MODULE 1

Foundations of AiML

You will learn about an 'Overview of AiML' in this module.

Module Learning Objectives

- At the end of this module, you will be able to:
- Understand the basics of Artificial Intelligence (AI) and its evolution
- Understand the basics of Machine Learning (ML), its origin and how it has evolved over time
- Enumerate the factors that connect and the commonalities that exist among Machine Learning (ML) and Artificial Intelligence (AI)
- Explain the applications of AI and ML in the modern context

Module Topics

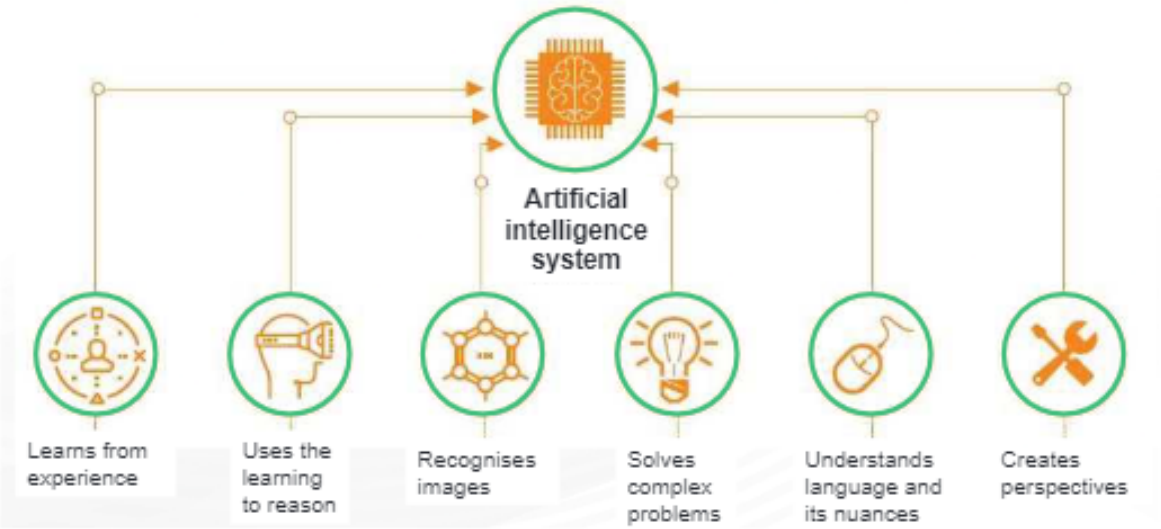
Let us take a quick look at the topics we will cover in this module:

- Introduction and history of AI
- Introduction and history ML
- Overlap between ML and AI
- Applications of AI and ML in the modern context

Copyright © 2018, Xebia Group. All rights reserved.

1. An Introduction to AI

Artificial intelligence is the ability of a computer program or a computer-assisted system to process the information and extract outcomes in a manner similar to the thought process of a human being.

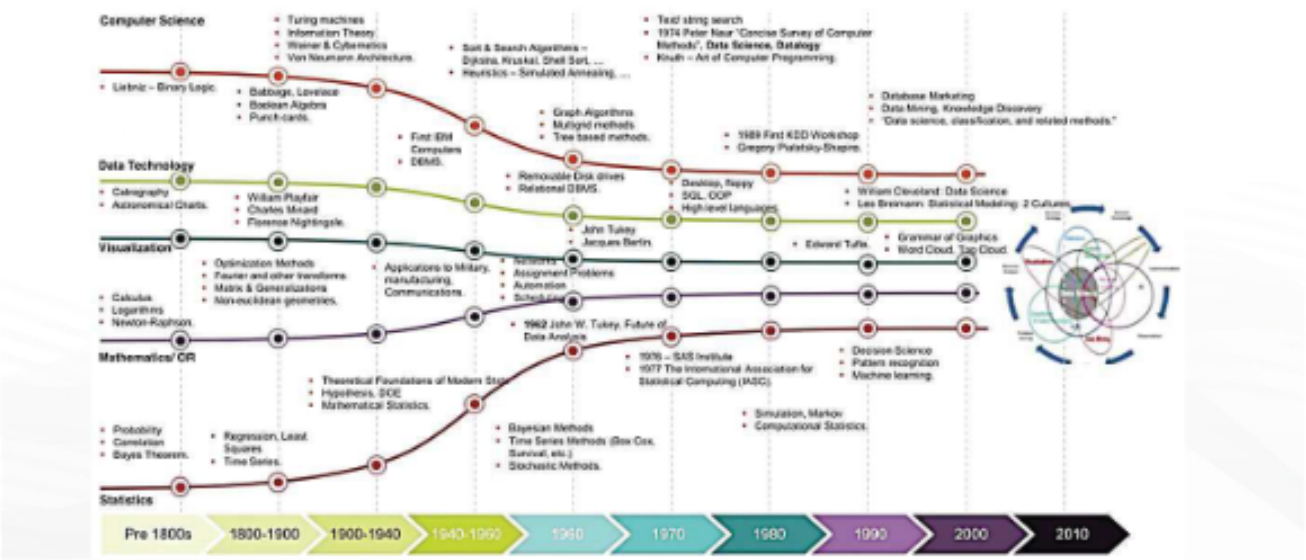


The major goal of AI is to develop the systems capable of handling complex problems in a way similar to human logic and reasoning. Thus, AI refers to the capability of computers or computer-enabled systems like robots to process the information and derive outcomes in a way that resembles human thought process in learning, decision making and solving problems.

John McCarthy (founder of MIT AI lab), who is regarded as the father of AI, defines AI as "the science and engineering of making intelligent machines, especially intelligent computer programs."

A report by Frost & Sullivan estimates that AI has the capacity to automate around 50% of jobs in the Western world within the next two decades.

1.1 History of AI



 Copyright © 2018, Xebia Group. All rights reserved.

Intelligence is said to be a rare gift that humans have among all other creatures on the planet. The ability to reason, solve problems, understand, learn and remember language, to dream and remember skills are all a gift of evolution.

History shows that it is possible to create super intelligence through technological advancements. What would super intelligence be like?

Nick Bostrom defined it as "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills."

Before understanding the ability to create intelligence, scientists discovered the ways in which machines learned. This process of learning is what became the stepping stone to artificial intelligence development. Instead of writing rules and algorithms are created to process information and learn from data. Instead of command and output, a logic is written for the machine to use as a method of processing a given data. This way the machine takes an input and produces an output using the logic or algorithm. A set of input and output associations are given to a machine, the machine learns these associations and produces a code. Once the machine assimilates this code, it can produce outputs based on this logic to other external inputs. This is very similar to how data mining works. Except that data mining seeks to derive knowledge out of existing information whereas machine learning works towards predicting an output or a future data from existing information and associations.

Over the years, this super intelligence has been defined as different aspirational goals by scientists and achieved segmentally.

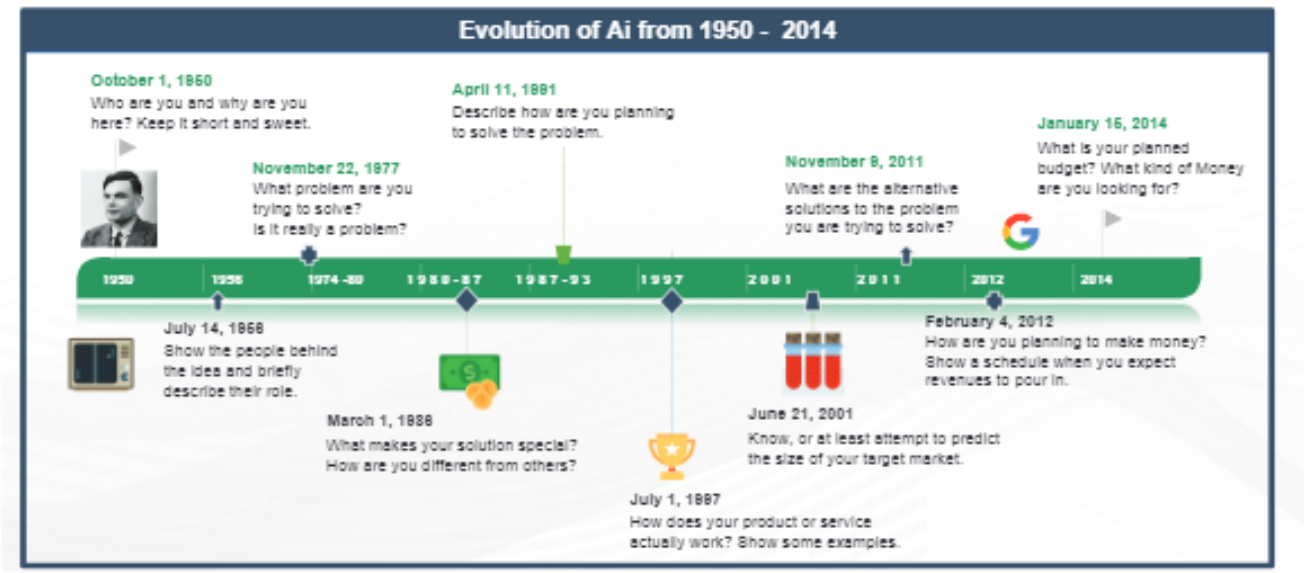
Artificial general Intelligence – This would be a machine that can learn, recall, makes decisions and communicate in the way that humans do. The singularity summit conducted in 2012 predicted that this would be achieved by 2040.

Weak/narrow artificial intelligence- This is designed to solve specific problems. For example, playing chess, or learning a language, translating, etc.

Adapted from Reference- https://hackernoon.com/understanding-understanding-an-intro-to-artificial-intelligence-be76c5ec4d2e

Copyright © 2018, Xebia Group. All rights reserved.

1.2 Timeline of Artificial Intelligence (AI)



1950 - Alan Turing came up with the first Academy paper on artificial intelligence. His major question, " being can machines think?" According to his paper computational logic became the core functioning principle of artificial intelligence. His focus in this paper was majorly on the wariest aspects of machine learning. All approaches were based on the ways in which humans learn. Through rewards and punishment, modeling, deductions based on learned principles also known as computational logic.

The breakthrough in Turing's work was his the first AI test called the imitation game.

Adapted from http://www.lirmm.fr/ecai2012/images/stories/ecai_doc/pdf/Turing/Muggleton_ECAI.pdf

1956- Turing's test stirred up an official meet in Dartmouth College, with eminent experts discussing artificial simulation.

This discussion led to the expansion of the idea and John McCarthy coined the term artificial intelligence for the first time in the paper that was submitted as a result of this rigorous brainstorming on the working of the machine based learning and execution of tasks in a similar manner as humans. The focus now shifted towards the ability of machines or computers to truly understand language and not just use wrote facts to process information. Was there a possibility for computers to go beyond learning and computing? Could computers develop intuitive capabilities after learning languages?

These fundamental questions and discussions led to funded research projects in the field aided by corporates and the government.

1974-1980- The research on AI was not necessarily giving quick results. It led to the AI effect.

The decade saw eager premature promises of a fully functional AI system which did not really materialize. The changes and problems solved by developed machines almost constantly brought in changes to our own understanding of intelligence. As funding sources went down due to the discounting of actual advancements that happened due to these redefined versions of intelligence. The AI ecosystem went on a lowdown and the PC revolution started taking place.

 Copyright © 2018, Xebia Group. All rights reserved.

1980-87 – A wave of specific expert systems arose in the UK and Japan. The Japanese government was attempting to make fifth generation computers that specialized in specific stimulation of skills among human experts in a variety of domains. This led to better functioning personalized computers.

1987- 93 – AI came back to the top with an MIT COG project that was a humanoid robot. Using Dynamic analysis and Preplanning Tool the robot could now play chess. This restored faith in the ability of AI and the scope of development. Computer scientist Vernor Vinge talked about the " coming technological Singularity" in his paper.

1997- This was the year when evidences of these specific narrow intelligences were employed in different machines and displayed as samples of working intelligent machines. IBM developed a super computer called DeepBlue that defeated world champion Garry Kasparov in a chess battle. This marked a turning point in the history of AI where the machine proved that it could think more strategically than a human brain.

After this turning point the AI started spreading across through different fields and machines.

Over the years AI has made its way into the process like predicting market trends through big data, targeting the specific sample size, making money using specific targeting and more. The data became the key to trigger and equip AI to produce results based on commands and requirements.

Adapted from - https://www.kdnuggets.com/2017/04/brief-history-artificial-intelligence.html

https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_overview.htm

2. Introduction to Machine Learning

Machine Learning

The science of making machines act without being explicitly programmed.

An application of AI that provides machines, the ability to learn automatically and improve from experiences.

Give machines, the access to data and let them learn for themselves.

If you want to make machines act, you need to teach them. You need to make them learn, by providing them with repeating patterns. For example, if you want the system to predict rainfall, you need to supply it at least with a decade's rainfall data, if not more. Using this data, machine will understand the pattern and amount of rainfall over the past years and will predict the rainfall for the current or upcoming years. So, you have to train the ML model with large volumes of similar data.

Arthur Samuel, who is credited with the coining of the term ML, defines ML as "the field of study that gives computers, the ability to learn without being explicitly programmed."

Copyright © 2018, Xebia Group. All rights reserved.

Tom Mitchell in 1997 provided a more engineering-oriented definition for ML - "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

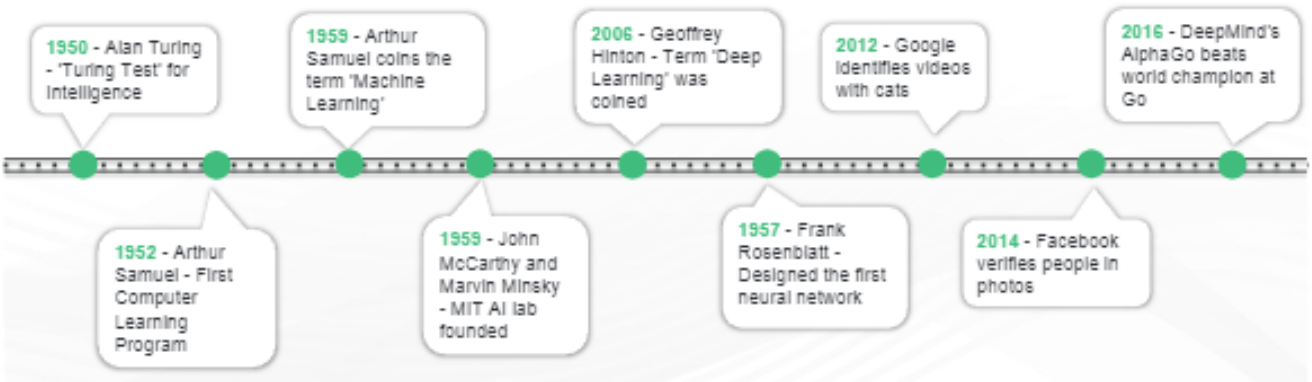
For example, you might have spam filters in your email. A spam filter is the ML program that can learn to flag spam given examples of spam emails (e.g., flagged by users) and examples of regular (non spam, also called "ham") emails. The examples that the system uses to learn are called the ***training set***. Each training example is called a ***training instance*** (or *sample*). In this case, the task T is to flag spam for new emails, the experience E is the ***training data***, and the performance measure P needs to be defined; for example, you can use the ratio of correctly classified emails. This particular performance measure is called ***accuracy*** and it is often used in classification tasks.

If you take AI as a tool, ML is the way to build the tool. ML provides machines, the ability to learn and improve from experiences without the need of being explicitly programmed.

Some of the examples for ML include web search engines, self-driving cars, speech and handwriting recognition, image recognition, recommendation engines, natural language processing, fraud detection, human genome mapping, etc. Learning algorithms are now surpassing human performance in lots of areas these days.

2.1 History of ML

The sequence of events that happened during the evolution of ML is given below:



Way back in 1950, an English computer scientist and mathematician named Alan Turing developed the Turing Test, that evaluates a machine's ability to showcase an intelligent behavior that is equivalent to or comparable with that of a human being. This is a notable incident in the history of Machine learning, as it proved machines' ability to mimic human thinking and intelligence.

Arthur Lee Samuel was an American pioneer in the field of computer gaming and AI, who coined the term "Machine Learning" in 1959. The Samuel Checkers-playing Program was among the world's first successful self-learning programs, and as such a very early demonstration of the fundamental concept of artificial intelligence (AI).

ML started flourishing as an individual discipline in the 1990s. ML and data mining often employ the same methods and overlap significantly, but while ML focuses on prediction, based on known

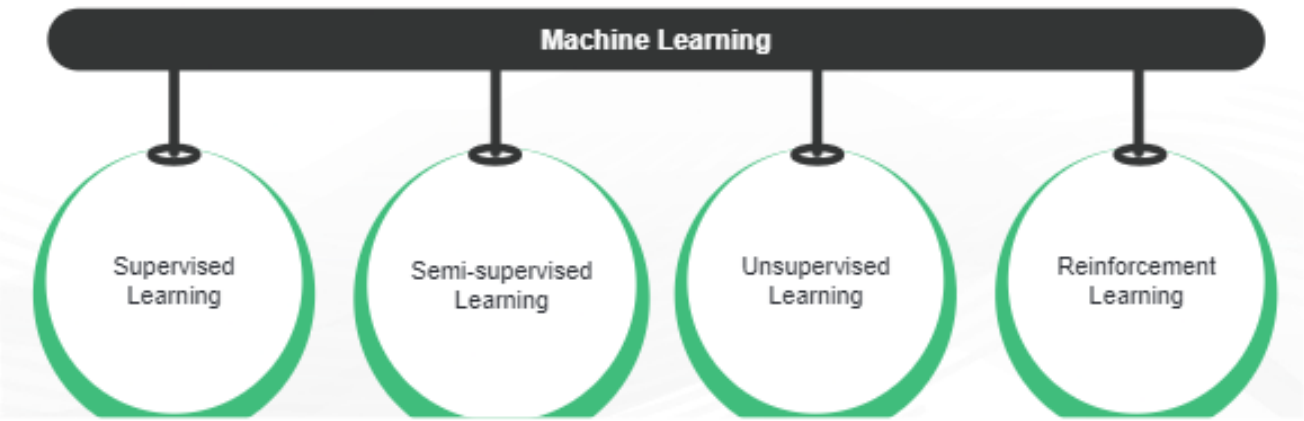
 Copyright © 2018, Xebia Group. All rights reserved.

properties learned from the training data, data mining focuses on the discovery of (previously) unknown properties in data.

Machine learning has been around for decades in some specialized applications, such as Optical Character Recognition (OCR). But the first ML application that really became mainstream, improving the lives of hundreds of millions of people, took over the world back in the 1990s: it was the spam filter.

2.2 Categories of ML Systems

There are the following four major categories of ML systems:



The primary goal of ML is to make computers learn and understand automatically with little or no human intervention and adjust the actions according to changing situations. There are several ways to teach machines. Most commonly used machine learning algorithms are:

- **Supervised Learning** - apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly. A typical supervised learning task is classification. The spam filter is a good example of this: it is trained with many example emails along with their class (spam or ham), and it must learn how to classify new emails.

- **Unsupervised learning** - used when the information used to train is neither classified nor labelled. The system tries to learn without a teacher. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

- For example, say you have a lot of data about your blog's visitors. You may want to run a clustering algorithm to try to detect groups of similar visitors. At no point, do you tell the algorithm which group a visitor belongs to, it finds those connections without your help. For example, it might notice that 40% of your visitors are males who love comic books and generally read your blog in the evening, while 20% are young sci-fi lovers who visit during the weekends, and so on.

Copyright © 2018, Xebia Group. All rights reserved.

- **Semi-supervised Learning** - Some algorithms can deal with partially labelled training data, usually a lot of unlabelled data and a little bit of labelled data. This is called semi-supervised learning. Some photo-hosting services, such as Google Photos, are good examples of this. Once you upload all your family photos to the service, it automatically recognizes that the same person A shows up in photos 1, 5, and 11, while another person B shows up in photos 2, 5, and 7. This is the unsupervised part of the algorithm (clustering). Now all the system needs are for you to tell it who these people are. Just one label per person, and it is able to name everyone in every photo, which is useful for searching photos.

- **Reinforcement learning** - a learning method that interacts with its environment by producing actions and discovers errors or rewards. The learning system, called an *agent* in this context, can observe the environment, select and perform actions, and get *rewards* in return (or penalties in the form of negative rewards). It must then learn by itself what is the best strategy, called a *policy*, to get the most reward over time. A policy defines what action the agent should choose when it is in a given situation.

Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal. For example, many robots implement Reinforcement Learning algorithms to learn how to walk. DeepMind's AlphaGo program is also a good example of Reinforcement Learning: it made the headlines in March 2016 when it beat the world champion Lee Sedol at the game of Go. It learned its winning policy by analyzing millions of game and then playing many games against itself. Note that learning was turned off during the games against the champion; AlphaGo was just applying the policy it had learned.

What did You Grasp?

1. Which of the following options is an example for supervised learning?

   A) Grouping of blog visitors
   B) Robotics
   C) Spam filters
   D) Photo sorting

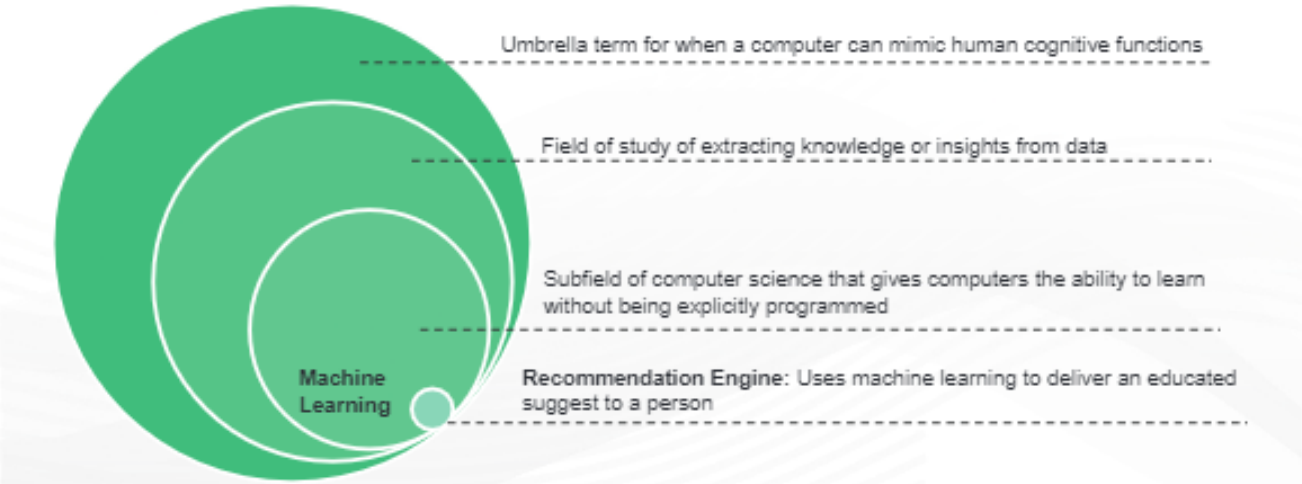
2. Who among the following, coined the term Deep Learning?

   A) John McCarthy
   B) Geoffrey Hinton
   C) Frank Rosenblatt
   D) Alan Turing

 Copyright © 2018, Xebia Group. All rights reserved.

3. How do DS, ML and AI Overlap with Each Other?

The following diagram illustrates how these terms overlap each other.



