = \frac{3700}{4427.188} = 0.8357$$

Coefficient of correlation r = 0.8357

Example 2.10.8 *Find Karl Pearson's correlation coefficient for the following paired data.*

X	38	45	46	38	35	38	46	32	36	38
Y	28	34	38	34	36	36	28	29	25	36

What inference would you draw from estimate ?

Solution :

X	$x = X - 38$	x^2	Y	$y = Y - 34$	y^2	XY
38	0	0	28	−6	36	0
45	7	49	34	0	0	0
46	8	64	38	4	16	32
38	0	0	34	0	0	0
35	−3	9	36	2	4	−6
38	0	0	26	−8	64	0
46	8	64	28	−6	36	−48
32	−6	36	29	−5	25	30
36	−2	4	25	−9	81	18
38	0	0	36	2	4	0
	$\sum x = 12$	$\sum x^2 = 226$		$\sum y = -26$	$\sum Y^2 = 266$	$\sum XY = 26$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\sum x^2 - ((\sum x)^2)/N}\sqrt{\sum y^2 - ((\sum y)^2)/N}}$$

$$r = \frac{26 - \frac{(12)(-26)}{10}}{\sqrt{226 - ((12)^2)/10)}\sqrt{266 - ((-26)^2)/10)}} = \frac{26 + 31.2}{\sqrt{226 - 14.4}\sqrt{266 - 67.6)}} = \frac{57.2}{14.546 \times 14.085}$$

$r = 0.2792$

Example 2.10.9 *Psychological tests of intelligence and of engineering ability were applied to 10 students. Here is a record of ungrouped data showing intelligence ratio (I.R) and engineering ratio (E.R) calculate the co-efficient of correlation ?*

Student	A	B	C	D	E	F	G	H	I	J
I.R	105	104	102	101	100	99	98	96	93	92
E.R	101	103	100	98	95	96	104	92	97	94

Solution :

$$\text{Mean} \quad \overline{X} = \frac{105 + 104 + 102 + 101 + 100 + 99 + 98 + 96 + 93 + 92}{10} = \frac{990}{10} = 99$$

$$\text{Mean} \quad \overline{Y} = \frac{101 + 103 + 100 + 98 + 95 + 96 + 104 + 92 + 97 + 94}{10} = \frac{980}{10} = 98$$

	Intelligence ratio			Engineering ratio			
	x	$X = x - \bar{X}$	X^2	y	$Y = y - \bar{Y}$	Y^2	XY
A	105	6	36	101	3	9	18
B	104	5	25	103	5	25	25
C	102	3	9	100	2	4	6
D	101	2	4	98	0	0	0
E	100	1	1	95	– 3	9	– 3
F	99	0	0	96	– 2	4	0
G	98	– 1	1	104	6	36	– 6
H	96	– 3	9	92	– 6	36	18
I	93	– 6	36	97	– 1	1	6
J	92	– 7	49	94	– 4	16	28
	$\sum x = 990$	$\sum X = 0$	$\sum X^2 = 170$	$\sum y = 980$	$\sum Y = 0$	$\sum Y^2 = 140$	$\sum XY = 92$

$$r = \frac{\sum XY}{\sqrt{(\sum X^2)(\sum Y^2)}} = \frac{92}{\sqrt{(170)(140)}} = \frac{92}{\sqrt{23800}} = \frac{92}{154.272} = 0.596$$

Example 2.10.10 *Calculate coefficient of correlation from the following data.*

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	3

Solution : In the problem statement, both series items are in small numbers. So there is no need to take deviations.