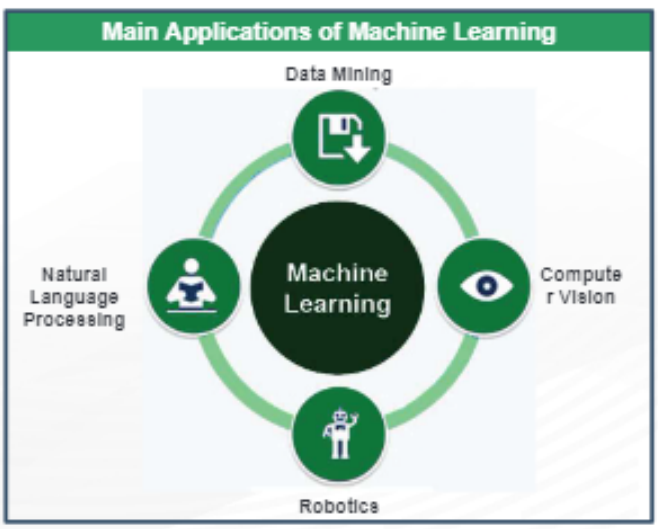
There is so much of hype these days around DS, ML and AI. Let's see, how these terms overlap with each other.

We learnt in an earlier section that DS encompasses a number of disciplines like mathematics, statistics, computer science, information science. Using ML is the practical application of AI. ML is the technology or enabler for AI. DS is likely, how you apply ML, i.e. the real-world application of ML, the goal being the creation of smart products that enrich people's lives.

AI is the broader concept of machines being able to carry out tasks in a way that we would consider "smart". ML is a current application of AI based on the idea that we should just be able to give machines access to data and let them learn for themselves.

4. Applications of ML

The following diagram illustrates the applications of ML.



Source: Frost & Sullivan

Copyright © 2018, Xebia Group. All rights reserved.

As you can see in the picture, there are four major applications of ML:

- Data Mining
- Computer Vision
- Robotics
- Natural Language Processing

Various sectors in which ML finds its applications:

- Agriculture - Farm mechanization, Precision agriculture
- Healthcare - Diagnostic APIs, Personalized treatment
- Manufacturing - Edge computing
- Automotive - Self-driving cars, Infotainment systems
- Social Media - Intelligent filters, Facial identification
- Financial Services - Innovative payment modes, Fraud detection

ML is the best choice for:

- The problems for which existing solutions require a lot of hand-tuning or long lists of rules: One ML algorithm can often simplify code and perform better.
- Complex problems for which there is no good solution at all using a traditional approach: The best ML techniques can find a solution.
- Fluctuating environments: a Machine Learning system can adapt to new data.
- Getting insights into complex problems and large amounts of data.

What did You Grasp?

1. State True or False.

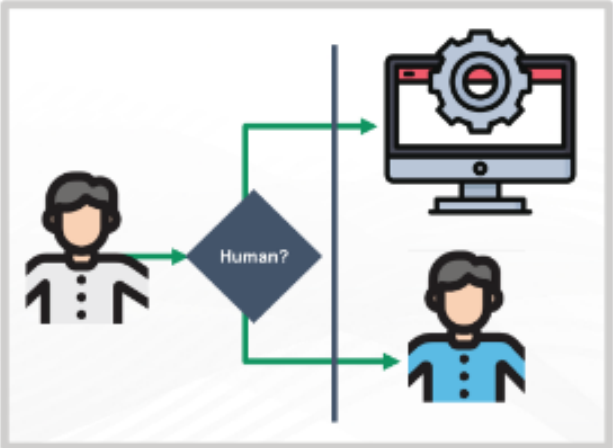
   Artificial Intelligence is the application of machine learning for creating smart systems.

   A) True
   B) False

 Copyright © 2018, Xebia Group. All rights reserved.

5. Artificial Intelligence with Turing Test Problem

- First test ever to check intelligence of a Machine. Determine whether a computer has achieved the intelligence of humans?
- Alan Turing (Father of Modern Computer)
- An English Mathematician wrote a landmark paper in 1950 that asked about the questions. Can machines think?
- He proposed a test to answer the question: " How will we know when we've succeeded?"
- Its Human vs the Machines!
- Can you guess who is who?



Introduction to Turing Test (1950)

So basically this test was about checking the level of intelligence machine has achieved till now. Is it equivalent to what humans have today? Can we distinguish between machine and humans without seeing who is on the other side? The test was based on a natural language conversation with a human and a machine. The evaluator is aware that one of them is surely a machine. All three people in this scenario are totally separated from each other. However, the so-called **total Turing Test** includes a video signal so that the interrogator can test the subject's perceptual abilities, as well as the opportunity for the interrogator to pass physical objects "through the hatch." To pass the total Turing Test, the computer will need

- **computer vision** to perceive objects, and
- **robotics** to manipulate objects and move about.

The machine will pass if the evaluator is unable to distinguish between both the participants.

6. Difference between Thinking Humanly and Thinking Rationally

THINKING HUMANLY	THINKING RATIONALLY
"The exciting new effort to make computers think...machines with minds, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning..." (Bellman, 1978)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
ACTING HUMANLY	**ACTING RATIONALLY**
"The art of creating machines that per-form functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole et al., 1998) "AI...is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Copyright © 2018, Xebia Group. All rights reserved.

The definitions on top are concerned with *thought processes* and *reasoning*, whereas the ones on the bottom address *behaviour*. The definitions on the left measure success in terms of fidelity to *human* performance, whereas the ones on the right measure against an *ideal* performance measure, called **rationality**. A system is rational if it does the "right thing," given what it knows.

Deep dive into the above table.

1) Acting Humanly - The computer would need to possess the following capabilities:

- **Natural language processing** to enable it to communicate successfully in English.
- **Knowledge representation** to store what it knows or hears.
- **Automated reasoning** to use the stored information to answer questions and to draw new conclusions.
- **Machine learning** to adapt to new circumstances and to detect and extrapolate patterns.

2) Thinking Humanly – If we say a program can think like humans, then firstly we must find out how humans think. Three ways to find that out :

- Through introspection—trying to catch our own thoughts as they go by.
- Through psychological experiments—observing a person in action.
- Through brain imaging—observing the brain in action.

If the program's input–output behaviour matches the corresponding human behaviour, that is evidence that some of the program's mechanisms could also be operating in humans.

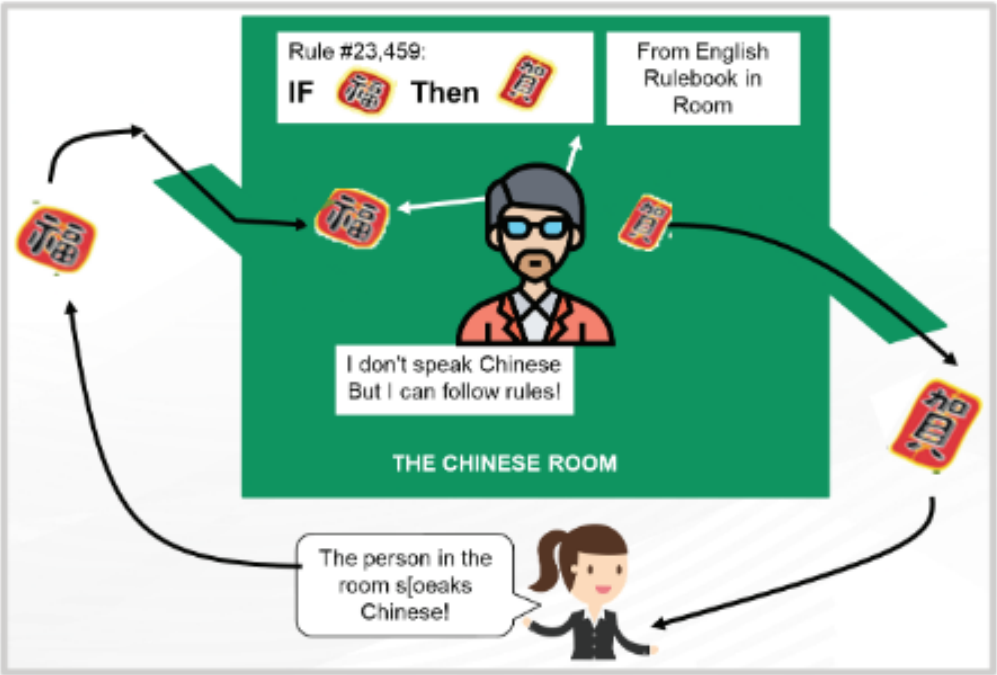
3) Thinking Rationally: (Right Thinking) – A Greek philosopher Aristotle can up with the first code of right thinking. His **syllogisms** provided patterns for argument structures that always yielded correct conclusions when given correct premises—for example, "John is a man; all men are mortal; therefore, John is mortal." These laws of thought were supposed to govern the operation of the mind; their study initiated the field called **logic**. In 19th Century people developed a precise notation for all kinds of location statements like and, or and negation. But was the computer program able to understand and decode it? There are two main obstacles to this approach. First, it is not easy to take informal knowledge and state it in the formal terms required by logical notation, particularly when the knowledge is less than 100% certain. Second, there is a big difference between solving a problem "in principle" and solving it in practice. Even problems with just a few hundred facts can exhaust the computational resources of any computer unless it has some guidance as to which reasoning steps to try first. Although both of these obstacles apply to *any* attempt to build computational reasoning systems, they appeared first in the logicist tradition.

4) Act rationally – An agent (Machine) is asked to do something more than just a simple computer program would do like operate autonomously, perceive the environment, adapt to changes or try to achieve certain goals. A **rational agent** is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome. It works on laws of thoughts which means that the agent should be able to make correct inferences because along with acting rationally the agent should also be able to reason logically to the conclusion that a given action will achieve one's goal

 Copyright © 2018, Xebia Group. All rights reserved.

and then act on it with its conclusion. On the other hand, correct inference is not *all* of rationality, in some situations, there is no provably correct thing to do, but something must still be done. There are also ways of acting rationally that cannot be said to involve inference. For example, recoiling from a hot stove is a reflex action that is usually more successful than aslower action taken after careful deliberation.

7. Chinese Room Problem



Let's discuss some of the key details of the Chinese room problem:

- We just learnt about the Turing test. Now John Searle in 1980 said that the Turing Test is not the right way to judge Intelligence of the machine.
- The **Chinese room argument** holds that a digital computer executing a program cannot have a "mind", "understanding" or "consciousness regardless of how intelligently or human-like the program may make the computer behave.
- According to him machines which don't have any understanding about the input and they are just processing the output to the program cannot be considered as intelligent.
- The Chinese Room name comes from the scenario in which he performed the test in order to put his point forward and contradict the Turing test.
- Simple Idea behind the test was to act as an translator from Chinese to English without understanding the context behind the communication.

Copyright © 2018, Xebia Group. All rights reserved.

- A machine programmed with the right inputs and right outputs having a mind of exactly the same sense human minds have is called a Strong AI.
- A machine that is only processing input and giving a corresponding output without getting the exact sense or understanding is called a weak AI.

John imagines himself (instead of machine) as non-Chinese person sitting inside the room isolated from another Chinese person who is outside the room tries to communicate. He has provided a list of Chinese characters and an instruction book explaining in detail the rules according to which strings (sequences) of characters may be formed, but without giving the meaning of the characters. That means he has a book with an English version of the computer program, along with sufficient paper, pencils, erasers, and filing cabinets.

In this thought experiment, a person in the "Chinese room" is passed questions from outside the room, and consults a library of books to formulate an answer.

Now, he receives all the messages posted through a slot in the door written in Chinese language. He will process all the symbols according to program instructions and produces the Chinese characters as output like:

- If he finds a Chinese symbol like , he returns symbol
- If he finds a Chinese symbol like ДЖЯ, he returns symbol Лиф

8. Agents

An **agent** is anything that can be viewed as perceiving its **environment** through **sensors** and acting upon that environment through **actuators**.

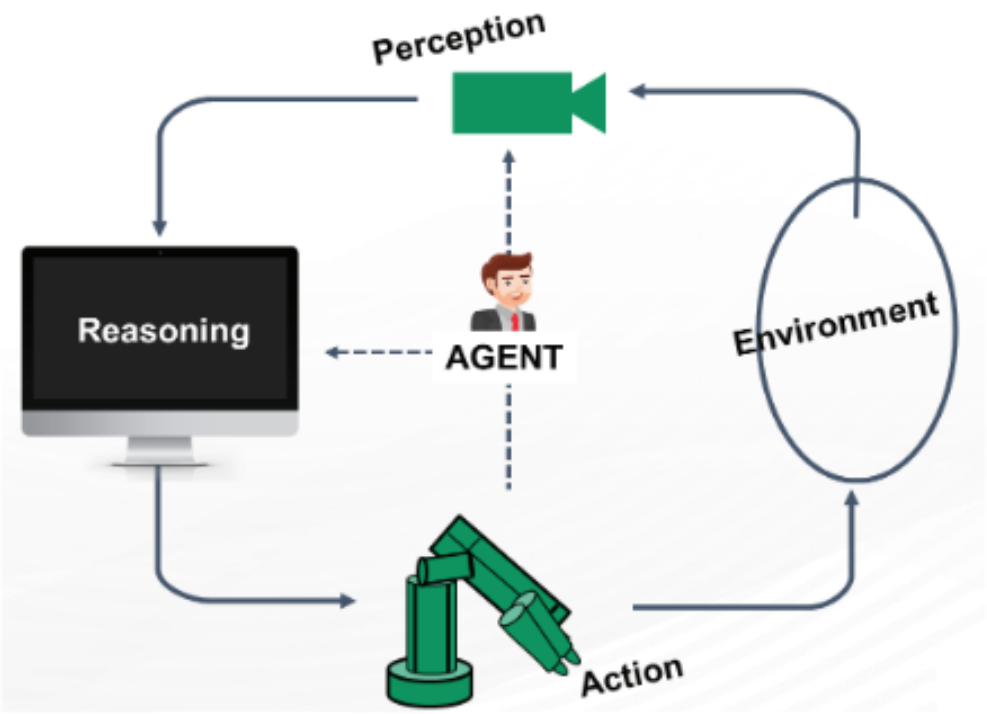
- A human agent has eyes, ears, and other organs for sensors and hands, legs, vocal tract, and so on for actuators. A robotic agent might have cameras and infrared range finders for sensors and various motors for actuators. A software agent receives keystrokes, file contents, and network packets as sensory inputs and acts on the environment by displaying on the screen, writing files, and sending network packets.
- **Percept** to refer to the agent's perceptual inputs at any given instant.
- An agent's behaviour is described by the **agent function** that maps any given percept sequence to an action.
- The agent function for an artificial agent will be implemented by an **agent program.**
- The agent function is an abstract mathematical description; the agent program is a concrete implementation, running within some physical system.

Adapted from: https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_agents_and_environments

Go through the points mentioned on the slide.

 Copyright © 2018, Xebia Group. All rights reserved.

8.1 Rational Agent



A **rational agent** is one that does the right thing—conceptually speaking.

Obviously, doing the right thing is better than doing the wrong thing, but what does it mean to do the right thing?

Considering the *consequences* of the agent's behavior. When an agent is plunked down in an environment, it generates a sequence of actions according to the percepts it receives. This sequence of actions, causes the environment to go through a sequence of states. If the sequence is desirable, then the agent has performed well. This notion of desirability is captured by a **performance measure** that evaluates any given sequence of environment states.

As a general rule, it is better to design performance measures according to what one actually wants in the environment, rather than according to how one thinks the agent should behave.

Notice that we said *environment* states, not *agent* states. If we define success in terms of agent's opinion of its own performance, an agent could achieve perfect rationality simply by deluding itself that its performance was perfect.

Copyright © 2018, Xebia Group. All rights reserved.

8.2 Task Environment

PEAS (Performance, Environment, Actuators, Sensors)	
Performance measure	How will you judge the performance of the agent
Environment	Place around where the agent is put in to perform
Actuators	Hardware / Device through which agent will interact with the environment
Sensors	Hardware / Devices through which agent will observe or take input from the environment

Agent Type	Performance Measure	Environment	Actuators	Sensors
Taxi driver	Safe, fast, legal, comfortable trip, maximize profits	Roads, other traffic, pedestrains, customers	Steering, accelerator, brake, signal, horn, display	Cameras, sonat, speedometer, GPS, odometer, accelerometer, engine sensors, keyboard

The facilitator will explain you about the task environment.

Activity

Internet Shopping Agent

- Performance measure- price, appropriateness, efficiency
- Environment- current and future websites, vendors, shippers

 Copyright © 2018, Xebia Group. All rights reserved.

- Actuators- display to user, follow URL, fill in the form
- Sensors- software for processing webpages

Interactive English Tutor

- Performance measure: Maximize student's score on the test
- Environment: Set of students
- Actuators: Screen display (exercises, suggestions, corrections)
- Sensors: Keyboard

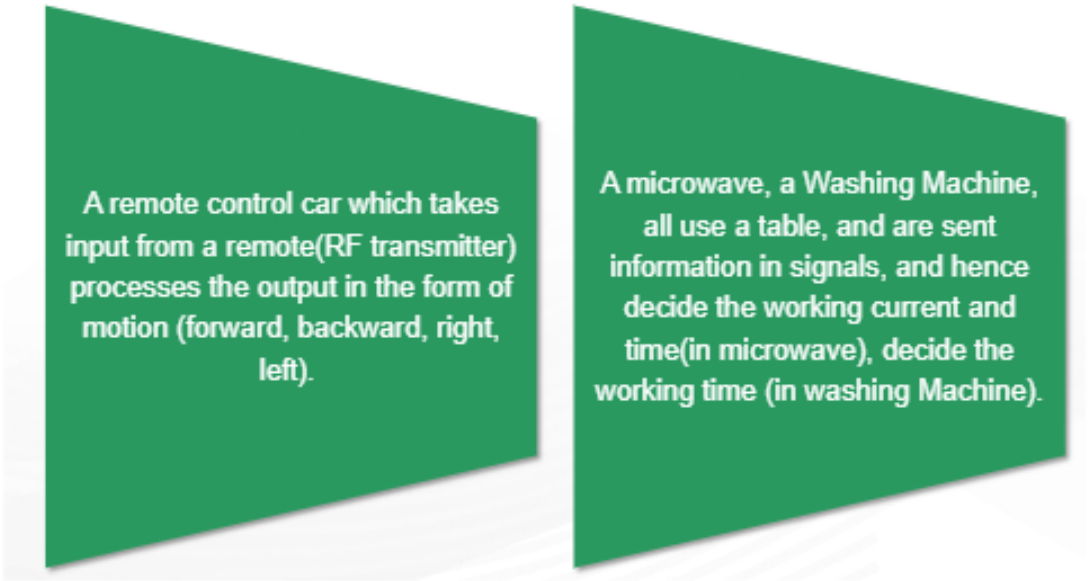
8.3 Types of Agents

Following are the different types of agents:

1. Table Driven Agents
2. Simple Reflex Agents
3. Model Based Agents
4. Goal Based Agents
5. Utility Based Agents
6. Learning Agents

Copyright © 2018, Xebia Group. All rights reserved.

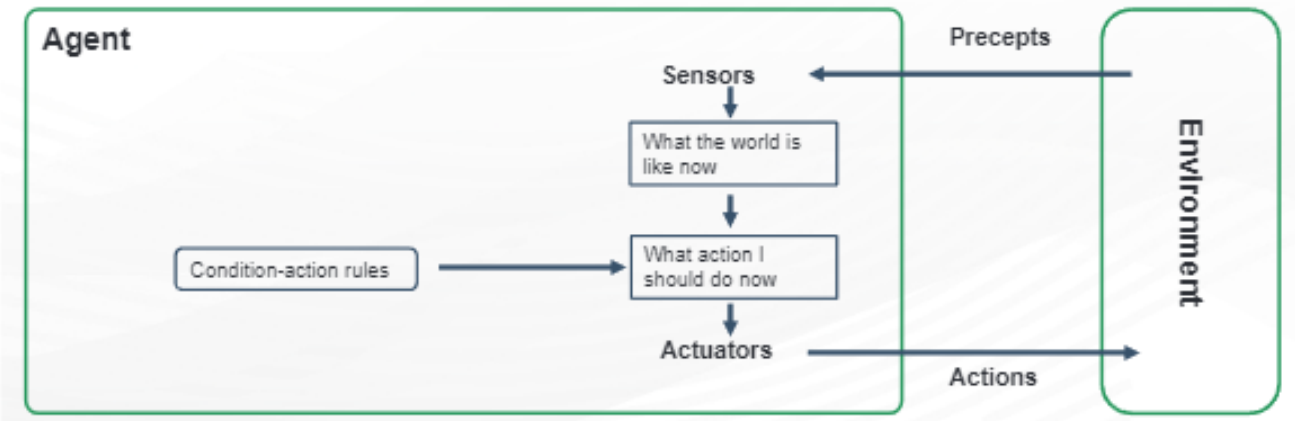
8.3.1 Table Based Agents



In a simple **Table**-driven **Agent**, a lookup **table** is used to match every possible percept sequence to the corresponding action.

8.3.2 Simple Reflex Agents

- A proportional controller which produces proportional output
- Input = x, Output = p * x, here p is the proportionality constant
- An A/C, keep the room temp in a close range of the set temperature (in A/C)
- If the temp goes above the set point the Compressor is turned off, if the temp goes below it, Compressor is turned on.



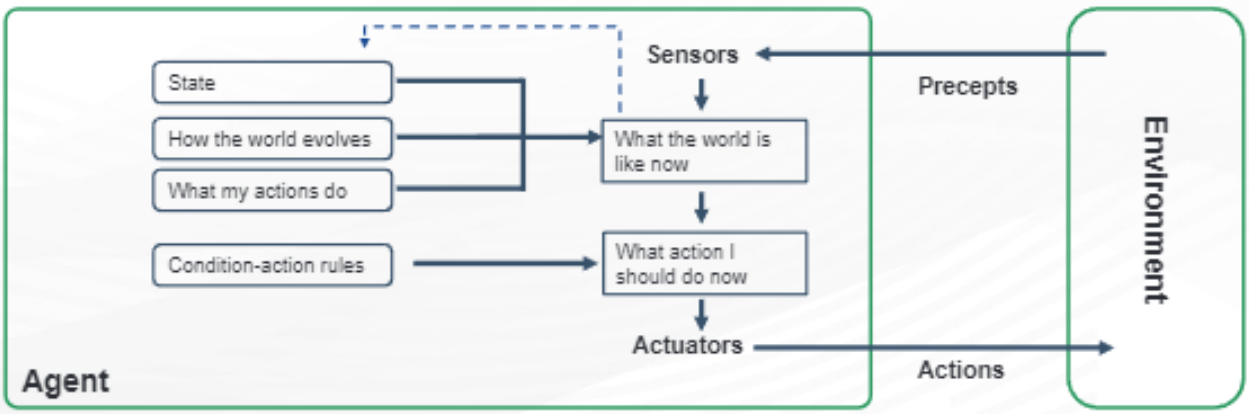
Imagine yourself as the driver of the automated taxi. If the car in front brakes and its brake lights come on, then you should notice this and initiate braking. In other words, some processing is done on the

 Copyright © 2018, Xebia Group. All rights reserved.

visual input to establish the condition we call "The car in front is braking." Then, this triggers some established connection in the agent program to the action "initiate braking." We call such a connection a **condition–action rule**, written as "if car-in-front-is-braking then initiate-braking."

8.3.3 Model Based Agent

- A Prediction system that uses a part of the given data. when predicting Share value of a certain company an intelligent software doesn't observe the entire data.. it takes a look at the partial information for the last week or last 2 months, and then decides if the market is bullish or bearish.
- A data entry system, or an attendance system this system doesn't look if the person returning the book or making the attendance is actually the person. So, theoretically we can cheat the machine by producing a fake fingerprint, and get attendance. It looks at the partial data.



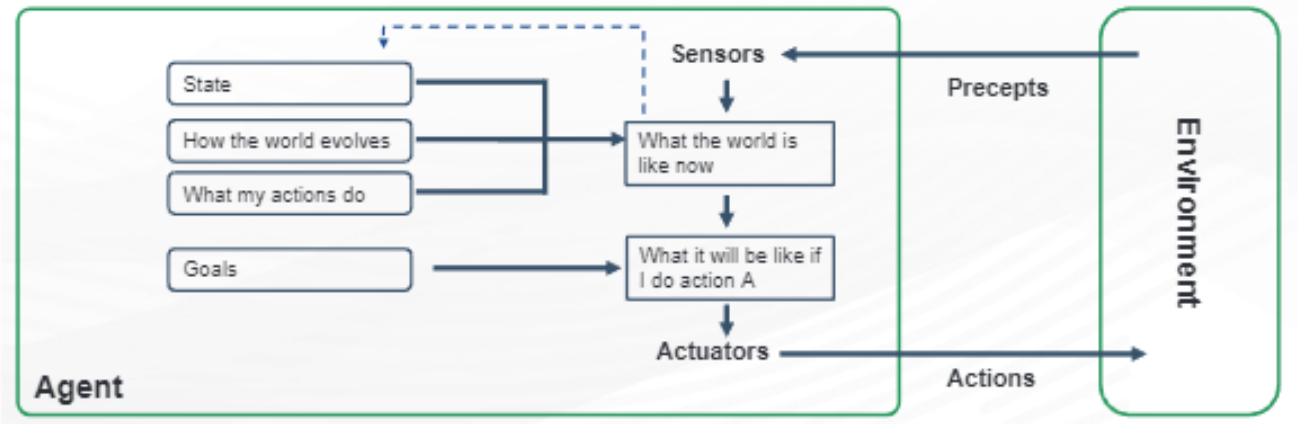
The most effective way to handle partial observability is for the agent to *keep track of the part of the world it can't see now*. That is, the agent should maintain some sort of **internal state** that depends on the percept history and thereby reflects at least some of the unobserved aspects of the current state. For other driving tasks such as changing lanes, the agent needs to keep track of where the other cars are if it can't see them all at once. And for any driving to be possible at all, the agent needs to keep track of where its keys are. Updating this internal state information as time goes by requiring two kinds of knowledge to be encoded in the agent program.

First, we need some information about how the world evolves independently of the agent—for example, that an overtaking car generally will be closer behind than it was a moment ago. Second, we need some information about how the agent's own actions affect the world—for example, that when the agent turns the steering wheel clockwise, the car turns to the right, or that after driving for five minutes northbound on the freeway, one is usually about five miles north of where one was five minutes ago. This knowledge about "how the world works"—whether implemented in simple Boolean circuits or in complete scientific theories—is called a **model** of the world. An agent that uses such a model is called a **model-based agent.**

Copyright © 2018, Xebia Group. All rights reserved.

8.3.4 Goal Based Agents

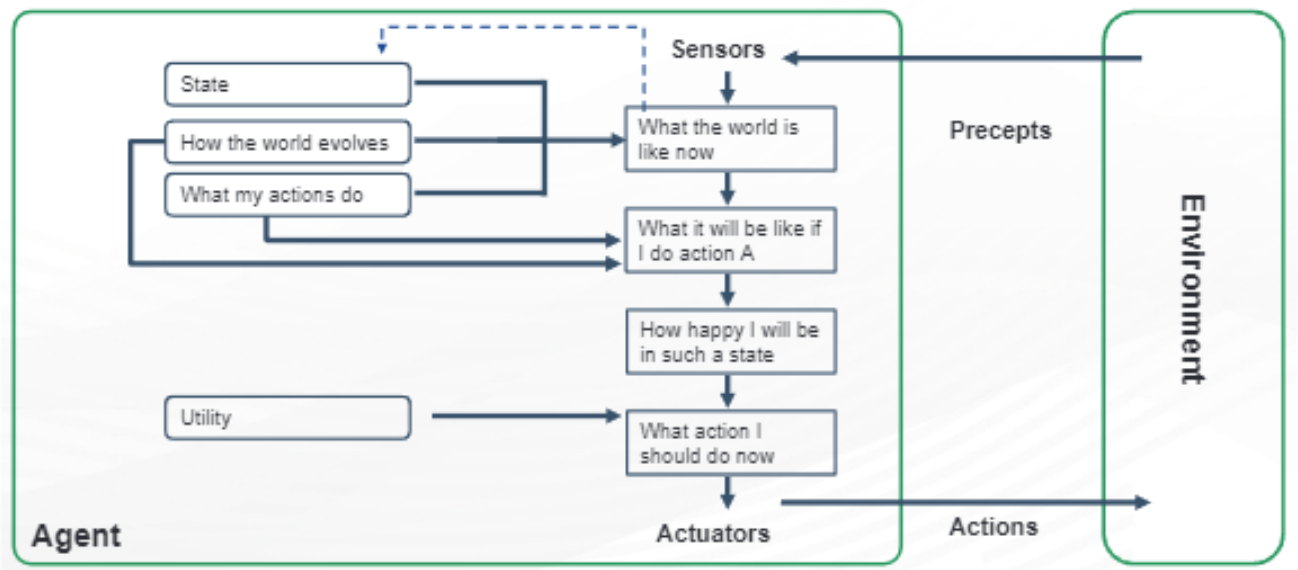
- The goal is fixed, but it can take any path to reach the goal, the goal here helps the system understand which paths are required and which can be ignored, Agent knows which actions it takes are bringing it closer to the goal and which are not needed.
- This is used in games where a fixed goal is to be achieved.
- A Simple Road tracking Agent, uses a fixed goal and suggests optimal ways of reaching said goal.



Knowing something about the current state of the environment is not always enough to decide what to do. For example, at a road junction, the taxi can turn left, turn right, or go straight on. The correct decision depends on where the taxi is trying to get to. In other words, as well as a current state description, the agent needs some sort of **goal** information that describes situations that are desirable. Here in this case that goal can be a destination where the passenger is heading.

8.3.5 Utility Based Agents

- The utility function is a function that maps a state to a measure of the utility of the state.
- An autonomous car.



 Copyright © 2018, Xebia Group. All rights reserved.

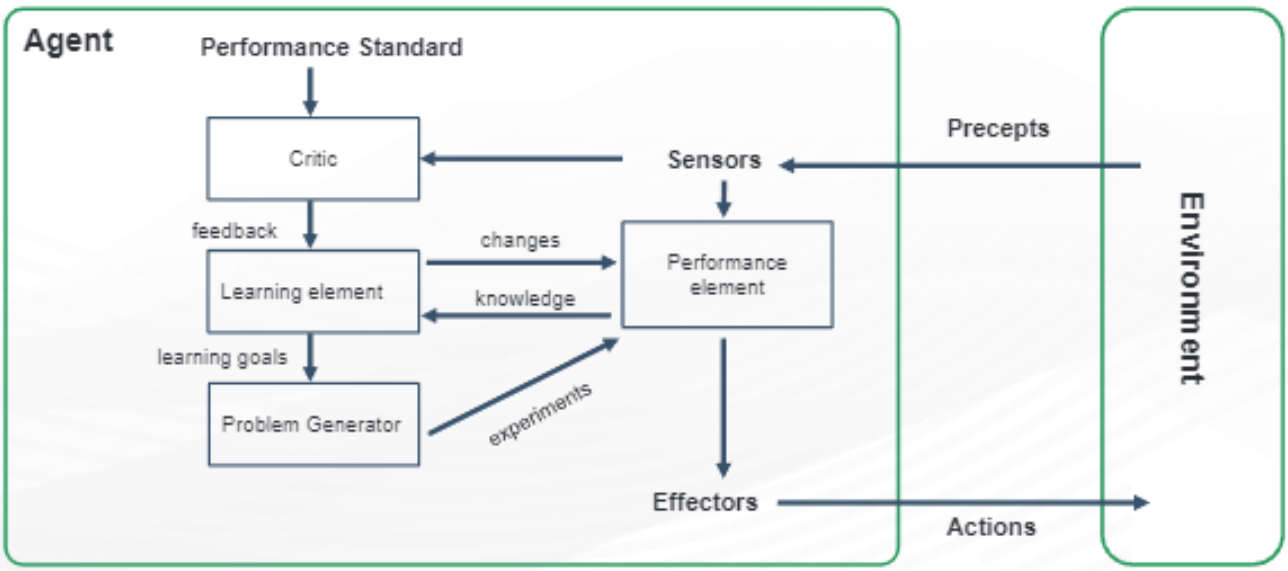
Goals alone are not enough to generate high-quality behavior in most environments. For example, many action sequences will get the taxi to its destination (thereby achieving the goal) but some are quicker, safer, more reliable, or cheaper than others. Goals just provide a crude binary distinction between "happy" and "unhappy" states. A more general performance measure should allow a comparison of different world states according to exactly how happy they would make the agent. Because "happy" does not sound very scientific, economists and computer scientists use the term **utility** instead.

We have already seen that a performance measure assigns a score to any given sequence of environment states, so it can easily distinguish between more and less desirable ways of getting to the taxi's destination. An agent's **utility function** is essentially UTILITY FUNCTION an internalization of the performance measure. If the internal utility function and the external performance measure are in agreement, then an agent that chooses actions to maximize its utility will be rational according to the external performance measure.

Adapted from - https://www.passeidireto.com/arquivo/18996209/livro-artificial-intelligence-a-modern-approach-3rd-edition-/26

8.3.6 Learning Based Agents

- Supervised Learning. Teaching using Examples and Data and Relative Output.
- Autonomous house system like JARVIS in Ironman Movie.



Copyright © 2018, Xebia Group. All rights reserved.

In many areas of AI, this is now the preferred method for creating state-of-the-art systems. Learning has another advantage, as we noted earlier: it allows the agent to operate in initially unknown environments and to become more competent than its initial knowledge alone might allow.

A learning agent can be divided into four conceptual components:

The most important distinction is between the **learning element**, which is responsible for making improvements, and the **performance element**, which is responsible for selecting external actions. The performance element is what we have previously considered to be the entire agent: it takes in percepts and decides on actions. The learning element uses feedback from the **critic** on how the agent is doing and determines how the performance element should be modified to do better in the future.

The critic tells the learning element how well the agent is doing with respect to a fixed performance standard. The critic is necessary because the percepts themselves provide no indication of the agent's success.

The last component of the learning agent is the **problem generator**. It is responsible for suggesting actions that will lead to new and informative experiences. The point is that if the performance element had its way, it would keep doing the actions that are best, given what it knows. But if the agent is willing to explore a little and do some perhaps suboptimal actions in the short run, it might discover much better actions for the long run. The problem generator's job is to suggest these exploratory actions.

Adapted from - https://www.ques10.com/p/13489/what-are-the-basic-building-blocks-of-learning-a-1/

What did You Grasp?

1. What type of Agent is a:
   A) Loan Default Prediction System
   B) Google Maps route detection Algorithm
   C) Determining the shortest path in Google Maps
   D) An IR sensor that sends an alert every time someone crosses it
2. The performance of an agent can be improved by
   A) Learning
   B) Observing
   C) Perceiving
   D) None of the mentioned

 Copyright © 2018, Xebia Group. All rights reserved.

9. Types of Environments

Fully observable vs. partially observable	If an agent's sensors give it access to the complete state of the environment at each point in time, then we say that the task environment is fully observable. An environment might be partially observable because of noisy and inaccurate sensors or because parts of the state are simply missing from the sensor data.
Single agent vs. multiagent	An agent solving a crossword puzzle by itself is clearly in a single-agent environment, whereas an agent playing chess is in a two agent environment.
Deterministic vs. stochastic	If the next state of the environment is completely determined by the current state and the action executed by the agent, then we say the environment is deterministic; otherwise, it is stochastic.

Adapted from: https://enacademic.com/dic.nsf/enwiki/4030454

A task environment is effectively fully observable if the sensors detect all aspects that are *relevant* to the choice of action; relevance, in turn, depends on the performance measure. Fully observable environments are convenient because the agent need not maintain any internal state to keep track of the world.

If the environment is partially observable, however, then it could *appear* to be stochastic. Most real situations are so complex that it is impossible to keep track of all the unobserved aspects; for practical purposes, they must be treated as stochastic.

Copyright © 2018, Xebia Group. All rights reserved.

Episodic vs. sequential

In an episodic task environment, the agent's experience is divided into atomic episodes. In each episode the agent receives a percept and then performs a single action. Crucially, the next episode does not depend on the actions taken in previous episodes. Many classification tasks are episodic. In sequential environments, on the other hand, the current decision could affect all future decisions.

Static vs. dynamic

If the environment can change while an agent is deliberating, then we say the environment is dynamic for that agent; otherwise, it is static. If the environment itself does not change with the passage of time but the agent's performance score does, then we say the environment is semidynamic.

Discrete vs. continuous

The discrete/continuous distinction applies to the state of the environment, to the way time is handled, and to the percepts and actions of the agent.

Known vs. unknown

In a known environment, the outcomes for all actions are given. Obviously, if the environment is unknown, the agent will have to learn how it works in order to make good decisions.

Adapted from: https://en.academic.com/dic.nsf/enwiki/4030464

Chess and taxi driving are sequential: in both cases, short-term actions can have long-term consequences. Episodic environments are much simpler than sequential environments because the agent does not need to think ahead.

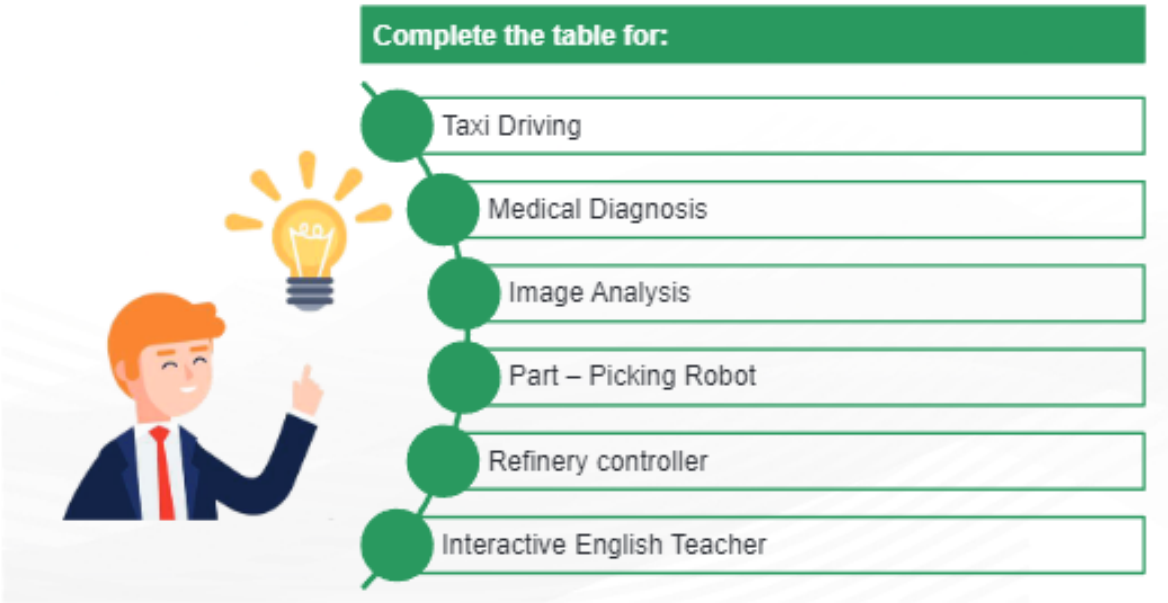
Static environments are easy to deal with because the agent need not keep looking at the world while it is deciding on an action, nor need it worry about the passage of time. Dynamic environments, on the other hand, are continuously asking the agent what it wants to do; if it hasn't decided yet, that counts as deciding to do nothing.

Taxi-driving actions are also continuous (steering angles, etc.). Input from digital cameras is discrete, strictly speaking, but is typically treated as representing continuously varying intensities and locations.

Task Environment	Observable	Agents	Deterministic	Episodic	Static	Discrete
Crossword puzzle	Fully	Single	Deterministic	Sequential	Static	Discrete
Chess with a clock	Fully	Multi	Deterministic	Sequential	Semi	Discrete
Poker	Partially	Multi	Stochastic	Sequential	Static	Discrete
Backgammon	Fully	Multi	Stochastic	Sequential	Static	Discrete

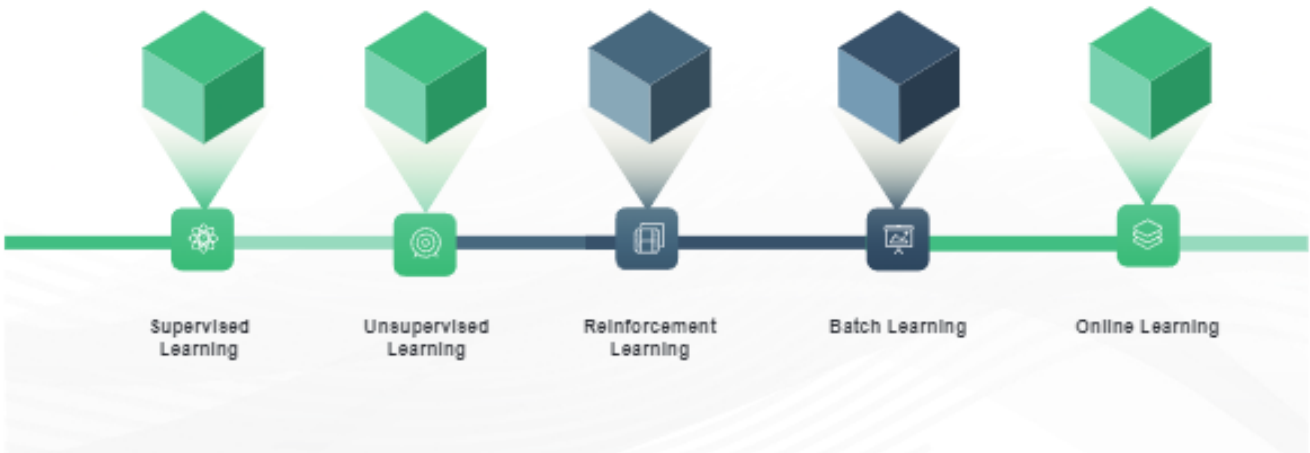
 Copyright © 2018, Xebia Group. All rights reserved.

Quiz



10. Types of Learning

The most important you need to focus on are as follows:



There are so many different types of Machine Learning systems that it is useful to classify them in broad categories based on:

- Whether or not they are trained with human supervision (supervised, unsupervised, semi supervised, and Reinforcement Learning)
- Whether or not they can learn incrementally on the fly (online versus batch learning)

Copyright © 2018, Xebia Group. All rights reserved.

- Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do (instance-based versus model-based learning)

10.1 Supervised Learning

Supervised Learning

Training data is fed with desired solutions, called *labels*. Supervised learning typically includes the task of *classification*.

Example: ***Spam filter***

In *supervised learning*, the training data you feed to the algorithm includes the desired solutions, called *labels*. A typical supervised learning task is *classification*. The spam filter is a good example of this: it is trained with many example emails along with their *class* (spam or ham), and it must learn how to classify new emails.

Another typical task is to predict a *target* numeric value, such as the price of a car, given a set of *features* (mileage, age, brand, etc.) called *predictors*. This sort of task is called *regression*. To train the system, you need to give it many examples of cars, including both their predictors and their labels.

10.2 Unsupervised Learning

Unsupervised Learning

In *unsupervised learning*, as you might guess, the training data is unlabeled. The system tries to learn without a teacher.

In *unsupervised learning*, as you might guess, the training data is unlabeled. The system tries to learn without a teacher.

For example, it might notice that 40% of your visitors are males who love comic books and generally read your blog in the evening, while 20% are young sci-fi lovers who visit during the weekends, and so on. If you use a *clustering* algorithm, it may also subdivide each group into smaller groups. This may help you target your posts for each group.

 Copyright © 2018, Xebia Group. All rights reserved.

10.3 Semi Supervised Learning

Semi Supervised Learning

Some algorithms can deal with partially labelled training data, usually a lot of unlabelled data and a little bit of labelled data. This is called *semi supervised learning*

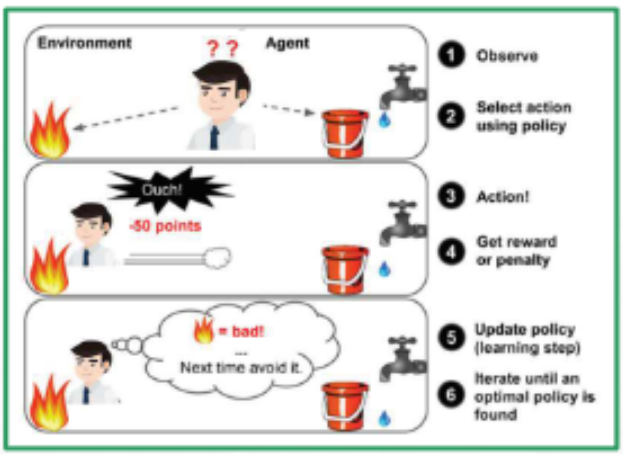
Some photo-hosting services, such as Google Photos, are good examples of semi supervised learning. Once you upload all your family photos to the service, it automatically recognizes that the same person A shows up in photos 1, 5, and 11, while another person B shows up in

photos 2, 5, and 7.

This is the unsupervised part of the algorithm (clustering). Now all the system needs is for you to tell it who these people are. Just one label per person,4 and it is able to name everyone in every photo, which is useful for searching photos.

10.4 Reinforcement Learning

Reinforcement Learning is a very different beast. The learning system, called an *agent* in this context, can observe the environment, select and perform actions, and get *rewards* in return (or *penalties* in the form of negative rewards). It must then learn by itself what is the best strategy, called a *policy*, to get the most reward over time. A policy defines what action the agent should choose when it is in a given situation.



For example, many robots implement Reinforcement Learning algorithms to learn how to walk. DeepMind's AlphaGo program is also a good example of Reinforcement Learning; it made the headlines in March 2016 when it beat the world champion Lee Sedol at the game of *Go*.

Copyright © 2018, Xebia Group. All rights reserved.

10.5 Batch Learning

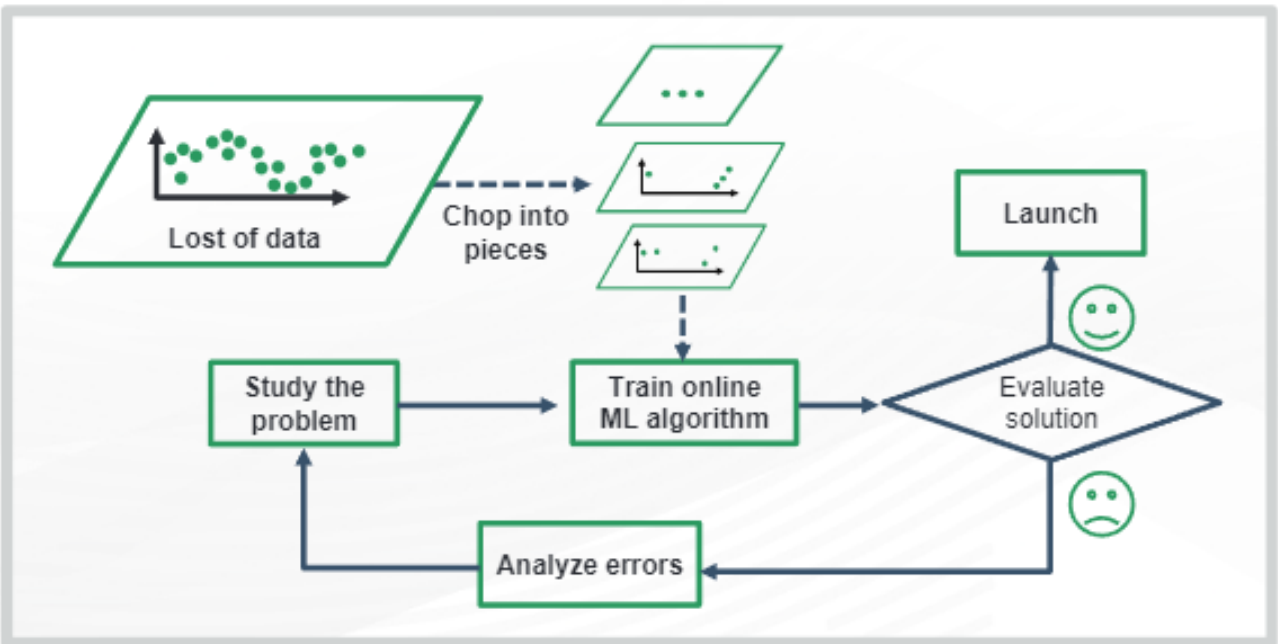
Batch Learning

In *batch learning*, the system is incapable of learning incrementally: it must be trained using all the available data. This will generally take a lot of time and computing resources, so it is typically done offline. First the system is trained, and then it is launched into production and runs without learning anymore; it just applies what it has learned. This is called *offline learning*.