Computation of coefficient of correlation

X	Y	X^2	Y^2	XY
12	14	144	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	3	49	9	21
$\sum X = 70$	$\sum Y = 63$	$\sum X^2 = 728$	$\sum Y^2 = 651$	$\sum XY = 676$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\sum x^2 - ((\sum x)^2)/N}\sqrt{\sum y^2 - ((\sum y)^2)/N}}$$

$$r = \frac{676 - \frac{(70)(63)}{7}}{\sqrt{728 - ((70)^2)/7}\sqrt{651 - ((63)^2/7)}} = \frac{676 - 630}{\sqrt{728 - 700} - \sqrt{651 - 567}} = \frac{46}{5.29 \times 9.165}$$

r = 0.9488

2.11 Analysis of Variance

- Analysis of Variance (ANOVA) is a statistical formula used to compare variance across the means of different groups. A range of scenarios use it to determine if there is any difference between the means of different groups.
- Fig. 2.11.1 shows ANOVA.
- The outcome of ANOVA is the 'F statistic'. This ratio shows the difference between the within group variance and the between group variance, which ultimately produces a figure which allows a conclusion that the null hypothesis is supported or rejected. If there is a significant difference between the groups, the null hypothesis is not supported, and the F-ratio will be larger.

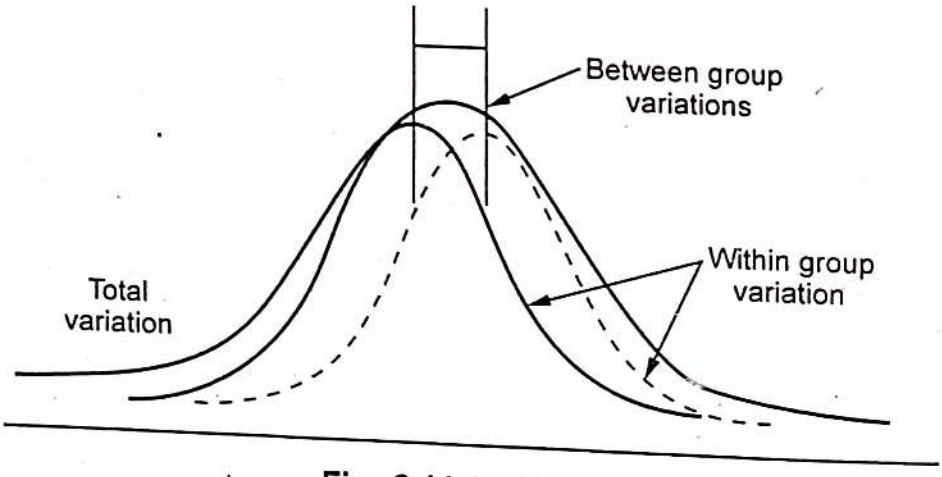


Fig. 2.11.1 ANOVA

- When to use ANOVA :
  1. Data must be experimental.
  2. If you do not have access to statistical software, an ANOVA can be computed by hand.
  3. With many experimental designs, the sample sizes must be equal for the various factor level combinations.
  4. A regression analysis will accomplish the same goal as an ANOVA.
  5. ANOVA formulas change from one experimental design to another.

- One of the biggest challenges in machine learning is the selection of the most reliable and useful features that are used in order to train a model. ANOVA helps in selecting the best features to train a model. ANOVA minimizes the number of input variables to reduce the complexity of the model. ANOVA helps to determine if an independent variable is influencing a target variable.
- An example of ANOVA use in data science is in email spam detection. Because of the massive number of emails and email features, it has become very difficult and resource-intensive to identify and reject all spam emails. ANOVA and f-tests are deployed to identify features that were important to correctly identify which emails were spam and which were not.

2.12 Multiple Choice Questions

Q.1 Two items are chosen at random from 12 items of which 4 are defective. A be the event that 'both items chosen are defective'. What is P(A) ?

a 1/11 b 14/33

c 10/11 d 1/33

Q.2 Two events are not independent if _____.

a events are not mutually exclusive

b events are mutually exclusive

c outcome of one trial does not depend on the outcome of the other trial

d none of these

Q.3 A random variable is said to be discrete if its range set is _____.

a finite b countably infinite

c either (a) or (b) d neither (a) nor (b)

Q.4 A pair of fair dice is tossed. What is the probability that the maximum of the two numbers is greater than 4 ?

a 4/36 b 20/36

c 2/36 d 6/36

Q.5 If X is the random variable representing the number of tails obtained when a coin is tossed four times, the maximum value taken by X is _____.

a 0 b 3

c 4 d 16