If you want a batch learning system to know about new data (such as a new type of spam), you need to train a new version of the system from scratch on the full dataset (not just the new data, but also the old data), then stop the old system and replace it with the new one.

10.6 Online Learning

In *online learning*, you train the system incrementally by feeding it data instances sequentially, either individually or by small groups called *mini-batches*. Each learning step is fast and cheap, so the system can learn about new data on the fly, as it arrives.



If you have limited computing resources: once an online learning system has learned about new data instances, it does not need them anymore, so you can discard them (unless you want to be able to roll back to a previous state and "replay" the data). This can save a huge amount of space.

 Copyright © 2018, Xebia Group. All rights reserved.

Online learning algorithms can also be used to train systems on huge datasets that cannot fit in one machine's main memory (this is called *out-of-core* learning). The algorithm loads part of the data, runs a training step on that data, and repeats the process until it has run on all of the data.

11. Areas of Artificial Intelligence

- Machine Learning - Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves.
- Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. The basic intuition of deep learning is that we try to replicate working of a human brain. The neuron is the base of our brain. Similarly, we try to make a neural network, which can replicate the working of the human brain. More than one layer of neurons is called as Deep Learning. Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost. It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers. Deep learning is getting lots of attention lately and for good reason. It's achieving results that were not possible before.

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labelled data and neural network architectures that contain many layers.

- Natural Language Processing – Performing a task related to test or language in which humans write or read. Like Google Translation, Consumer complaint classification.
- Computer Vision – Giving sight to machines related to images or videos or we can say performing tasks related to images/videos.
- Robotics – Adding Speech for response and vision for sight to any robot makes it even smarter as compared to simple mechanical robots.

Copyright © 2018, Xebia Group. All rights reserved.

11.1 Why is Deep Learning Gaining so much Popularity?

Go through the points mentioned on the slide.

12. Characteristics of Artificial Intelligence

- Artificial Intelligence is growing rapidly in the field of computer vision and natural language processing.
- Suppose you want to write a program to differentiate between a dog and cat. How will you go about it?

Characteristics

- AI automates repetitive learning and discovery through data
- AI adapts through progressive learning algorithms
- AI analyses more and deeper data
- AI achieves incredible accuracy
- AI gets the most out of data

Computer Vision – Performing a task related to images / videos

Natural Language Processing – Performing a task related to test or language in which humans write or read.

Speech Recognition – Performing task related to audio or speech.

 Copyright © 2018, Xebia Group. All rights reserved.

Can a machine differentiate between a cat and dog?

You will have to hard code the rules that make a cat and dog different from each other. There could be thousands of rules you could state and still there might be a possibility you would miss a few points. Whereas, using Artificial intelligence techniques you give this power to your program. You have to just provide it with relevant data. It's now the responsibility of your code to learn the differentiating features from your data.

1. Instead of automating manual tasks, AI performs frequent, high-volume, computerized tasks reliably and without fatigue, which brings up better results and improvement in performance.
2. AI finds structure and regularities in data so that the algorithm acquires a skill The algorithm becomes a classifier or a predictor. So, just as the algorithm can teach itself how to play chess, it can teach itself what product to recommend next online.
3. Getting deeper into data helps the machine identify patterns which are not possible even for humans to find out. Also with increase in technologies like Big data and cloud computing, which can handle and process huge amounts of data making an immense rise in Artificial Intelligence.
4. Accuracy is achieved higher and higher with improvement in the algorithms you use. Also, with few self learning algorithms you improve the machine's accuracy itself.
5. When algorithms are self-learning, the data itself can become intellectual property. The answers are in the data; you just have to apply AI to get them out. Since the role of the data is now more important than ever before, it can create a competitive advantage. If you have the best data in a competitive industry, even if everyone is applying similar techniques, the best data will win.

12.1 Applications of Artificial Intelligence

There are immense applications of Artificial Intelligence

- Self driving cars (Tesla and Waymo)
- Chatbots
- Email Spam or Ham detection
- Disease (Cancer, Malaria) detection from Human Cell Images
- Google Assistant, Amazon Alexa, Siri, Google Home
- Sentiment Analysis
- Credit card / Loan default prediction.
- OCR (Optical character recognition)
- Face Recognition
- Recommendation Engines
- Latest and most interesting application bought up by Google

Copyright © 2018, Xebia Group. All rights reserved.

Self driving cars is one of the applications of computer vision where the car has enough intelligence to differentiate between a person, another vehicle or pit hole identifying which it has to change to speed or apply brakes.

Chatbots are one of the most rapidly growing application as it helps a lot of businesses to improve their customer satisfaction and also help exponentially in cost cutting. No actual person is sitting on the other to answer your queries, it's all done by machine. Understanding your problem and providing you with the appropriate solution. Eg., Zomato, Swiggy.

Email Spam Detection – Have you ever noticed how many emails you receive daily in your inbox. Do you open all of them? Thanks to Google (Gmail) that their system is smart enough to segregate your useful mails (ham) and spam. The main idea to look for specific patterns in mails to categorize them. According to Google reports their system is able to identify spam and ham with 99.9% accuracy which is amazing.

Disease Detection – The model is trained on thousands of images and is asked to learn the features which you think might cause cancer. After we are satisfied with the performance of the model we use it to test on the new images and classify them.

Google Assistant, Alexa and Siri are all voices based assistants working on speech to text and text to speech. It is intelligent enough to understand your voice and decode into commands. This will help you do task only with your voice. More the data provided to them the accuracy of the assistant improves. This type of learning is called reinforcement learning where the model is continuously learning with new input data.

Sentiment Analysis – Machine is able to understand the sentiment of the text or audio provided to it. Detecting feelings at a high level makes the machine act in a similar way. Three main sentiments are Positive, Negative or Neutral. Sentiment Analysis is highly used in product/restaurant reviews done by customers on e-commerce websites to find out how people are reacting to their product.

Credit Card default detection – Banks use this model to find out if a person will be able to payback the loan or not using the huge amount of past data and finding out patterns which will be helpful for the bank to save on their assets.

OCR – It is an application of artificial Intelligence where the machine is able to extract text from an image. Like getting your information from your Aadhar card you gave to the bank. Or Extracting text from your personal certificates.

Face Recognition – Smartphones having a face unlock and companies taking attendance using Face recognition.

Recommendations – Movies recommended by Netflix and products recommend by Flipkart and Amazon work on Machine Learning. Based on your past history and your likes they recommend you movies/products.

 Copyright © 2018, Xebia Group. All rights reserved.

12.2 Real world Application by Big Companies

1. Banks using Loan and Credit card default prediction
2. Tesla, Uber and Waymo into Self driving cars
3. Haptik.ai – India's leading startup in chatbots
4. Cure.ai – Working in disease detection only with your X ray images.
5. Netflix and Amazon Prime – Movie recommendations
6. Google Duplex, Google Photos tagging

Go through the points mentioned on the slide.

Case Study-I

- Artificial Intelligence is highly being used in E-commerce industry and helping them earn millions.
- Here is a small case study how Amazon, Alibaba are using AI to increase their revenue.
  https://apiumhub.com/tech-blog-barcelona/artificial-intelligence-ecommerce/

Case Study-II

- With the rise of Digital Media, the one thing that also rises is the uses of fake videos and pictures. Adobe, with its tool Photoshop, leads the race of most use image manipulation software in fraud cases.
- Here is a small case study of Adobe coming up with a new feature or tool that uses machine learning to automatically detect when images of faces have been manipulated.

  https://techgrabyte.com/adobes-ai-photoshopped-images/

Copyright © 2018, Xebia Group. All rights reserved.

Test Your Understanding

1. Artificial Intelligence is all about?
   A) Coding
   B) Hardware
   C) Relevant Data
   D) Learning

2. Artificial Intelligence has its expansion in the following application?
   A) Planning and Scheduling
   B) Game Playing
   C) Robotics
   D) All of the Above

In a nutshell, we learnt

1. The basics of AI and its history
2. The basics of ML and its history
3. Different ML algorithms
4. The overlap of ML and AI
5. Practical applications of AI and ML

 Copyright © 2018, Xebia Group. All rights reserved.

MODULE 2

Types of Data

You will learn about the 'Types of Data' in this module.

Module Learning Objectives

At the end of this module, you will be able to:

- List the different types of data.
- Define Structured Data and its organization.
- Define Unstructured Data.
- Define Semi-structured Data.
- Compare and contrast the different features of Structured, Unstructured and Semi-structured Data.
- Understand the concept of noise from a data viewpoint.

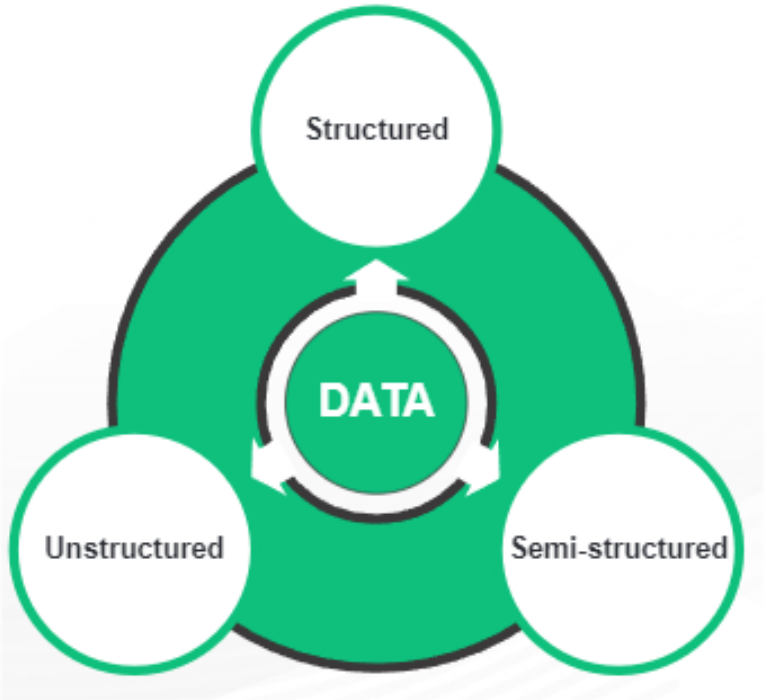
Module Topics

The following topics that will be covered in the module:

1. Structured Data
2. Semi-structured Data
3. Unstructured Data
4. Comparison of Structured, Unstructured and Semi-structured Data
5. Presence of Noise in Data.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1 Types of Data



In the previous module, you learnt about Data, its definition and sources of Data. In this module, you're going to learn about the different types of Data. There are three major types of Data.

- **Structured Data:** Information stored relational databases belong to this category. Data has a strict structure and the schema has to be predefined before populating data in a relational database. It is the Database Management System (DBMS) that checks and ensures that the populated data follows the structure as specified in the schema.
- **Semi-structured Data:** Data that is collected in an ad-hoc manner before the storage and management methods are known. This data is portable and does not require the schema to predefined, though schema definition is not impossible. Data might have some structure, but all information will not have an identical structure. This type of data is also referred to as self-describing data.
- **Unstructured data:** Data that doesn't have any fixed structure. There is very less indication of the type of data. There is no predefined data model for this type of data. Since there is no predefined schema for storing unstructured data, complex queries are needed to search and retrieve information from unstructured data.

In the upcoming sections, we will learn about each of these data types in detail along with examples, their advantages and disadvantages.

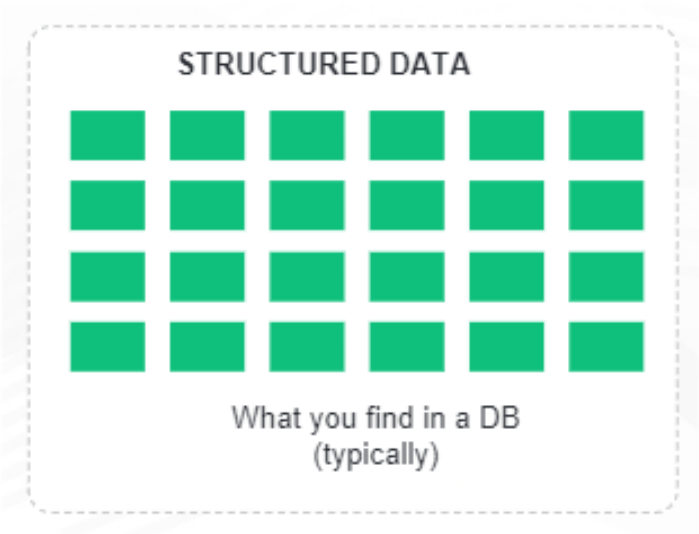
This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1.1 Structured Data

Structured Data:

- Is highly organized and readily searchable by queries or algorithms.
- Can be quickly consolidated into facts.
- Follows a predefined schema.
- Usually resides in fixed fields.
- A typical example is a Relational Database Management System (RDBMS).
- A Schema is defined before the content is created and the data is populated.



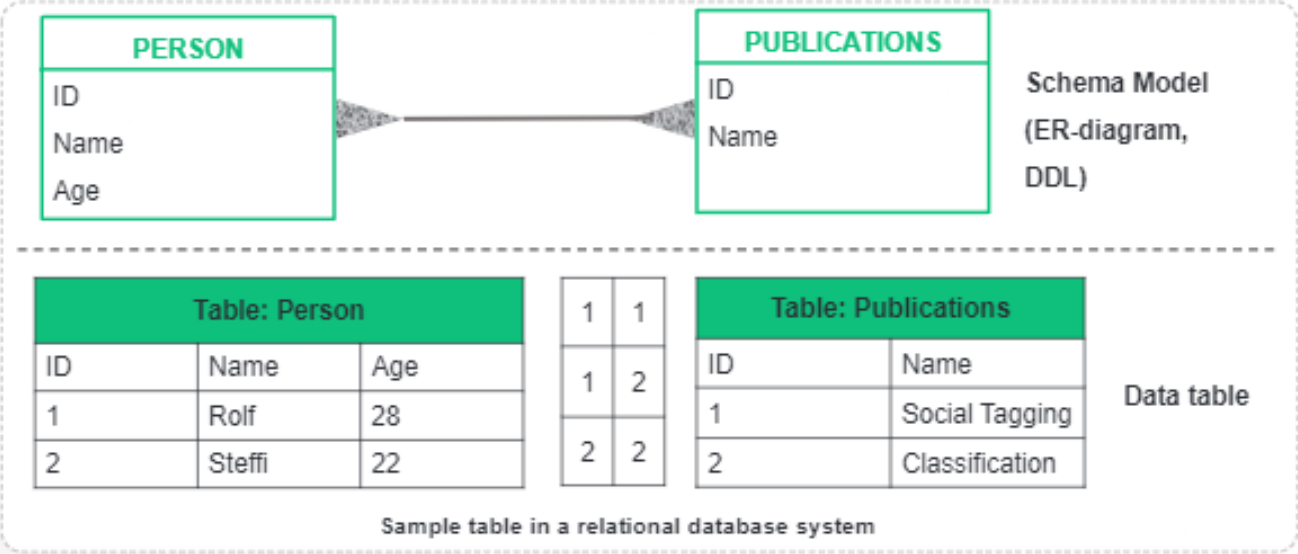
Structured Data is a simple and highly organized type of Data. Data is written in a defined format, that is easy for the machines to understand and is easily searchable even by the basic algorithms. Structured Data is usually present in relational databases (RDBMS) and is retrieved using queries and algorithms using data-type and the field names.

Structured Data can be entered, stored, queried and analyzed using simple and efficient ways. Field-name and type of data should strictly be defined before populating in an RDBMS. Hence, Structured Data may face restrictions in terms of the number of characters or specific terminologies. Simple or complex queries in Excel spreadsheets or Structured Query Language (SQL) are used to retrieve information from a relational database.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1.2 Organization of Structured Data

The following diagram and tables illustrates the structured data.



Sample table in a relational database system

Structured Data usually resides in Relational databases or Data Warehouses. The Structured Data contains usually text files, displayed in titled columns and rows which can easily be queried and processed by Data Mining tools.

Let's take the example of a Relational Database System to understand the organization of structured data. The figure above gives an idea of an Entity-Relationship (ER) diagram and its concrete tables within a Relational Database Management System (RDBMS).

In a Relational Database System, the data is stored in tables with predefined columns. The Column title describes the type of information stored in a table. Designing a database schema is an elaborate process. Schema is the one which defines the type and structure of the data and its relations. In a relational database, the schema is defined well before content is created and the data is populated. The well-defined schema of a fully structured data helps the in-efficient data processing and improved storage and navigation of content.

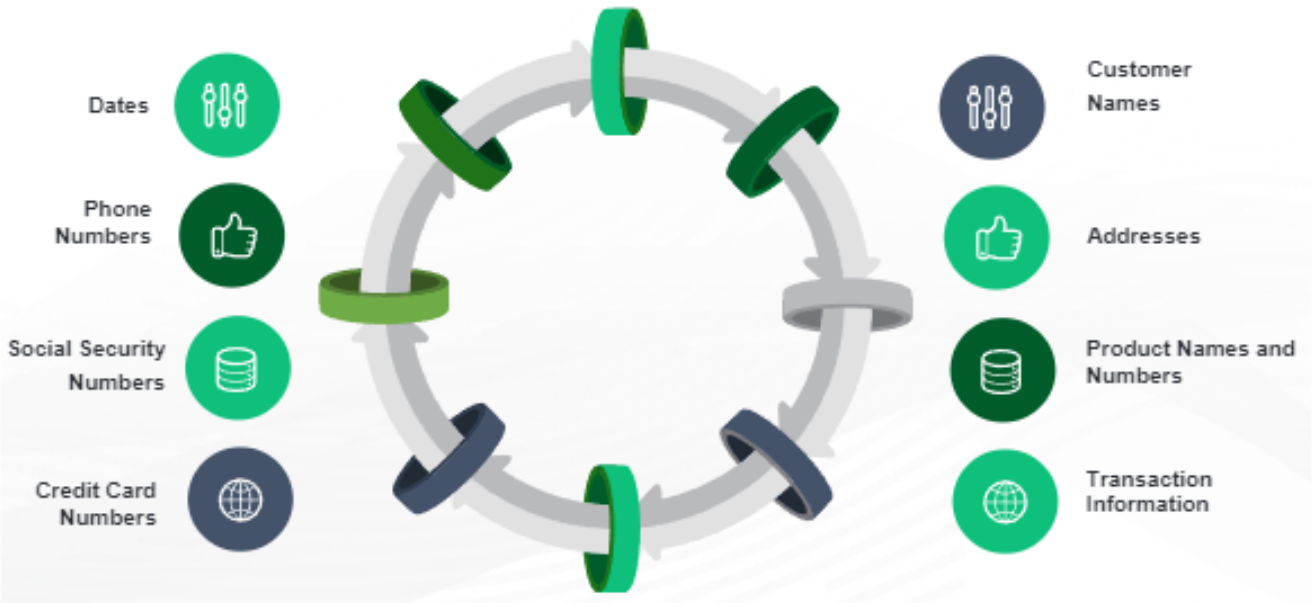
In a Relational Database, extending previously-defined schemas that already have content, might be difficult. For example, to extend a single table row with a new attribute, another table column must be created. This will not be suitable for tables with thousands of other rows that do not need another attribute.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1.3 Examples of Structured Data

Some of the examples of Structured data:

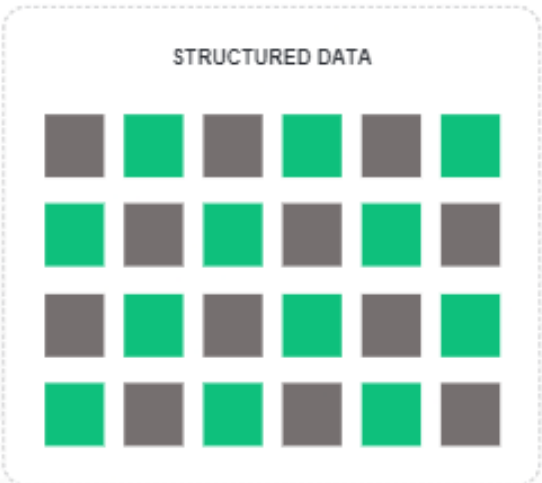


You may note down a few examples of Structured Data.

1.4 Applications of Structured Data

Relational database applications with Structured Data:

- Airline reservation systems
- Inventory control
- Sales transactions
- ATM activity



This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Some of the applications where data is structured and well-organized are:

- Airline reservation systems
- Inventory control
- Sales transactions
- ATM activity

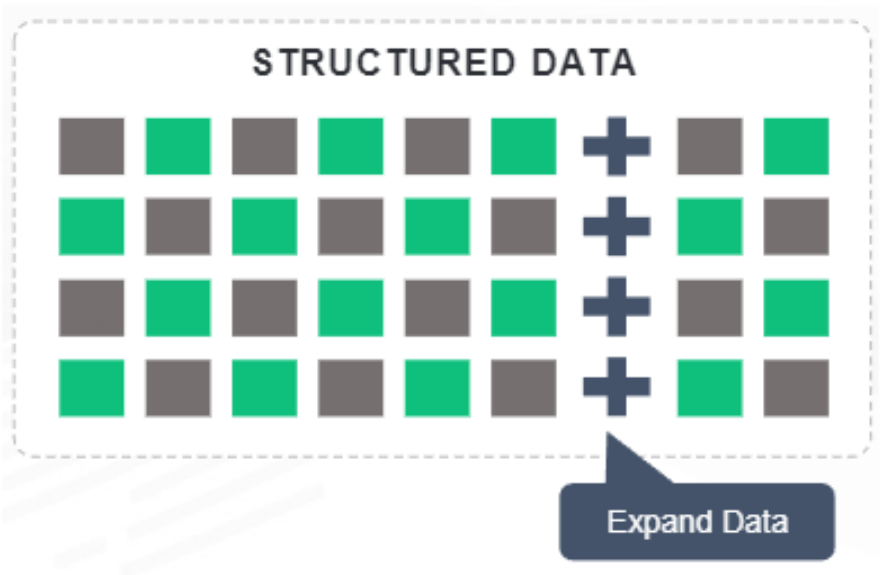
1.5 Expansion of Structured Data

The following are some reasons of expansion of Structured data:

- Creation of multiple backups of data and database.
- Having different databases like test database, development database, reporting database, production database etc.
- Having multiple copies of production database.
- Duplication of data: having multiple copies of the same data. One of the common causes of data expansion.

Techniques for controlling data expansion:

- Data Compression
- Deduplication
- Cloning
- Archiving



This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Understand how Structured Data expands. Data growth is unavoidable. Data expansion actually results in consumption of more storage space and there are associated complexities with data occupying more space. The cost of data storage and management also increases.

One of the major reasons for data expansion is unnecessary data duplication. Having multiple copies of the same data might be useful from a backup point of view. But this requires a lot of storage space and updating a copy of data doesn't mean that all copies of it will be updated. This results in data reliability issues as well. Data expansion can be controlled by implementing the following techniques:

- Data compression
- Deduplication
- Cloning
- Archiving

1.6 (a) Advantages and Disadvantages of Structured Data

The following are advantages of Structured data:

Advantages

- Optimized query evaluation.
- Improved storage.
- Construction of indexes for describing the database content to the user and facilitating query formulation is simpler.
- Proscribing certain updates is easy.
- Supports strongly typed languages.
- Data analysis is easy compared to unstructured data.

Structured Data has its own advantages and disadvantages.

Some of the advantages include:

- Query evaluation is optimized since data is well organized. Simple queries have the ability to fetch the desired information in less time.
- The Storage is simpler as the structure is defined.
- Describing the database content to the user becomes simpler and construction of queries and indexes to retrieve the information is also much simpler.
- Data analysis is much simpler compared to Unstructured Data.
- Strongly typed languages are supported.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1.6 (b) Advantages and Disadvantages of Structured Data

The following are disadvantages of Structured Data:

Disadvantages

- Schema needs to be predefined before populating data.
- Not suitable if the amount of data is very high.
- Cannot be used for data that keeps changing frequently.
- Subsequent extensions of a previously defined schema with content is difficult.

Some of the disadvantages include:

- Data cannot be populated to the database, without a predefined schema.
- This requires much time and effort even before data is loaded. In case of an update to the database, a complete database is required to be modified from the scratch, especially .
- If the amount of data is very high and from multiple sources, it is not possible to give a predefined structure to the data.
- Frequently changing data cannot be categorised under a defined structure.

What did You Grasp?

1. State True or False.

   Structured Data requires a predefined schema before the data is stored in database.

   A) True
   B) False

2. Which of the following options is an example of Structured Data?

   A) Email
   B) Videos
   C) Images
   D) Phone numbers

2 What is Unstructured Data?

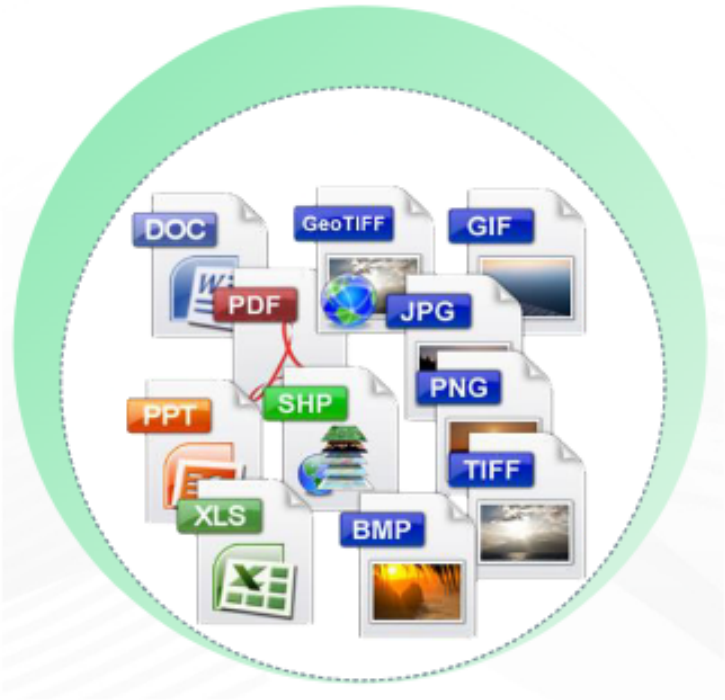
Unstructured Data:

- No identifiable structure for this kind of data
- Have internal structure, but no predefined schema

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- Cannot be stored in rows and columns like a relational database
- No fixed data model, a massive unorganized conglomerate of various information
- Require more storage space than structured data



Around 80% of business data is Unstructured Data. Unstructured Data is defined as data that has no identifiable structure. There is an internal structure to unstructured data, but it is not structured by pre-defined data models or schema. Unstructured Data may be textual or non-textual and human-generated or machine-generated. This primarily involves text files, audio, video and images. Without preprocessing, we cannot store unstructured data in a table.

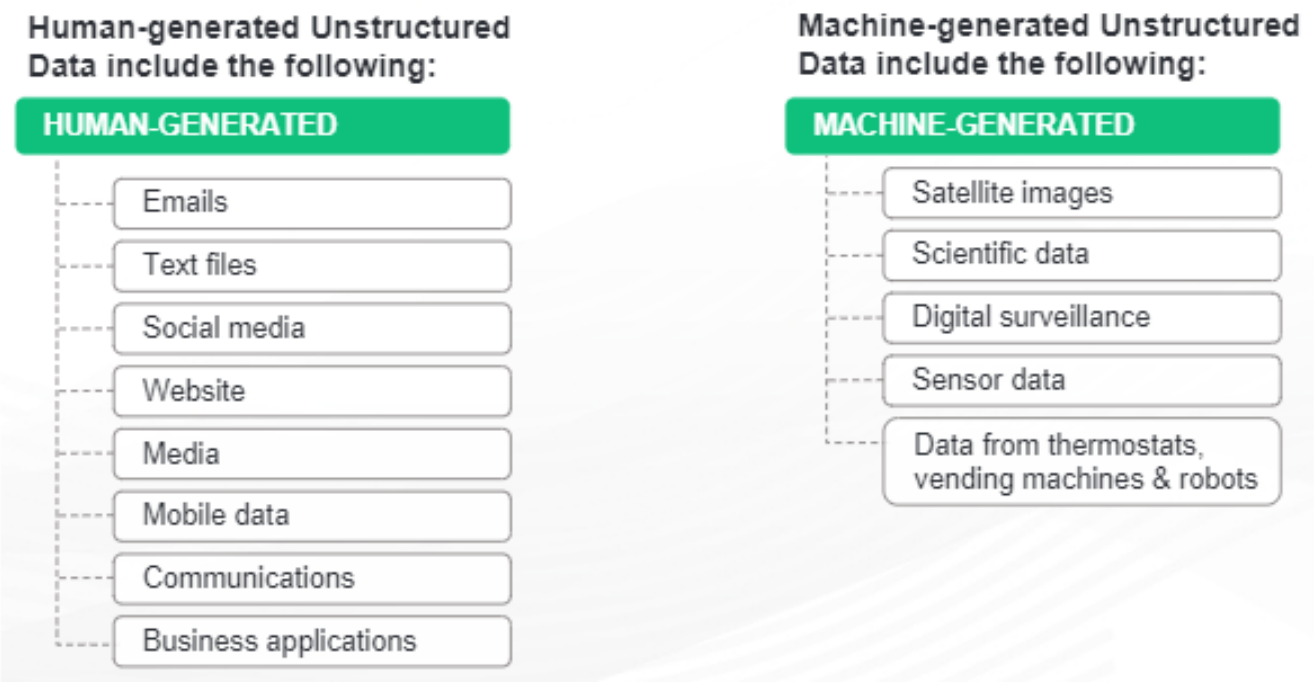
The term 'Unstructured data' needs to be understood properly. We can call a data 'unstructured' if it has some form of structure but is not helpful to processing. For example, the email messages have data with some implied structure, still, normal data mining tools cannot parse the information in an email. Therefore, it comes under Unstructured Data category.

Searching information from the Unstructured Data is not straightforward, as compared to Structured Data. Complex queries and algorithms are needed to retrieve information from an unstructured data.

Full-text search is one of the ways of searching text documents, but this is not suitable for pictures or videos.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

2.1 Examples of Unstructured Data



Human-generated Unstructured Data include the following:

- **Text files:** Word processing, spreadsheets, presentations, email, logs.
- **Email:** Email has some internal structure, thanks to its metadata, and we sometimes refer to it as semi-structured. However, its message field is unstructured and traditional analytics tools cannot parse it.
- **Social Media:** Data from Facebook, Twitter, LinkedIn.
- **Website:** YouTube, Instagram, photo sharing sites.
- **Mobile data:** Text messages, locations.
- **Communications:** Chat, IM, phone recordings, collaboration software.
- **Media:** MP3, digital photos, audio and video files.
- **Business applications:** MS Office documents, productivity applications.

Machine-generated unstructured data include the following:

- **Satellite imagery:** Weather data, landforms, military movements.
- **Scientific data:** Oil and gas exploration, space exploration, seismic imagery, atmospheric data.
- **Digital surveillance:** Surveillance photos and video.
- **Sensor data:** Traffic, weather, oceanographic sensors.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Common examples of the types of files considered as unstructured data:

- Books
- Health records
- Satellite images
- Adobe PDF files
- A warranty request created by a customer service representative
- Notes in a web form
- Objects from presentations
- Blogs
- Text messages
- Word documents
- Videos
- Photos and other images

2.1 (a) Advantages and Disadvantages of Unstructured Data

The following are advantages of Unstructured Data:

Advantages

- More information is contained in Unstructured Data.
- Almost 80% of all the data available is unstructured.
- Mostly all important and critical data fall under unstructured data.
- Analysis of Unstructured Data provides competitive advantages.
- Reveals customer trends and improves customer satisfaction.
- Insights, patterns and concepts are deeply buried in Unstructured data.

Like structured data, unstructured data also has several advantages and disadvantages. Let's look at some of these.

Advantages of Unstructured Data are:

- The information here is not organized into a defined 'Schema'. Almost all that is outside a Relational database is Unstructured Data. There are many types and sources of unstructured data. This means more information is contained in them
- Amount of unstructured data is astonishing. This category makes up nearly 80% of all digitally stored data which is very advantageous for data-driven businesses.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- Unstructured Data critical to businesses are commonly present in websites, presentations or documents. All the medical records are unstructured. The test reports and values are mostly stored as unstructured data.
- Despite the fact that unstructured data is not well-organized or easy to access, analyzing this data can offer competitive advantages to businesses. Especially in this era of Artificial Intelligence and Machine Learning, data is everything. The way data is visualized, analyzed, processed and insights are derived is critical for businesses. This information can give a clear picture of the customer behaviour and requirements, which is a major driver for business growth.
- Analyzing social trends such as tweets, Facebook posts and transcripts from support calls will give a clear view of how customers perceive a value regarding the products. Understanding the issues proactively and acting upon them can dramatically improve customer satisfaction. The requests for new features can be captured, grouped and prioritized in ways that were previously not possible.
- Mining the sales information, social media sentiments, news and survey data helps us to understand the pattern of sales, why it went up or down. Insights, patterns, and concepts are usually buried deep in human-language communication data like, emails, notes, surveys, social network posts, tweets, etc. which will not be available to businesses otherwise.

2.1 (b) Advantages and Disadvantages of Unstructured Data

The following are disadvantages of Unstructured data:

Disadvantages

- Controlled navigation within unstructured content is not possible.
- Complex queries or algorithms are needed to search and retrieve information from unstructured data.
- Analytical tools for mining unstructured data are nascent and companies cannot fully leverage the potential of vast amounts of valuable data.
- Storing huge amounts of unstructured data results in higher costs.
- Understanding, analyzing and processing of unstructured data is difficult for non technical business people.

Disadvantages of Unstructured data:

- Since data is diverse and has no defined structure, controlled navigation is not possible with unstructured data.
- Since data is not grouped or organized using a defined schema, it is difficult to retrieve data using simple queries. Complex algorithms need to be developed to search and retrieve the necessary information from the huge pile of data.
- Unless data is analyzed, there is no value to it. Analyzing unstructured data requires sophisticated tools. Many enterprises lack these tools and are not able to leverage the potential of data completely.
- Storing unstructured data is a complex and an expensive exercise for the organizations. Though the cost of storage devices have come down, the associated cost of maintaining data centres is still a nightmare to organizations. Apart from the cost, highly skilled personnel is also required for storing, maintaining and analyzing the data. Many organizations still lack this.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- Non-technical business people still find the concept of unstructured data difficult. A rich technical expertise is required for analyzing and processing unstructured data. The business people will not be able to understand the intricacies, consequently, they will not be able to explain this to their customers.

What did You Grasp?

1. Which of the following options is true with respect to unstructured data?

   A) There is no internal structure to this type of data.
   B) The data model is fixed.
   C) Requires less storage space than structured data.
   D) Preprocessing is required to store unstructured data.

2. Which of the following options is/are human-generated data?

   A) Email
   B) Sensor data
   C) Text files
   D) Scientific data
   E) Videos

3 What is Semi-structured Data?

Semi-structured data:

- Often explained as schema-less or self-describing
- Contains semantic tags, but do not comply with the standards or structure of typical relational databases
- No separate description of the type or structure of the data
- Does not require a schema definition, the definition is not impossible, but it is optional
- Data can have different attributes, and new attributes can be added anytime

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

Semi-structured data is the data that does not reside in a relational database but has some organizational properties that make it easily analyzable. When processed, it can be stored in a Relational database. Though Semi-structured data do not comply with the standards of a Relational database, tags and the other types of markups in them helps in the identification of the individual, distinct entities within the data.

Semi-structured data can have n-level hierarchies that differentiate it from the structured data, where data is represented as a flat table. Documents and databases can be semi-structured. The Semi-structured data also form only 5-10% of all informative data but has critical business use cases.

In semi-structured data, some of the aspects are structured, while the others are not. The Semi-structured data serves as an important source for Big Data analytics but may not be suitable for traditional databases.

3.1 Examples of Semi-structured Data

Examples of Semi-structured Data include:

- **Markup language XML** - a set of documents encoding rules that define a human- and machine-readable format.
- **Open Standard JSON** - a lightweight, plain-text, data-interchange format based on a subset of the JavaScript programming language.
- **NoSQL** - some noSQL databases contain semi-structured data.

Examples of semi-structured data include:

- **XML:** XML is a semi-structured document language. XML defines the set of rules for encoding documents. XML's initial focus was on the documents, but later it found many use cases like a representation of arbitrary data structures and serving as the base language for communication protocols. The tag-driven structure of XML is highly flexible, and developers can use it for globalizing the data structure, storage and transportation on the web.
- **JSON:** JSON is a lightweight, plain-text, data-interchange format based on a subset of the JavaScript programming language. One of the semi-structured data interchange formats. JSON structure can contain name/value pairs (or object, hash table, etc.) and an ordered value list (or array, sequence, list). The structure of JSON is interchangeable among languages and is very helpful in transmitting data between web applications and servers.
- **NoSQL:** Many NoSQL ("not only SQL") databases contain semi-structured data. In NoSQL databases, a schema is not separated from the data and is very flexible. For example, the text with varying lengths cannot be stored in a relational database but stored in a NoSQL DB. Some newer NoSQL databases like MongoDB and Couchbase store data in the JSON format which is semi-structured.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

3.2 Applications of Semi-structured Data

Applications containing semi-structured data include:

- Big Data Infrastructure.
- Web applications
  - LinkedIn
  - Salesforce
  - Reader recommendations in Amazon

3.3 (a) Advantages and Disadvantages of Semi-structured Data

The following are advantages of Semi-structured data:

Advantages

- Data not constrained by a fixed schema
- Flexibility - data can be easily modified
- Portable
- Possible to view structured data as semi-structured data
- Convenient data transportation configuration

Semi-structured data is an evolving form and is still under research. Let's look at some of the advantages and disadvantages of semi-structured data.

Advantages of Semi-structured data are:

- Data is not constrained by a fixed schema like that in a structured data and is very flexible. This type of data can be used to represent information that cannot be constrained by a schema.
- Semi-structured data offers greater flexibility and it can be easily modified even if the data changes frequently.
- This type of data is portable that the data exchange is also possible between contrasting databases.
- Semi-structured data are very supportive in screening structured data as semi-structured data.
- This type of data is convenient for transportation configuration.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

3.3 (b) Advantages and Disadvantages of Semi-structured Data

The following disadvantages of Semi-structured data:

Disadvantages

- Queries are less efficient than in a constrained structure
- Data diversity
- Extensibility
- Storage

Disadvantages of Semi-structured data are:

- Records in semi-structured database are stored with only one of a kind IDs that are referenced with indicators to their specific locality on a disk. Despite the fact that queries are very well-organized, this approach is not practical while doing searches over scores of records, for the reason that it is forced to seek in the various regions of the disk by following the indicators.
- In federated systems, data diversity is a serious issue. This involves complexities such as unit and semantic incompatibilities, grouping incompatibilities, and non-consistent overlapping of sets.
- It is important that extensibility from the context of data is an indication of data presentation and not data processing. The Data processing should be able to happen without the aid of database updates.
- Transfer formats like XML are universally in text or in Unicode. They are helpful for data transfer but not for storage. The presentations are instead, stored by deep-seated and accessible systems that support such standards.

What did You Grasp?

1. Which of the following options is/are true about semi-structured data?

   A) **Semantic tags need to comply with the standards or a defined structure.**
   B) **Tags and other markups help in identification of entities in the data.**
   C) **Data attributes cannot be modified once added.**
   D) **Structured data cannot be viewed as Semi-structured data.**

2. State True or False.
   Schema definition cannot be done at all, for Semi-Structured data.

   A) **True**
   B) **False**

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Activity

Your facilitator will give different examples of Data. Discuss among yourselves and identify which category they belong to. Justify your answer, why they belong to that category.

4 Comparison of Structured, Unstructured and Semi-structured Data

The following table describes the comparison between all types of data.

	Structured data	Unstructured data	Semi-structured data
Technology	Relational database tables.	Character and binary data like image, text, video, audio etc.	XML/RDF.
Transaction management	Matured transaction management, various concurrency techniques.	No transaction management, no concurrency.	Transaction management adapted from RDBMS, not matured.
Version management	Versioning over tuples, rows, tables, etc.	Versioned as a whole.	Not very common; versioning over triples or graphs possible.
Flexibility	Schema-dependent, very flexible.	Very flexible, absence of schema.	Flexible, tolerant schema.
Scalability	Scaling DB schema is difficult.	Highly scalable.	Schema scaling is simple.
Robustness	Very robust, enhancements since 30 years.	-	New technology, not widely spread.
Query performance	Structured query, allows complex joins.	Only textual queries possible.	Queries over anonymous nodes are possible.
Content type	Sensitive Content & System Maintenance Data, Core Component Data.	Textual Content.	Non-sensitive Core Component Data, Flexible Content & Individual Data.

Comparison of Structured, Unstructured and Semi-structured data will give you a clear idea of which data will be insightful under what scenario.

What did You Grasp?

1. State True or False.
   Structured data is highly scalable compared to unstructured data.

   A) True
   B) False

2. Which of the following options is right, with respect to queries in Structured data?

   A) Structured queries that allow complex joins
   B) Only textual queries are possible
   C) Queries over anonymous nodes are possible
   D) None of the above

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

5 Noisy Data

Noisy Data:

- Contains a large amount of meaningless information, hence considered to be corrupted.

Noise:

- An anomaly and is considered to be an error. According to Zhu et al. (2004), noise in data negatively affects data mining results.
- A persistent phenomenon that is superimposed on desired information.
- Unnecessarily occupies more storage space.
- Two major sources of noise:
  - Implicit errors
  - Random errors

We've learned about the different types of data. The amount of data generated every day is growing in an exponential phase. In order to derive value out of data, it needs to be processed and analyzed. It is also important to distinguish between useful and not-so-useful data before it is processed. Unless useful data is identified early in the process, organizations will end up spending a lot of time, money and effort, for the value output that is much less than the expected. One such phenomenon that affects the results of Data Mining is the presence of Noise in the data.

Noisy data is thus defined as the data that contains a large amount of meaningless information, i.e noise. Noisy data is considered to be the corrupted data. Noise is an anomaly and thus considered to be an error in the data. Zhu et al, in their research article titled "Class Noise vs. Attribute Noise: A Quantitative Study", which was published in the journal, Artificial Intelligence Review in 2004, explained that noise negatively affects data collection and data preparation processes, hence Data Mining results are affected. According to them, there are two major sources of Data Noise:

- **Implicit errors:** Implicit errors are introduced by data measurement tools, such as different types of sensors.
- **Random errors:** Random errors are the ones that are generated by Batch Processing Systems or experts who collect them. One example is an error that is introduced during Data Digitization process.

The performance of data classifiers depends, to a larger extent, on the quality of data used for training the machine learning models. Thus, noise in data may have a direct impact on the predictions made by machine learning models.

There is another phenomenon called 'outliers' which should not be confused with noise. The outliers are not meaningless but are out of range data. They are still legitimate, but they do not comply with the general behaviour of data. They sometimes are considered while data processing, according to the need. One example of their application is fraud detection systems.

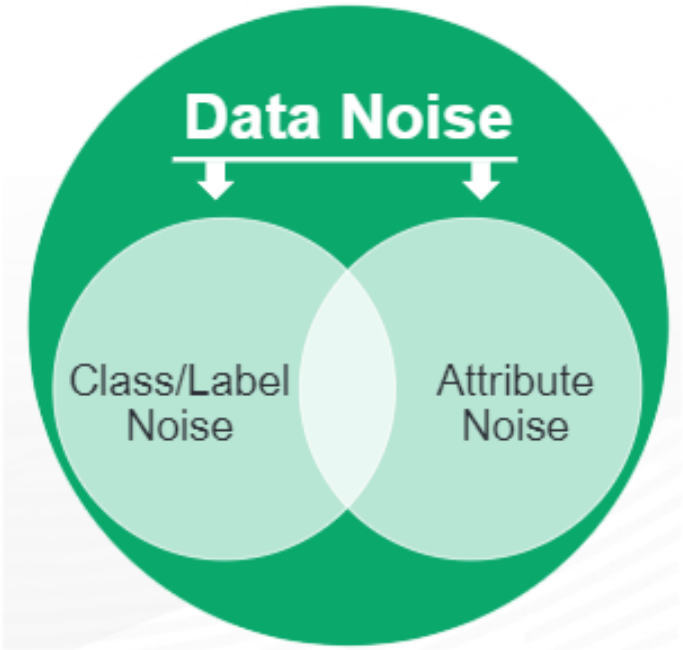
This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Detecting noise in data is very important for improving machine learning performance. Understanding of data and its knowledge can be expanded by means of identifying the exceptional cases in data.

5.1 Types of Data Noise

There are two major categories of data noise as depicted in the illustration below:



There are two major categories of Data Noise:

1. **Class/Label Noise:** Class noise occurs when data is incorrectly labelled. There are several reasons for class noise. Some of the common reasons include: Subjectivity during labelling, data entry errors, insufficient information used to label a data, etc. There are two types of class noise:
   a) **Contradictory examples** - refers to duplicate data that have different class labels.
   b) **Misclassifications** - data that are labelled as a class that is different from the real one.
2. **Attribute Noise:** Attribute noise refers to the corruption that occurs in one or more of attribute values. Some of the examples of attribute noise are:
   a) Erroneous attribute values
   b) Missing or unknown attribute values
   c) Incomplete attributes or "do not care" values

Zhu et al discovered that attribute noise is more harmful than class noise. They also found that eliminating class and attribute noise will improve classifier performance.

Usage of noise filters can greatly reduce noise in data.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

What did You Grasp?

1. State True or False.

   Implicit errors are generated by batch processors.

   A) True
   B) False

2. Which of the following options is an example for class noise?

   A) Missing attribute values
   B) Misclassifications
   C) Incomplete values
   D) Erroneous values

In a nutshell, we learnt:

1. Data types - Structured, Unstructured and Semi-structured Data.
2. Structured Data, its organization, examples, advantages and limitations.
3. Unstructured Data, examples, advantages and limitations.
4. Semi-structured Data and examples.
5. Comparison between Structured, Unstructured and Semi-structured Data.
6. Noise in data, sources of noise, types of Data Noise and ways to minimize

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

MODULE 3

Basic Statistical Concepts

You will learn about the 'Basic Statistical Concepts' in this module.

Module Learning Objectives

At the end of this module, you will be able to:

- Explain the basic concepts of statistics along with its classification: descriptive and inferential statistics.
- Understand the measures of scales: nominal, ordinal, interval and ratio scales.
- Describe the measures of location: mean, median and mode.
- Enumerate the measures of variability: range, quartiles and interquartile range, variance and standard deviation.
- Explain the measures of shape: skewness and kurtosis.

Module Topics

The following topics that will be covered in the module:

1. Introduction to Statistics
2. Scale of Measurements (Nominal, Ordinal, Ratio and Interval)
3. Measures of Location
4. Measures of Variability/Spread
5. Measures of Shape

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1.2 Classification of Statistical Methods

The following illustration depicts the two major categories of statistical analysis.

We just learnt about the introduction to Statistics. Let's now learn about the classification of statistical methods. As depicted in the illustration above, there are two major categories of statistical methods. Each one is important for specific purposes. Some of the measures may be similar, but the goal and methodologies differ. Two major categories of statistical analysis are:

1. **Descriptive Statistics:** As the name suggests, the purpose of descriptive statistics is to collect, organize, summarize and describe data. It deals with a larger chunk of data and describes it with the help of summary charts and tables. The goal is to describe data and not derive conclusions out of it. Descriptive statistics is used to describe the characteristics of a sample set.

2. **Inferential Statistics:** Inferential statistics applies complex mathematical calculations to derive an inference and make decisions about a large population, based on the study of a smaller sample. Inferential statistics is used for making predictions.

1.3 Descriptive Statistics

Descriptive Statistics:

- Is a set of methods to describe the collected data.
- Aims to describe the characteristics of a sample data set and understanding the specific set of observations.
- Summarizes and represents data using graphs, charts and tables.
- Involves following methods:
  - Measurement of central tendency
  - Measurement of dispersion/spread
  - Measurement of skewness and kurtosis
  - Exploration of relationships of paired data

Descriptive statistics is simple and straightforward. It describes a data set. For example, you will use descriptive statistics, if you want to study about a group of people and their shopping patterns. You will simply collect the details about their shopping behaviour, days on which they shop more, their items of interest, etc. You will summarize the data and present it using graphs and charts, which will

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

give a picture of the group's properties. The sample might be a representative of a larger population, but descriptive statistics do not intend to draw conclusions or make predictions about the larger population using the sample data.

Some more scenarios where descriptive statistics is applied: the portion of salary that is spent on vegetables each month; batting average of MS Dhoni in the last five years; the number of seats that a political party won during the last election.

There is no uncertainty associated with descriptive statistics, as we describe only the sample or population that we actually measure and there's no assumption. In this process, we actually take a considerably large number of data points and reduce them using summary values and graphs. The major advantage of this method over excel sheets and tables of data is that it allows one to gain more insights and visualize the data patterns.

Let's now have a look at the methods that are used in descriptive statistics.

- **Measurement of the central tendency:** Central tendency refers to the measure of finding out where most of the values in a given data set fall, i.e., the general trends that prevail among the values. Mean (average), Median (mid-value) and Mode (most commonly occurring value) are referred to as the measures of central tendency.
- **Measurement of dispersion/spread:** Measures of dispersion is the way of describing how the data is distributed and relate to each other. It helps us in finding out how far the data extends from the centre. If the dispersion is low, it indicates that the data is tightly clustered around the centre value and a higher dispersion value indicates scattered data. Using a graph, frequency distribution can be plotted. Common measures of dispersion/spread include:
    - Range - the entire range of values present in a data set.
    - Frequency distribution - frequency of occurrence of a particular value in a data set.
    - Quartiles - subgroups formed within a data set, after the values are divided into four equal parts across the range.
    - Variance - the amount of spread that exists among the data.
    - Standard deviation - the spread of data in accordance with the mean.
- **Measurement of skewness and kurtosis:** This measure tells about the symmetric nature of values in a given data set. If data is asymmetric, it will be skewed to the right or left of the central value. Similarly, Kurtosis tells us about the shape of the tails of the distribution on the far left or far right.
- **Exploration of relationships of paired data:** Relationships of data can be explored by measuring the correlation or by using scatterplots. Correlation refers to the change in one of the variables in a particular direction, in response to the change in the other variable. A correlation coefficient is the measure of the direction and strength of this tendency to vary together.

In the forthcoming sections and modules, we'll see each of these in detail.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1.4 Inferential Statistics

Inferential Statistics:

- Is a set of methods used to make a generalization, estimate, prediction or decision.
- Aims to test a hypothesis and derive conclusions about a population based on the sample.
- Analysis results are generalized from a sample to a larger population.
- Includes the following analysis tools:
  - Hypothesis tests
  - Confidence Intervals
  - Regression analysis

In inferential statistics, you draw a conclusion, i.e., an inference about a larger population by analysing a sample data set taken from the population. Once you derive a conclusion from the sample, you will then generalize it to the population. There needs to be the confidence that the sample accurately reflects the population.

Let's again take the example of people's shopping patterns. You have taken data of 100 people from a group of 1000. Around 50 of them buy chocolates a lot in April in a given year. With this data, you predict that 50% of all the population will buy more chocolates in April in the coming year. There's an uncertainty in this prediction, as people's buying behavior might change.

Let's look at the other examples given the previous slide as well: with the current month's spend on vegetables, you can estimate the amount that will be spent on vegetables the whole year; you can predict the result of a match in accordance with MS Dhoni's batting average; with previous election results you can try to predict the number of seats a party will win in the upcoming elections. Again, all these are associated with uncertainties.

In inferential statistics, random sampling is done, instead of picking up the group of choice, to have more confidence that the sample represents the population. When we have a small data set as the sample from a large population, the mean of the sample might not reflect the actual mean of the population. This difference between the sample statistics and the population statistics is referred to as 'sampling error'. Inferential statistics incorporates the estimates of these sampling errors into the results.

Most common tools used in inferential statistics are:

- **Hypothesis tests:** Consider a clinical trial where a drug is tested in a specific group of people and controls. Using Hypothesis tests, we can find out whether the drug will have the similar effect on the entire population. Hypothesis tests are thus used to make a claim about the population by analyzing the sample data set.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- **Confidence intervals:** Confidence intervals give the range of values for an unknown parameter, like mean or standard deviation, of the population by measuring the sample. This is expressed in terms of an interval and the degree of confidence that the parameter is, within the interval. Confidence intervals incorporate the sample errors and uncertainties to create the range within which the population values are likely to fall.
- **Regression analysis:** Regression analysis is used to describe a relationship between a set of independent variables and a dependent variable. Regression analysis also makes use of hypothesis tests to find if the relationship identified in the sample data exists in the population as well.

All these tools will be explained in detail in upcoming sections/modules. Some of the other techniques that statisticians use to explore the relationship among data, thereby to create inferential statistics are: linear regression analyses, logistic regression analyses, ANOVA, correlation analysis, structural equation modeling, and survival analysis. Common tests of significance include the chi-square and t-test. These methods tell the statistician if the results of the sample analysis represent the population.

We've now covered the introduction to statistics, where we learnt about the classification and the measures involved. We're now going to learn about the various measures of statistics.

What did You Grasp?

1. State True or False.
   Measurement of skewness is part of descriptive statistics.

   A) True
   B) False

2. Select all that apply.
   Which of the following options is/are true about inferential statistics?

   A) Inferential statistics involves random sampling
   B) Inferential statistics is used to describe a data set
   C) Inferential statistics is associated with uncertainties
   D) Measure of central tendency belongs to inferential statistics

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

2 Scale of Measurements (Nominal, Ordinal, Ratio and Interval)

2.1 Introduction to Measurement Scales

Scales of measurement:

- Ways in which variables or numbers are defined and categorized.
- Properties of measurement scales determine their appropriateness for use of analyses. These include:
  - Identity
  - Magnitude
  - Equal intervals
  - Absolute zero

Before doing any statistical procedure, we need to measure something. A scale gives the unit of measurement like categories, inches, seconds, etc. and the range of possible values. The scale affects the way in which data values and statistics using those data values are interpreted. It does not change the way that the data values can be analyzed statistically (although some of those statistical analyses would make no sense when you try to interpret them).

The unit can be either continuous (allows partial units, or decimal, values) or discrete (only whole units), but the value of any particular unit has the same meaning for all objects that are assigned that value on the scale. A scale can have an infinite range of possible values, i.e., a measure can take on any real-number value, or may be limited to as few as two possible values. If there is only one possible value, it is a constant, and no need for a scale.

Measurement scales can have the following properties.

- Identity - Each value on the measurement scale has a unique meaning.
- Magnitude - The ability to know if the ability to know if one score is greater than, less than, or equal to another score. Values on the measurement scale have an ordered relationship to one another.
- Equal Intervals refers to the magnitude that is represented by a unit of measurement that is the same no matter where it falls on the scale. Units along the scale are equal to one another.
- Absolute Zero refers to the scale that has a point equal to zero or a point where there is no score. Below absolute zero, no value exists.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

2.2 Nominal Scales

Nominal scales:

- The simplest scale of measurement, having only the identity property.
- Not a measure of quantity, but measures the quality, i.e., identity and difference.
- Indicates a category - A categorical variable, called nominal variable - for mutually exclusive categories.
- In a nominal scale, numbers identify and classify objects.

The nominal scale of measurement only satisfies the identity property of measurement. Values assigned to variables represent a descriptive category but have no inherent numerical value with respect to magnitude. A nominal scale uses categories that have no particular order. The units of a nominal scale are categories that do not overlap; an object cannot be in more than one category. Automobile make (Ford, Chevy, Honda, Toyota, etc.), teams (Cougars, Owls, Tigers, etc.), and gender (female, male) are nominal scale variables.

Gender is an example of a variable that is measured on a nominal scale. Individuals may be classified as "male" or "female", but neither value represents more or less "gender" than the other. Religion and political affiliation are other examples of variables that are normally measured on a nominal scale.

2.3 Ordinal Scales

Ordinal scale:

- Refers to order in measurement, indicating the direction, in addition to providing nominal information.
- Labels represent an order that indicates either preference or ranking.
- Numbers indicate the relative positions of objects, but not the magnitude of differences between them.
- Has both identity and magnitude properties.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

The ordinal scale has the property of both identity and magnitude. Each value on the ordinal scale has a unique meaning, and it has an ordered relationship to every other value on the scale. An ordinal scale uses categories that have a particular order. The units of an ordinal scale are categories that do not overlap and defined such that a higher category has more of some property and characteristic than the preceding category.

An example of an ordinal scale in action would be the results of a horse race, reported as "win", "place", and "show". We know the rank order in which horses finished the race. The horse that won finished ahead of the horse that placed, and the horse that placed finished ahead of the horse that showed. However, we cannot tell from this ordinal scale whether it was a close race or whether the winning horse won by a mile.

Class year (freshman, sophomore, junior, senior), game innings or quarters (1st, 2nd, 3rd, etc.), and meal (breakfast, brunch, lunch, hors-d'oeuvres, salad, soup, meat, dessert, etc.) are ordinal scale variables.

Nominal and ordinal scales involve discrete variables.

2.4 Interval Scales

Interval scales:

- Have identity, magnitude and equal intervals.
- Magnitude between the consecutive intervals are equal.
- Scale that represents quantity and has equal units.
- Do not have a true zero, i.e., zero is just an additional point of measurement in the scale.

The interval scale of measurement has the properties of identity, magnitude, and equal intervals. The unit of measurement in an interval is equal across the whole range of possible values. There is no overlapping between the units and they relate an order to the property being measured. As mentioned above, an interval scale does not have a "true zero" value. The value of zero does not mean that there is no characteristic; zero is just a number used to describe a position on the scale. Negative values are also allowed by interval variables.

A perfect example of an interval scale is the Fahrenheit scale to measure temperature. The scale is made up of equal temperature units so that the difference between 40 and 50 degrees Fahrenheit is equal to the difference between 50 and 60 degrees Fahrenheit.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

With an interval scale, you know not only whether different values are bigger or smaller, you also know how much bigger or smaller they are. For example, suppose it is 60 degrees Fahrenheit on Monday and 70 degrees on Tuesday. You know not only that it was hotter on Tuesday, you also know that it was 10 degrees hotter.

Other examples of interval scales include attitude scales, calendar years.

2.5 Ratio Scales

Ratio scale:

- The most sophisticated scale of measurement.
- Has all the four properties, identity, magnitude, equal intervals and absolute zero.
- Has a consistent unit of measurement.
- True zero value indicates absence of a characteristic.
- Measure of the magnitude of the data value using equal intervals between values and absolute zero.

Ratio scales are the most sophisticated scales of measurement and possess all the four properties of real numbers, like identity, magnitude, equal intervals and absolute zero. Ratio scales have an origin that makes them different from the other three scales of measurement. Ratio scores have an infinite range and thus are continuous.

A ratio scale has a consistent unit of measurement. The true zero value indicates an absence of a characteristic. The units do not overlap and relate an order to the property being measured. Ratio scales also allow comparisons between values, like twice as high as, or one-half of, etc.

Some of the examples for ratio scales are mass, length, height, time, distance, energy, and many laboratory values. For example, the unit of energy (calories) has the same meaning at all parts of the scale. A zero value here indicates the absence of energy. Negative values do not have any meaning unless the definition of the measure is changed. For example, negative calories indicate an energy deficit rather than energy gain.

Interval and Ratio scales involve continuous variables.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

What did You Grasp?

1. Select all that apply.
   In a centigrade scale for temperature measurement, which of the following properties is/are satisfied?

   A) Identity
   B) Magnitude
   C) Equal intervals
   D) Absolute zero

2. State True or False.
   An ordinal scale exhibits only identity and magnitude.

   B) True
   C) False

3 Measures of Location

3.1 Measures of Location

Measures of Location:

- One of the tools covered under descriptive statistics.
- Measure of Location summarizes the data set by giving the range of the data values that describes its location relative to the entire data set arranged according to the magnitude.
- The measure of central tendency is the commonly used method to identify the centre of the data set.
- Used to describe a data set by identifying the central position within that data set.
- Measures of central tendency:
  - Mean
  - Median
  - Mode

Measure of location comes under descriptive statistics, which is also referred to as summary statistics. Location measures summarize the data and find one value that is typical for the data set, i.e., a single value that is a representative of the entire data set. Thus the measure of location is useful where it is not possible to list all the data, instead, a representative data value is used to describe the complete data set.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Most commonly applied measure of location is the measure of central tendency, i.e., identifying the centre of the data set, around which all other values are clustered. Measures of central tendency include: Mean, Median and Mode. We'll see each one of these in detail in the upcoming slides.

Other measures of location include:

- Minimum - the smallest value in the data set.
- Maximum - the largest value in a data set.

3.2 Mean

Mean:

- The most popular measure of central tendency.
- Also referred to as the average value in the data set.
- Defined as the sum of all values in a data set, divided by the total number of values in a data set.
- Example: Let x1, x2, x3, ……. Xn be the values in the data set, and there are 'n' number of values in the data set:

  Mean =

$$\frac{x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n} \quad \text{or} \quad \frac{\sum x}{n}$$

You might have heard the term average countless number of times. What does it refer to? In statistics, we use the term mean in place of average. Mean is equal to the sum of all values in a data set divided by the total number of values in the data set. The Greek letter 'mu' μ is used to represent mean.

The formula for calculating mean is given in the slide.

Let's take an example now: the temperature in Mumbai on five consecutive days is 35°C, 38°C, 40°C, 34°C and 38°C. The mean temperature of these five days is the sum of the temperatures on all five days divided by 5 i.e., 35+38+40+34+38/5, 187/5, which is equal to 37°C.

You might notice that the mean value is not one of the values in the data set. But mean is important in statistical analyses, at it minimises errors in the prediction of any value in the given data set. Also, mean is the only measure of central tendency, where the sum of the deviations of each value from the mean is always zero.

For mean, the distance between two numbers is defined to be the square of their difference. The sum of the squares of the differences between the data and the mean is smaller than the sum of squares

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

of the differences between the data and any other number. Similarly, the root mean square (rms) of the differences from the mean is smaller than the rms of the list of differences from any other number

Mean has a disadvantage, i.e., mean value if affected by the influence of outliers or out of range values. For example, if the distribution of data is not uniform and has extremely low or high numerical values compared to the mean. In this case, mean might not reflect the actual characteristic of the data set.

3.3 Median

Median:

- Middle value in a data set, after it is arranged in the order of magnitude.
- Values are first arranged from the smallest to the largest.
- Median is the middle value of the data set if the total number of values is an odd number; average of the middle two values, if the total number of values is an even number.
- Median = [(n+1)/2]th value, if there are odd number of values.
- Median = average of the [n/2]th and the [(n/2)+1]th values, if there are an even number of values.

Median is defined as the middle point of the data set after it is organized in the order of magnitude. It is estimated by first ordering the data from smallest to largest and then counting upwards for half the observations. The estimate of the median is either the value at the centre of the ordering in the case of an odd number of values, or the simple average of the middle two values if the total number of values is even.

To be specific, if there are an odd number of observations, median is the [(n+1)/2]th value, and if there are an even number of values, it is the average of the [n/2]th and the [(n/2)+1]th values.

Example:

Suppose we have the data below:

65, 55, 89, 56, 35, 14, 56, 55, 87, 45, 92

Rearrange that data into order of magnitude (smallest first):

14, 35, 45, 55, 55, **56**, 56, 65, 87, 89, 92

Median is the middle value - in this case, 56 (highlighted in bold). It is the middle value because there are 5 values before it and 5 values after it.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Now, let's imagine if there are only 10 values? Take the middle two scores and average the result. So, if we look at the example below:

65, 55, 89, 56, 35, 14, 56, 55, 87, 45

Rearrange that data into order of magnitude (smallest first):

14, 35, 45, 55, 55, 56, 56, 65, 87, 89

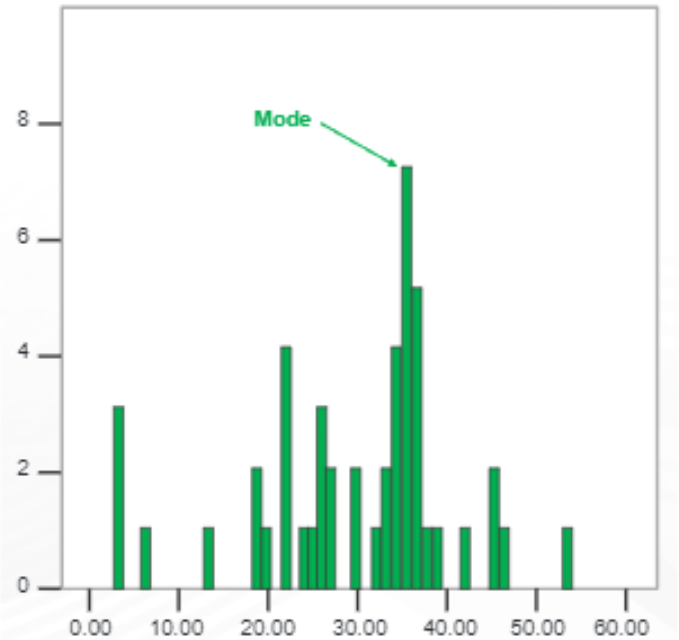
Only now we have to take the 5th and 6th score in our data set and then find the mean of the two (use the formula in the previous slide) to get a median of 55.5.

For median, the distance between two numbers is defined to be the absolute value of their difference. That is, the sum of the absolute values of the differences between a median and the data is no larger than the sum of the absolute values of the differences between any other number and the data.

3.4 Mode

Mode:

- Most frequently occurring value in the data set, i.e. value of the random sample that occurs with the greatest frequency.
- Used in a qualitative fashion.
- Applicable to all levels of data measurement, i.e., nominal, ordinal, interval and ratio scales.
- Bimodal - if values occur most frequently, two modes.
- Multimodal - Data sets with more than two modes.



The mode of a set of data is the most common value among the data. In other words, mode is the most frequently occurring value in a data set. If the data are grouped, the grouping with the highest frequency is the mode. It is rare that several data coincide exactly, unless the variable is discrete, or the measurements are reported with low precision.

Mode is not used much in statistical analysis since it depends on the accuracy with which the data are measured; although it may be useful for categorical data to describe the most frequent category. The expression 'bimodal' distribution is used to describe a distribution with two peaks in it and multimodal has more than two modes in it. This can be caused by mixing populations. For example, height might appear bimodal if one had men and women on the population.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Looking at the temperature example from Mean calculation: 35°C, 38°C, 40°C, 34°C and 38°C. Mode of this data set is 38°C since it occurs more number of times than the other values.

For mode, the distance between two numbers is defined to be zero if the numbers are equal, and one if they are not equal. That is, the number of data that differ from a mode is no larger than the number of data that differ from any other value. Equivalently, a mode is a number from which the fewest possible data differ: a "most common" value.

What did You Grasp?

1. Calculate the mean of the following data set - 60, 56, 61, 68, 51, 53, 69, 54.

   A) 58
   B) 56
   C) 59
   D) 61

2. Find the median of the given data set - 6, 5, 4, 1, 7, 3

   A) 3
   B) 4.5
   C) 7.5
   D) 5

4 Measures of Variability/Spread

4.1 Measures of Variability/Spread

Measures of Spread:

- A measure of how the data is dispersed or spread around the mean, i.e. the average.
- Is used in quantitative data, as the variables can be arranged in a logical order, with a low and high value.
- Measures of spread:
  - Range
  - Quartiles and Interquartile range
  - Variance
  - Standard deviation

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Measures of spread summarise the data in a manner that described the scatteredness of the values and how they differ from the mean value. Measures of variability need to have the following properties.

- The measure should be proportional to the spread of the data (small when the data are clustered together, and large when the data are widely spread).
- The measure should be independent of the number of values in the data set (otherwise, simply by taking more measurements the value would increase even if the scatter of the measurements was not increasing).
- The measure should be independent of the mean.

4.2 Range

Range:

- The simplest measure of variability.
- The distance between the smallest value and the largest value in a dataset.
- Represents the width of the smallest interval that contains all the data.
- The range of a list is calculating the difference between the largest value and the smallest value.
- Affected by outliers, as variance may either be too low or too high because of outliers.

Range is defined as the distance between the smallest value to the largest value. Range of a data set is calculated by finding the difference between the largest and the smallest values in the data set.

For example, the range of the following data set: 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8 is **4**, the difference between the highest value (8) and the lowest value (4).

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

4.3 Quartiles and Interquartile Range

Quartiles range:

- Divide an ordered dataset into four equal parts, and refer to the values of the point between the quarters:
  - The lower quartile (Q1) is the point between the lowest 25% of values and the highest 75% of values. It is also called the 25th percentile.
  - The second quartile (Q2) is the middle of the data set. It is also called the 50th percentile, or the median.
  - The upper quartile (Q3) is the point between the lowest 75% and highest 25% of values. It is also called the 75th percentile.
  - The interquartile range (IQR) is the difference between the upper (Q3) and lower (Q1) quartiles and describes the middle 50% of values when ordered from lowest to highest.

The Interquartile Range (IQR) is the difference between the upper quartile (75th percentile) and the lower quartile (25th percentile). It is the width of the interval that contains the middle 50% of the data. It is insensitive to the most extreme values of the data (assuming that there are more than four data). The IQR is resistant: changing just one value has a limited effect on it. Neither the range nor the IQR is a range of numbers, despite their names—each is a single number.

A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts) or percentiles (hundred equal parts).

For example, consider the data set: 4, 5, 5, Q1 5, 6, 6 Q2 6, 6, 7 Q3 7, 7, 8. The quartiles are as follows:

Q1 = (5+5) / 2 = 5

Q2 = (6+6) / 2 = 6

Q3 = (7+7) / 2 = 7

The IQR for Dataset A is = 2

IQR = Q3 - Q1 = 2

4.4 Variance

Variance:

- Variance is a measure of how the spread-out a data set is.
- Calculated as the average squared deviation of each number from the mean of a data set.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- The formula for calculating variance (S2) of a data set is:

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n}$$

Variance is a measure of how the spread-out a data set is. It is calculated as the average squared deviation of each number from the mean of a data set.

Use this step-by-step approach to find the variance of a data set.

- Calculate the mean.
- Subtract the mean from each value.
- Square each of the resulting values.
- Add these squared results together.
- Divide this total by the number of values (variance, S2).

4.5 Standard Deviation (SD)

SD:

- The mean distance of all values from the overall mean.
- The square root of variance.
- SD(s) of a sample data set can be calculated using the formula:

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

(where xi is the individual value, x is the mean of the sample set, n is the number of values in the distribution.)

In datasets with a small spread, all values are very close to the mean, resulting in a small variance and standard deviation. Where a dataset is more dispersed, values are spread further away from the mean, leading to a larger variance and standard deviation.

The smaller the variance and standard deviation, the more the mean value is indicative of the whole dataset. Therefore, if all values of a dataset are the same, the standard deviation and variance are zero. Because of its close links with the mean, standard deviation can be greatly affected if the mean gives a poor measure of central tendency.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Standard deviation is also influenced by outliers one value could contribute largely to the results of the standard deviation. In that sense, the standard deviation is a good indicator of the presence of outliers. This makes standard deviation a very useful measure of spread for symmetrical distributions with no outliers. The units of the SD are the same as the original units of measurement. For example, if the list is comprised of measurements of heights in inches, the SD has units of inches.

What did You Grasp?

1. State True or False.

   The range of a data set is not affected by outliers.

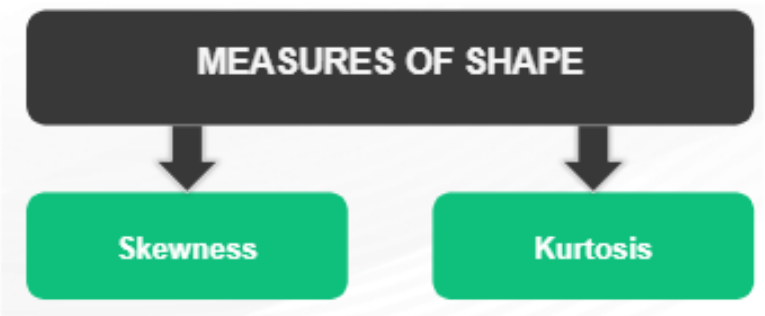
   A) True
   B) False

5 Measures of Shape

5.1 Measures of Shape

Measures of shape:

- Describe the pattern of distribution of values within a data set.
- Data distribution may be:
  - Symmetric] - two sides of the distribution are a mirror image of each other. e.g.. Normal distribution
  - Skewed - asymmetrical distribution where the values tend to be more frequent around the high or low ends of the x-axis.



This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Measures of shape help us in identifying the pattern of data distribution within a data set. Distribution of values in a data set may be symmetrical or asymmetrical. Common examples of symmetry and asymmetry are the 'normal distribution' and the 'skewed distribution', respectively.

The distribution shape of quantitative data can be described as there is a logical order to the values, and the 'low' and 'high' end values on the x-axis of the histogram are able to be identified. The distribution shape of a qualitative data cannot be described as the data are not numeric.

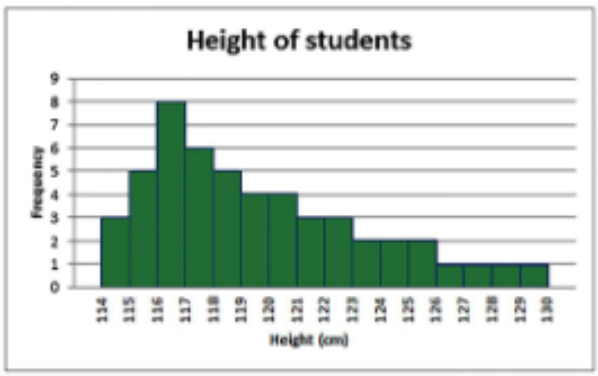
As mentioned in the slide, in a symmetrical distribution, the two sides of a distribution are mirror images of each other. Example for a truly symmetrical distribution is a normal distribution. When a histogram is constructed on values that are normally distributed, the shape of columns form a symmetrical bell shape. This is why this distribution is also known as a 'normal curve' or 'bell curve'.

In an asymmetrical distribution, the two sides will not be mirror images of each other. A skewed distribution can be positively or negatively skewed. We'll see about skewed distributions in the next slide.

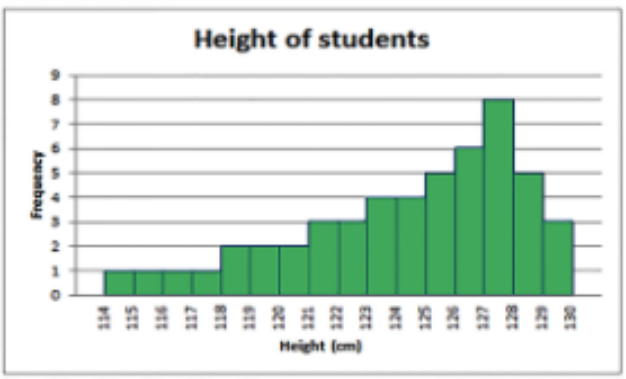
5.2 Skewness

Skewness:

- The amount and direction of skew, i.e. deviation from the horizontal symmetry.
- Two types of skewed distribution:



Positively skewed distribution



Negatively skewed distribution

Skewness refers to the tendency of the values in a data set to be clustered around the high or low ends of the axis. By constructing a histogram and looking at the shape of the distribution, skewness can be identified.

A distribution is said to be positively skewed when the tail on the right side of the histogram is longer than the left side. Most of the values tend to cluster toward the left side of the x-axis (i.e. the smaller values) with increasingly fewer values at the right side of the x-axis (i.e. the larger values).

A distribution is said to be negatively skewed when the tail on the left side of the histogram is longer than the right side. Most of the values tend to cluster toward the right side of the x-axis (i.e. the larger values), with increasingly lesser values on the left side of the x-axis (i.e. the smaller values).

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Key features of the skewed distribution:

- asymmetrical shape
- mean and median have different values and do not lie at the centre of the curve
- there can be more than one mode
- the distribution of the data tends towards the high or low end of the dataset

5.3 Kurtosis

Kurtosis:

- Measure of how tall and sharp the central peak is, relative to a standard bell curve.
- Also called as the measure of tailedness of a data distribution.
- Outliers impact the kurtosis of a data distribution.
- Formula to calculate kurtosis:

$$\frac{\sum_{i=1}^{N} \frac{\left(X_i - \bar{X}\right)}{N}}{s^4}$$

Where Xi is the individual value, X is the mean, N is the total number of values and s is the standard deviation

Kurtosis is defined as the measure of flatness or peakedness of the given distribution for the random variable along its tail. The outliers in the given data have more effect on this measure. Kurtosis does not have any unit.

Based on the value of kurtosis, the distribution can be classified into three categories.

- The distribution with kurtosis equal to 3 is known as mesokurtic. A random variable which follows normal distribution has kurtosis 3.
- If the kurtosis is less than three, the distribution is called as platykurtic. Here, the distribution has shorter and thinner tails than the normal distribution. Moreover, the peak is lower and also broader when compared to the normal distribution.
- If the kurtosis is greater than three, the distribution is called as leptokurtic. Here, the distribution has longer and fatter tails than the normal distribution. Moreover, the peak is higher and also sharper when compared to the normal distribution.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

What did You Grasp?

1. Which of the following options is/are true about a skewed distribution?

   A) Skewed distribution is a symmetrical distribution
   B) Skewed distribution can have more than one node.
   C) A positively skewed distribution has its right tail longer than the left side
   D) Data distribution tends towards the centre of the data set.

In a nutshell, we learnt:

1. An introduction to statistics along with its classification: descriptive and inferential statistics.
2. The four measures of scales: nominal, ordinal, interval and ratio scales.
3. Measures of location: mean, median and mode.
4. Measures of variability: range, quartiles and interquartile range, variance and standard deviation.
5. Measures of shape: skewness and kurtosis.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1 Introduction to Statistics

1.1 Introduction to Statistics

Statistics:

- The branch of mathematics that deals with the collection, analysis, interpretation and presentation of masses of numerical data.
- The discipline that utilizes data samples to support claims about populations (a larger data set).
- Utilizes models, representations and synopses for a set of experimental data and applies the results to a larger population.

Think about the following questions:

- What is the population of India?
- What is the literacy rate in India?
- Which state in India excels in healthcare?
- Which state in India has the largest number of automobile manufacturing units?
- Which Indian state recorded most number of accidents in a given year?

We come across a lot of questions like this on a day to day basis. There exist answers to all these questions. And all this is statistics, but it is not limited to data like this. In most of the cases like this, data from the entire population is not necessary. A portion of the population, that is, a sample is enough to give insights about the population. Imagine you want to find the average salary of an entry-level software developer in India. Will you go and meet each one of them in person to get the data? You will send a survey to a set of people and analyse it to get an average number. You will then apply it to the entire population. So, what is statistics?

As given in the slide, "Statistics is that branch of mathematics, which deals with the collection, analysis, interpretation and presentation of masses of numerical data." Statistics takes data from samples, perform interpretations and applies the results to populations, i.e., a larger data set. The sample is thus a representative of a population. In a nutshell, statistics is a process that is used to characterize a data set. It involves the process of gathering and evaluating data and summarizing the data into a mathematical form.

Statistics has applications in almost every discipline like education, healthcare, manufacturing, government sectors IT, Physical and Social sciences, Psychology, Humanities and so forth.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

MODULE 4

Probability Theory

You will learn about the 'Probability Theory' in this module.

Module Learning Objectives

At the end of this module, you will be able to:

- Understand the principles of counting including addition rule, product rule, permutation and combination.
- Understand the introduction and important definitions of Probability Theory.
- Describe conditional probability.
- Explain Bayes Theorem and its applications.
- Explain different methods of discrete probability distribution like discrete uniform distribution, Poisson distribution, Bernoulli distribution and Binomial distribution.
- Understand the concepts of Covariance and Correlation.
- Understand the different methods of Continuous Probability Distribution.
- Explain the concept of Central Limit Theorem.
- Understand Hypothesis Testing and the types of errors.

Module Topics

The following topics that will be covered in the module:

1. Principles of Counting
2. Introduction and Definitions of Probability Theory
3. Conditional Probability

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- **Multiplication rule:**

  The multiplication rule states that 'If E1 is an experiment with n1 outcomes and E2 is an experiment with n2 possible outcomes, and E1 and E2 are independent events, then the experiment which consists of performing E1 first and then E2 consists of **n1*n2** possible outcomes.

  For example, tossing a coin twice, will produce the following number of outcomes:

  For E1, n1 = 2; for E2, n2 = 2; n1*n2 = 4

- **Permutation:**

  The number of ways to choose a sample of r elements from a set of n distinct objects where order does matter and replacements are not allowed. When n = r this reduces to n!, a simple factorial of n. Permutation is computed using the formula, nPr = n! divided by (n-r)!, i.e., **n!/(n-r)!**

- **Combination:**

  The number of ways to choose a sample of r elements from a set of n distinct objects where order doesn't matter and replacements are not allowed. Combination is computed using the formula, nCr = n! divided by r!*(n-r)!, i.e., **n!/r!*(n-r)!**

1.1 Permutations

Permutation: An arrangement or order of a set of objects, and the order if arrangement is important. There are four theorems of Permutation.

- **Theorem 1:** Arranging n Objects: This theorem states that n objects can be arranged in n! Ways.
- **Theorem 2:** Number of Permutations: The number of permutations of n distinct objects taken 'r' at a time, denoted by nPr, where repetitions are not allowed, is given by:

nPr = $n(n-1)(n-2)\ldots(n-r+1) = \frac{n!}{(n-r)!}$

Notation of permutations can be: P_r^n or ${}^nP_r$

Recall the factorial notations that you have learned in school mathematics. n!, which is read as n-factorial, is defined as:

n! = n x (n-1) x (n-2) x ….. 3 x 2 x 1.

The theorems in Permutation are based on factorial notation.

Example for theorem 1: In how many ways can 3 persons be seated?

Since there are 3 persons, the number of ways is 3!.

From the factorial notation, 3! = 3 x 2 x 1 = **6 ways**

Theorem 2: Note that nPn = n!, since 0! = 1.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Example for theorem 2: In how many ways can 5 boxes be arranged in 3 spaces in a shelf?

By theorem 2, 5P3 = 5!/(5-3)! = 5!/2! = **60**

Theorem 3: Permutations of different kinds of objects: The number of different permutations of n objects of which n1 are of one kind, n2 are of a second kind, ... nk are of a k-th kind is:

$$\frac{n!}{n_1! \times n_2! \times n_3 \times \ldots \times n_k!}$$

Theorem 4: Arranging objects in a circle: There are (n−1)! ways to arrange n distinct objects in a circle (where the clockwise and anti-clockwise arrangements are regarded as distinct.)

Example for theorem 3: In how many ways can the seven letters of the word "arrange" be arranged in a row?

There are two 'a's, two 'r's, one n, one g and one e.

Based on theorem 3, the number of ways of arrangement is: 7!/(2!)(2!)(1!)(1!)(1!) = 5040/(2)(2)(1)(1)(1) = 5040/4 = **1260 ways**.

Example for theorem 4:

In how many ways can 5 people be arranged in a circle?

Based on theorem, (5-1)! = 4! = **24 ways**.

1.2 Combinations

Combinations:

- Combinations are for unordered selections.
- A combination of n objects taken r at a time is a selection which does not take the arrangement of the objects into account, i.e the order of arrangement is not important.
- The number of ways (or combinations) in which r objects can be selected from a set of n objects, where repetition is not allowed, is denoted by:

$$C_r^n = \frac{n!}{r!(n-r)!} \quad \text{Or} \quad C_r^n = \frac{P_r^n}{r!}$$

The ways of combination can be computed even if the objects are unordered. Some points to note are:

- nC0 = 1
- nCn = 1
- nCr = nCn-r

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Example: Find the number of ways in which 3 components can be selected from a batch of 20 different components.

20C3 = 20!/(3!)(20-3)! = 20!/(3!)(17!) = **1140 ways**.

What did You Grasp?

1. In how many ways can a group of 4 boys be selected from 10 if the eldest boy is excluded?

   A) 84
   B) 126
   C) 210
   D) 44

2. Introduction to Probability Theory

Chance, Likely, Possibly...etc,

What do these words indicate?

These words indicate 'Lack of certainty'.

- There is no perfect yardstick to measure the lack of certainty or simply 'uncertainty'.
- But, it can be measured mathematically on the basis of some assumptions. This numerical measure is referred to as "PROBABILITY".

Will it rain today? It may.

Will India win the match? There's a chance.

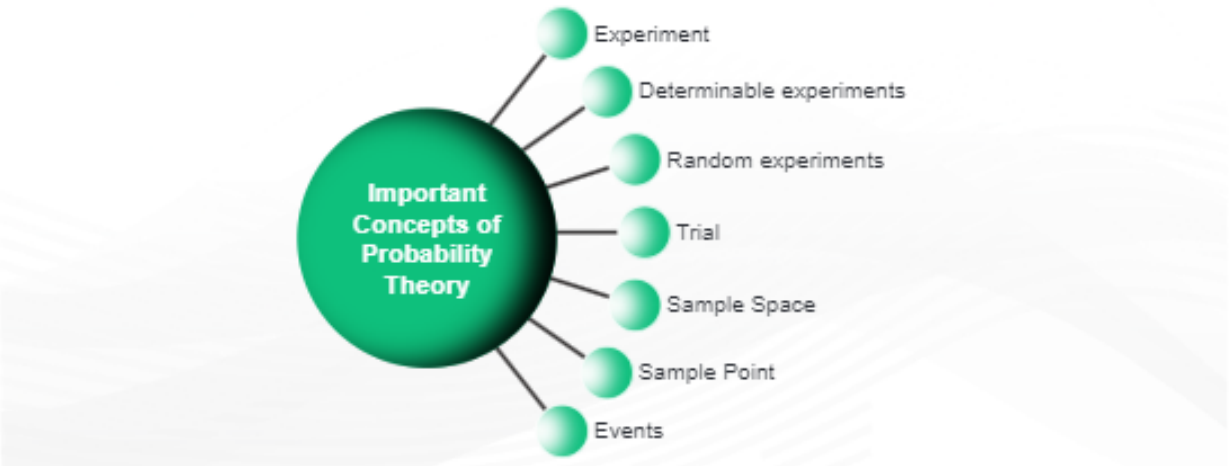
We come across the words may be, possibly, likely etc. on a daily basis. Things may happen or may not. There is always an uncertainty associated with these terms. Is there a way to measure uncertainty? There is no perfect way to measure uncertainty, but mathematically, based on assumptions, uncertainty can be measured. Here comes the probability theory. You can use the terms chance, odds, uncertainty, prevalence, risk, expectancy etc., to interpret probability.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

In 1929, Bertrand Russell said during one of his lectures, ' Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means." Probability theory is used to evaluate the reliability of conclusions and inferences based on data, thus it is fundamental to statistics.

2.1 Key Concepts and Definitions of Probability Theory

The important concepts of probability theory:



Important concepts of probability are listed below:

- **Experiment:** There are two categories of experiments
  - **Determinable experiments** - Experiments whose outcome may be predictable. For example, if we heat water, we know for sure that it will boil and change into vapour at some point in time.
  - **Random experiments** - Experiments whose outcomes are not predictable, i.e., a non-deterministic model. If in each trial of an experiment conducted under identical conditions, the outcome is not only always the same but may be any of the possible outcomes, such experiment is termed as a Random experiment. Examples of such events are - tossing a coin, rolling a die, selecting a card from a well-shuffled deck of 52 cards, etc.
- **Trial:** A trial is an action which results in one or several outcomes. Example: Tossing a coin, rolling a die.
- **Sample Space:** The set of all outcomes of a random experiment is called sample space. Denoted by 'S".

  For rolling a dice, the sample space is, S = {1, 2, 3, 4, 5, 6}

  For tossing a coin, S = {H, T}

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- **Sample Point:** Sample point is defined as each individual outcome of a random experiment.

  Flipping a coin: {H}, {T}

  Rolling a dice: {1}, {2}, {3}, {4}, {5}, {6}

- **Events:** An event is defined as any subset of the sample space. For example, the events in rolling a die can include: getting an even number, getting a prime number, etc.

2.2 Independent Events

Independent Events:

- If two events A and B are said to be "Independent" P(AB) = P(A)*P(B)
- Note 1: If A and B are independent, then,
  - A' and B' are independent.
  - A and B' are independent.
  - A' and B are independent.
- Here, A' denotes A complement, B' denotes B complement.
- *Note 2: Mutually Exclusive events are certainly not independent.*

The interpretation of two events being independent is that occurrence of one event does not affect the probability of the other event, i.e., P(A|B) = P(A). P(B|A) means "the probability of the event B given that A has already occurred" (you will learn about this in conditional probability).

Intersections of Independent Events

The probability of the intersection of a series of independent events A1, A2, A3,..An is simply given by

P(A1 ∩ A2 …… An) = P(A1) P(A2) ……. P(An).

Example:

Problem: A fair die is tossed twice. Find the probability of getting a 4 or 5 on the first toss and a 1, 2, or 3 in the second toss.

Solution:

P(A1) = P(4 U 5) = 2/6 = ⅓

P(A2) = P(1 U 2 U 3) = 3/6 = ½

Since these are independent events,

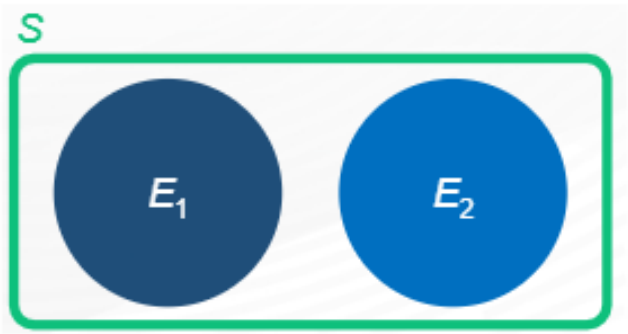
P(A1 ∩ A2) = P(A1) P(A2) = (⅓)(½) = ⅙ .

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

2.3 Mutually Exclusive Events

Mutually exclusive (disjoint):

- Events are the events that cannot happen simultaneously.
- Two events A and B are said to be mutually exclusive if they do not contain any element in common. So the probability of the two events will be zero.
- P(A and B) = 0, i.e., P(A ∩ B) = 0.
- When rolling a die, if A = getting an even number and B = getting an odd number, then A and B are said to be mutually exclusive events, as we cannot get both at the same time.



Two or more events are said to be mutually exclusive if the occurrence of any one of them means the others will not occur (That is, we cannot have 2 or more such events occurring at the same time). In the diagram, we see that there is no overlap in two mutually exclusive events. Hence, the probability of two mutually exclusive events is zero.

Now, suppose "A or B" denotes the event that "either A or B both occur", then

(a) If A and B are not mutually exclusive events: P(A or B) = P(A) + P(B) − P(A and B)

We can also write: P(A U B) = P(A) + P(B) − P(A ∩ B)

Complements of Events

The event A' (pronounce as A 'dash'), the complement of event A, is the event consisting of everything in the sample space S that is not contained within the event A.

2.4 Collectively Exhaustive Events

- In a sample space S, if events A and B are said to be Collectively Exhaustive events, then

  A U B =S (sample space)

  P(A U B) = 1

  Eg: When rolling a dice, if

- A = Getting an odd number, and

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- B = getting an even number
- A and B are Collectively Exhaustive Events.

In probability, exhaustive is a condition of two or more events which serves a great role in finding the probability as it changes if the events are exhaustive or not. A set of events are said to be exhaustive if there is a certain chance of occurrence of at least one of them when they are all considered together.

For example: Consider the experiment of a fair die being thrown. Then there are six outcomes and all of them are equally likely to occur. Also ,the events of getting different numbers taken together are exhaustive as together at least one of them is certain to happen. For getting a 2 or 5, sure will get one of the numbers during the experiment. So events are exhaustive.

If we compare this with mutually exhaustive events, where no more than one event can occur at a given time, the set of all possible die rolls is both collectively exhaustive and mutually exclusive. The events 1 and 6 are mutually exclusive but not collectively exhaustive. The events "even" (2,4 or 6) and "not-6" (1,2,3,4, or 5) are collectively exhaustive but not mutually exclusive.

What did You Grasp?

1. State True or False.
   A sample space refers to each individual outcome of a random experiment.

   A) True
   B) False

2. If the probability that person A will be alive in 20 years is 0.7 and the probability that person B will be alive in 20 years is 0.5, what is the probability that they will both be alive in 20 years?

   A) 0.5
   B) 0.45
   C) 0.35
   D) 0

3. It is known that the probability of obtaining zero defectives in a sample of 40 items is 0.34 whilst the probability of obtaining 1 defective item in the sample is 0.46. What is the probability of obtaining not more than 1 defective item in a sample?

   A) 1
   B) 0.2
   C) 0.8
   D) 0.35

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

2.5 Conditional Probability

- Conditional Probability of an event A, given that event B, has already happened is given by: P(A|B).
- If A and B are two events in a sample space S, given that P(B) > 0, then the conditional probability of A, assuming B has already happened, is given by:

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

- If an event is taken as a sample space for another event which is a subset of the former event.

The formula for conditional probability is derived based on the following logic. When we know that event B has occurred, every outcome that is outside B should be discarded. Thus, our sample space is reduced to the set B. Now the only way that A can happen is when the outcome belongs to the set A ∩ B. We divide P(A ∩ B) by P(B), so that the conditional probability of the new sample space becomes 1, i.e., P(B|B)=P(B∩B) divided by P(B) = 1.

Example:

A pair of four-sided dice is rolled and the sum is determined. What is the probability that a sum of 3 is rolled before a sum of 5 is rolled in a sequence of rolls of the dice?

Solution:

The sample space of this random experiment is

S = {(1, 1) (1, 2) (1, 3) (1, 4)

(2, 1) (2, 2) (2, 3) (2, 4)

(3, 1) (3, 2) (3, 3) (3, 4)

(4, 1) (4, 2) (4, 3) (4, 4)}.

Let A be the event of getting a sum of 3 and B be the event of getting a sum of 5. The probability that a sum of 3 is rolled before a sum of 5 is rolled can be thought of as the conditional probability of a sum of 3, given that a sum of 3 or 5 has occurred. That is, P(A/A U B). Hence,

P(A/A U B) = P(A ∩ (A U B))/P(A U B) = P(A)/P(A) + P(B) = 2/2 + 4 = ⅓ .

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

What did You Grasp?

1. What is the probability that the total of two dice will be greater than 8, given that the first die is a 6?

   A) ⅓
   B) 1/9
   C) ⅙
   D) ⅔

2.6 Law of Total Probability

If $A_1, A_2, A_3, \ldots$ An is a partition of a sample space, (they are mutually exclusive and totally exhaustive events) then the probability of any event B can be obtained from the probabilities $P(A_i)$ and $P(B|A_i)$ using the formula (based on conditional probability):

i.e. $$P(B) = P(A_1)P(B \mid A_1) + \cdots + P(A_n)P(B \mid A_n)$$

$$P(B) = \sum_i P\,(B|Ai)P(Ai)$$

Let's see an example of Total Probability.

I have three bags that each contain 100 marbles.

Bag 1 has 75 red and 25 blue marbles.

Bag 2 has 60 red and 40 blue marbles.

Bag 3 has 45 red and 55 blue marbles.

I choose one of the bags at random and then pick a marble from the chosen bag, also at random. What is the probability that the chosen marble is red?

Solution:

Let R be the event that the chosen marble is red. Let Bi be the event that I choose Bag i. We already know that

$P(R|B_1)=0.75$,

$P(R|B_2)=0.60$,

$P(R|B_3)=0.45$

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

We choose our partition as B_1, B_2, B_3. Note that this is a valid partition because, firstly, the Bi's are disjoint (only one of them can happen), and secondly, because their union is the entire sample space as one the bags will be chosen for sure, i.e., $P(B_1 \cup B_2 \cap B_3) = 1$.

Using the law of total probability, we can write

$P(R) = P(R|B_1)P(B_1)+P(R|B_2)P(B_2)+P(R|B_3)P(B_3)$

$=(0.75)\frac{1}{3} + (0.60)\frac{1}{3} + (0.45)\frac{1}{3}$

$=0.60$

2.7 Bayes' Theorem

- If A_1, A_2, A_3, …. An is a partition of a sample space, (they are mutually exclusive and totally exhaustive events) then the posterior probabilities of the event Ai, conditional on any event B can be obtained from the probabilities, $P(Ai) > 0$ and $P(B) > 0$.
- Suppose that we know $P(A|B)$, but the probability $P(B|A)$ is unknown. Using the definition of conditional probability, we have:
- $P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$
- Dividing by $P(A)$, we get the Bayes' rule, that is :

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

There are many situations where the ultimate outcome of an experiment depends on what happens in various intermediate stages. This issue is resolved by the Bayes' Theorem.

2.7.1 Bayes' Rule

The following are two Bayes' rules:

- **Rule 1:** For any two events A and B, where $P(A) \neq 0$,

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

- **Rule 2:** If B_1, B_2, B_3 …. Bn, form a partition of the sample space S, and A is any event with $P(A) \neq 0$,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

Bayes' Theorem (also known as Bayes' rule) is a deceptively simple formula used to calculate conditional probability. For two events, A and B, Bayes' theorem allows you to figure out P(A|B) (the probability that event A happened, given that test B was positive) from p(B|A) (the probability that test B happened, given that event A happened).

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Example:

Imagine you want to find out a patient's probability of having liver disease if they are an alcoholic.

- **A** could mean the event "Patient has liver disease." Past data tells that 10% of patients entering your clinic have liver disease. P(A) = 0.10.
- **B** could mean the litmus test that "Patient is an alcoholic." 5% of the clinic's patients are alcoholics. P(B) = 0.05.

You might also know that among those patients diagnosed with liver disease, 7% are alcoholics. This is your B|A: the probability that a patient is alcoholic, given that they have liver disease, is 7%.

By Bayes' theorem:

P(A|B) = (0.07 * 0.1)/0.05 = 0.14

In other words, if the patient is an alcoholic, their chances of having liver disease is 0.14 (14%). This is a large increase from the 10% suggested by past data. But it's still unlikely that any particular patient has liver disease.

2.8 Random Variables

- A random variable, usually written X, is a variable whose possible values are numerical outcomes of a random phenomenon.
- Though it is named as a variable, it is a function from Sample space S to R (set of real numbers).

$$X: S \rightarrow R$$

- A random variable is a real-valued variable whose value is determined by an underlying random experiment.
- There are two types of random variables, discrete and continuous.

Objects in a sample space may not be numbers. Thus, we use the notion of a random variable to quantify the qualitative elements of the sample space.

To analyze random experiments, we usually focus on some numerical aspects of the experiment. For example, in a soccer game, we may be interested in the number of goals, shots, shots on goal, corners kicks, fouls, etc. If we consider an entire soccer match as a random experiment, then each of these numerical results gives some information about the outcome of the random experiment. These are examples of random variables. In a nutshell, a random variable is a real-valued variable whose value is determined by an underlying random experiment.

For example, if we toss a coin 5 times, the random experiment and the sample space can be written as: S = {TTTTT, TTTTH, TTTHH, ... HHHHH}.

The sample space S here has $2^5 = 32$ elements. Suppose that in this experiment, we are interested in the number of heads. We can define a random variable X whose value is the number of observed

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

heads. The value of X will be one of 0, 1, 2, 3, 4 or 5 depending on the outcome of the random experiment.

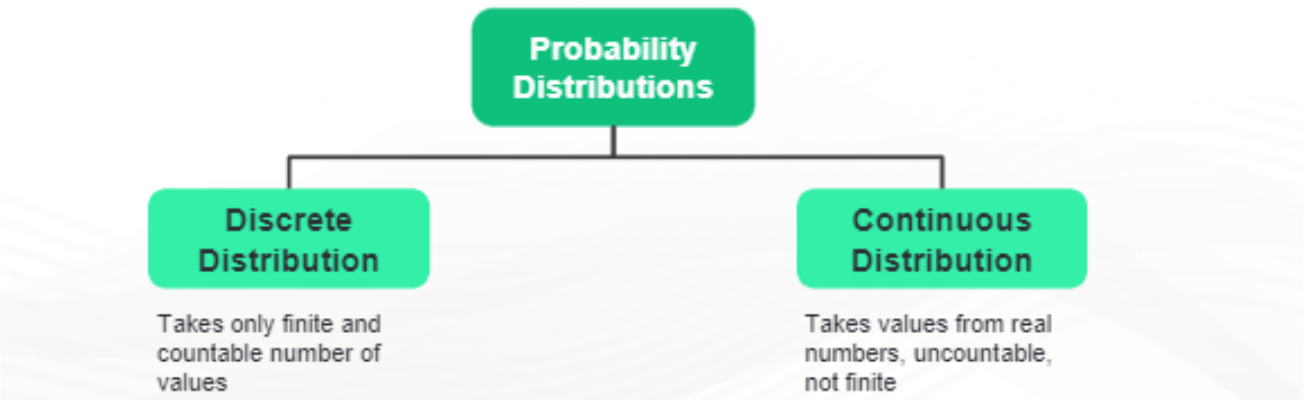
Since a random variable is a function, we can talk about its range. The range of a random variable X, shown by Range(X) or R_X, is the set of possible values for X. In the above example, Range(X) = R_X = {0, 1, 2, 3, 4, 5}.

A random variable is characterized by either its probability density function or its cumulative distribution function. The other characteristics of a random variable are its mean, variance and moment generating function.

There are two types of random variables: discrete random variables and continuous random variables, and based on this there are two types of probability distribution functions.

3. Probability Distributions

There are two major categories of probability distributions.



A discrete random variable is one whose set of assumed values is countable (arises from counting). A discrete probability distribution is a table (or a formula) listing all possible values that a discrete variable can take on, together with the associated probabilities.

A continuous random variable is one whose set of assumed values is uncountable (arises from a measurement.).

3.1 Discrete Probability Distribution

- A discrete distribution describes the probability of occurrence of each value of a discrete random variable.
- A discrete random variable is a random variable that has countable values.
- When we say that the probability distribution of an experiment is discrete then the sum of probabilities of all possible values of the random variable must be equal to 1. That is, if X is a discrete random variable, then,

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

$$\sum P(Xi)=1$$

(Here, 'i = 1, 2, 3, ... n' is the set of all values that the variable X can take.)

A discrete random variable is one which may take on only a countable number of distinct values such as 0,1,2,3,4,..... Discrete random variables are usually (but not necessarily) counts. If a random variable can take only a finite number of distinct values, then it must be discrete.

Examples of discrete random variables include the number of children in a family, the Friday night attendance at a cinema, the number of patients in a doctor's surgery, the number of defective light bulbs in a box of ten, etc.

The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function.

Some of the discrete probability distributions are Poisson distribution, Bernoulli distribution, Binomial distribution, etc. We'll learn about some of these distributions in detail in the upcoming sections.

3.1.1 Important Formulae

The following are two important formulae of Discrete probability distribution:

Mean $$\mu_x = x_1p_1 + x_2p_2 + \cdots + x_kp_k = \sum x_ip_i$$

Variance $$\sigma_x^2 = \sum (x_i - \mu_x)^2 p_i$$

The standard deviation σX is the square root of the variance.

Suppose a variable X can take the values 1, 2, 3, or 4. The probabilities associated with each outcome are described by the following table:

For the outcomes 1, 2, 3, 4, the probabilities are 0.1, 0.3, 0.4 and 0.2, respectively.

The probability that X is equal to 2 or 3 is the sum of the two probabilities: P(X = 2 or X = 3) = P(X = 2) + P(X = 3) = 0.3 + 0.4 = 0.7. Similarly, the probability that X is greater than 1 is equal to 1 - P(X = 1) = 1 - 0.1 = 0.9, by the complement rule.

3.1.2 Discrete Uniform Distribution

- A random variable has a uniform distribution when each value of the random variable is equally likely, and values are uniformly distributed throughout some interval.
- A discrete uniform distribution is a symmetric probability distribution whereby a finite number of values are equally likely to be observed; every one of n values has equal probability 1/n.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- The cumulative distribution function (CDF) of the discrete uniform distribution can be expressed, for any k ∈ [a,b], as:

$$F(k; a, b) = \frac{\lfloor k \rfloor - a + 1}{b - a + 1}$$

An example of a discrete uniform distribution on the first N integers is the statistical experiment of rolling one die, where the random variable X represents the outcome of the die.

For the standard six-sided die, we have the probability P(X) = ⅙ for each outcome. Furthermore, the expected value is E(X) = 6+1 / 2=3.5, so over the long run, the average of the outcomes should be midway between 3 and 4. We also find that the variance is Var(X) = 6²−1 / 12 = 35 / 12 ≈ 2.9167, and the standard deviation of the outcomes is σX = sqrt(35 / 12) ≈ 1.7078.

3.1.3 Poisson Distribution

- Developed by the French mathematician Simeon Denis Poisson in 1837.
- Probability distribution of a Poisson random variable X representing the number of successes occurring in a given time interval or a specified region of space is given by the formula:

$$P(X) = \frac{e^{-\mu}\mu^{x}}{x!}$$

(x = 0, 1, 2, 3… n

e = 2.71828 (but use your calculator's 'e' button)

μ = mean number of successes in the given time interval or region of space)

Poisson Distribution was developed by the French mathematician Simeon Denis Poisson in 1837.

A Poisson random variable must satisfy the following conditions:

- The number of successes in two disjoint time intervals is independent.
- The probability of a success during a small time interval is proportional to the entire length of the time interval.

The formula for calculating the probability distribution of a Poisson random variable is given in the slide.

Applications of Poisson distribution include:

- The number of deaths by horse kicking in the Prussian army (the first application)
- Birth defects and genetic mutations
- Rare diseases (like Leukemia, but not AIDS because it is infectious and so not independent) - especially in legal cases

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- Car accidents
- Traffic flow and an ideal gap distance
- Number of typing errors on a page
- Hair found in McDonald's hamburgers
- Spread of an endangered animal in Africa
- Failure of a machine in one month

3.1.3.1 Example of Poisson Distribution

Problem:

If electricity power failures occur according to a Poisson distribution with an average of 3 failures every 20 weeks, calculate the probability that there will not be more than one failure during a particular week.

Solution:

The average number of failures per week is: μ = 3/20 = 0.15

$$P(x0) + P(x1) = \frac{e^{-0.15}\,0.15^0}{0!} + \frac{e^{-0.15}\,0.15^1}{1!}$$

= 0.98981

In this example, we've added the probabilities for 0 failures and 1 failure, since it is given that "Not more than one failure".

3.1.3.2 Mean and Variance of Poisson Distribution

If μ is the average number of successes occurring in a given time interval or region in the Poisson distribution, then the mean and the variance of the Poisson distribution are both equal to μ.

($E(X) = \mu$ and $V(X) = \sigma^2 = \mu$)

In a Poisson distribution, only one parameter, μ is needed to determine the probability of an event.

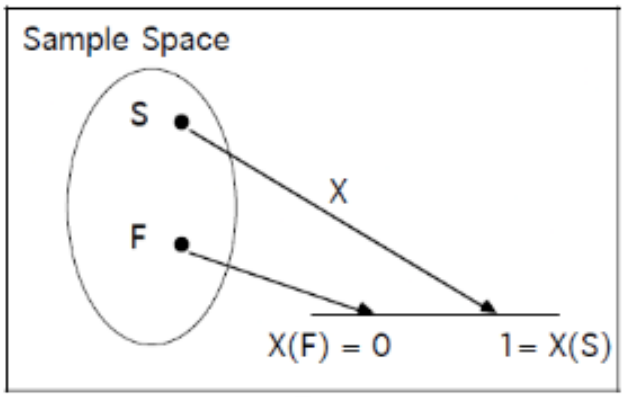
This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

What did You Grasp?

1. A life insurance salesman sells on an average 3 life insurance policies per week. Use Poisson's law to calculate the probability that in a given week he will sell 2 or more policies, but less than 5 policies.

   A) 0.61611
   B) 0.32926
   C) 0.62358
   D) 0.52369

3.1.4 Bernoulli Distribution



Bernoulli Distribution:

The probability density function of a random variable is

$f(0) = P(X = 0) = 1 - p$

$f(1) = P(X = 1) = p$,

(where p denotes the probability of success. Hence,

$f(x) = p^x (1 - p)^{1-x}$, $x = 0,)$

The random variable X is called the Bernoulli random variable if its probability density function is of the form

$f(x) = p^x (1 - p)^{1-x}$, $x = 0, 1$, where p is the probability of success.

We denote the Bernoulli random variable by writing $X \sim BER(p)$.

Example:

What is the probability of getting a score of not less than 5 in a throw of a six-sided die?

Solution:

Although there are six possible scores {1, 2, 3, 4, 5, 6}, we are grouping them into two sets, namely {1, 2, 3, 4} and {5, 6}. Any score in

{1, 2, 3, 4} is a failure and any score in {5, 6} is a success. Thus, this is a Bernoulli trial with:

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

P(X = 0) = P(failure) = 4/6

P(X = 1) = P(success) = 2/6.

Hence, the probability of getting a score of not less than 5 in a throw of a six-sided die is 2/6.

3.1.4.1 Bernoulli's Theorem

Bernoulli Theorem:

If X is a Bernoulli random variable with parameter p, then the mean, variance and moment generating functions are respectively given by:

$$\mu_X = p$$
$$\sigma_X^2 = p\,(1-p)$$
$$M_X(t) = (1-p) + p\,e^t.$$

Proof for mean, standard deviation and moment generating function:

$$\begin{aligned}\mu_X &= \sum_{x=0}^{1} x\,f(x)\\ &= \sum_{x=0}^{1} x\,p^x\,(1-p)^{1-x}\\ &= p.\end{aligned}$$

$$\begin{aligned}\sigma_X^2 &= \sum_{x=0}^{1} (x-\mu_X)^2\,f(x)\\ &= \sum_{x=0}^{1} (x-p)^2\,p^x\,(1-p)^{1-x}\\ &= p^2\,(1-p) + p\,(1-p)^2\\ &= p\,(1-p)\,[p + (1-p)]\\ &= p\,(1-p)\end{aligned}$$

$$\begin{aligned}M(t) &= E\left(e^{tX}\right)\\ &= \sum_{x=0}^{1} e^{tx}\,p^x\,(1-p)^{1-x}\\ &= (1-p) + e^t\,p.\end{aligned}$$

Understand Bernoulli's theorem and its proof. Note that for the Bernoulli distribution all its moments about zero are same and equal to p.

3.1.5 Binomial Distribution

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

$$P(X) = C_x^n p^x q^{n-x}$$

where,

n = the number of trials; x = 0, 1, 2, ... n; p = the probability of success in a single trial; q = the probability of failure in a single trial (i.e. q = 1 − p);

C_x^n is a combination.

Binomial Distribution applies to when chances are two alternatives.ie, p = success and q = failure and p + q = 1.

Example:

The probability of winning a match for team A is 0.60. Find the probability of winning 3 matches out of 5.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Solution:

Probability of winning, p = 0.60

Probability of losing, q = 0.40

Probability of winning 3 matches out of 5, $P(x=3)5 = 5C3(0.6)^3(0.4)^2$

$P(x=3) = 5C3 \times 0.216 \times 0.16 = 10 * 0.03456$

Hence, the probability is **0.3456**.

3.1.5.1 Properties of a Binomial Experiment

A Binomial experiment has the following properties:

- The experiment consists of n repeated trials.
- Each trial results in an outcome that may be classified as a success or a failure (hence the name, binomial).
- The probability of a success, denoted by p, remains constant from trial to trial and repeated trials are independent.
- The number of successes X in n trials of a binomial experiment is called a binomial random variable.

Go through the pointers in the slide and understand the properties of a Binomial distribution. The Binomial distribution can arise whenever we select a random sample of n units with replacement. Each unit in the population is classified into one of two categories according to whether it does or does not possess a certain property.

3.1.5.2 Constants of Binomial Distribution

If p is the probability of success and q is the probability of failure in a binomial trial, then the expected number of successes in n trials (i.e. the mean value of the binomial distribution) is:

$E(X) = \mu = np$

The variance of the binomial distribution is:

$V(X) = \sigma 2 = npq$

Standard deviation = $\sigma = \sqrt{npq}$

In a binomial distribution, only 2 parameters, namely n, and p, are needed to determine the probability.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

What did You Grasp?

1. A die is tossed 5 times. What is the probability of no fives turning up?

   A) 0.5787
   B) 0.2356
   C) 0.3472
   D) 0.2569

3.2 Covariance

Covariance:

Provides a measure of the strength of the correlation between two or more sets of random variates.

Is the relationship between two variables in a given data set.

Example: Let E(x) be the expected value of a given variable x, and E(y) be the expected value of variable y, then the covariance between x and y is given by:

$$\text{cov}(x, y) = E[xy] - E[x]\,E[y]$$

The covariance generalizes the concept of variance to multiple random variables. Instead of measuring the fluctuation of a single random variable, the covariance measures the fluctuation of two variables with each other.

3.2.1 Properties of Covariance

Given a constant 'a' and random variables X, Y, and Z, the following properties are:

$$\begin{aligned}
&\mathrm{Cov}(X, X) = \mathrm{Var}(X) \geq 0 \\
&\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X) \\
&\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y) \\
&\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z).
\end{aligned}$$

Go through the slide and understand the properties of covariance.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

3.3 Correlation

Correlation:

- Depends upon two variables, change in one variable effects a change in second variable.
- Its value lies in the range of -1 and +1. Whereas in covariance two variables vary together, which can be negative or positive.
- Is the measure of the strength of the relation between two variables. How strongly two variables are connected is defined as the correlation.
- Is related to covariance by the given formula: $cor(x,y)= cov(x,y)/\sigma x\ \sigma y$

Correlation is used to test relationships between quantitative variables or categorical variables. In other words, it's a measure of how things are related. The study of how variables are correlated is called correlation analysis.

Some examples of data that have a high correlation:

Your caloric intake and your weight.

Your eye color and your relatives' eye colors.

The amount of time your study and your GPA.

Some examples of data that have a low correlation (or none at all):

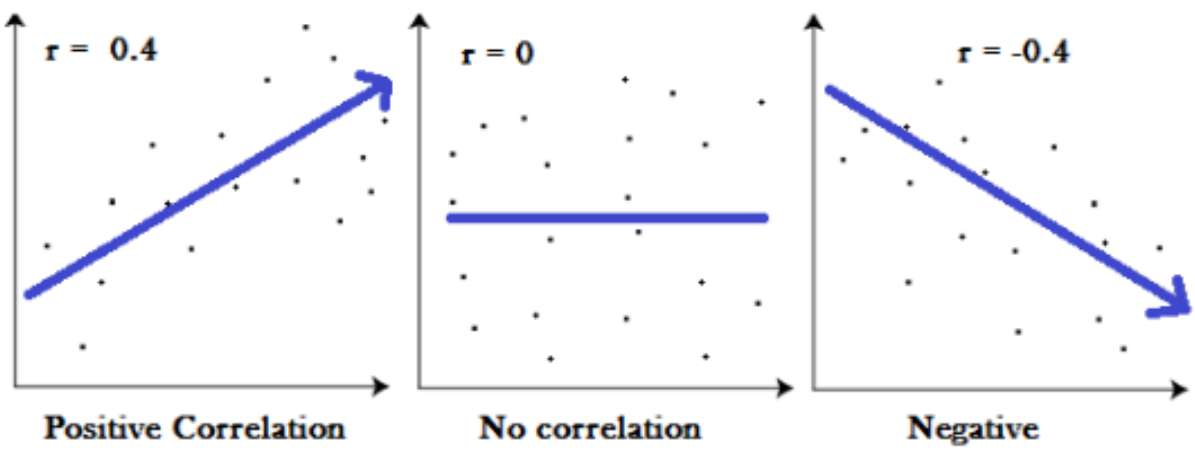
A dog's name and the type of dog biscuit they prefer.

The cost of a car wash and how long it takes to buy a soda inside the station.

3.3.1 Correlation Coefficient

A correlation coefficient is a way to put a value to the relationship.

The following graph shows the correlation of -1, 0 and 1.



This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Correlation coefficients have a value of between -1 and 1. A "0" means there is no relationship between the variables at all, while -1 or 1 means that there is a perfect negative or positive correlation (negative or positive correlation here refers to the type of graph the relationship will produce).

4. Continuous Random Variables

Continuous Random Variable:

- A variable, which takes an infinite number of possible values.
- Are usually measurements.
- Examples include height, weight, the amount of sugar in an orange, the time required to run a mile.
- Is not defined at specific values. Instead, it is defined over an interval of values, and is represented by the area under a curve (in advanced mathematics, this is known as an integral).
- 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.

Suppose a random variable X may take all values over an interval of real numbers. Then the probability that X is in the set of outcomes A, P(A), is defined to be the area above A and under a curve. The curve, which represents a function p(x), must satisfy the following:

1: The curve has no negative values ($p(x) > 0$ for all x)

2: The total area under the curve is equal to 1.

A curve meeting these requirements is known as a Density curve.

4.1 Continuous Probability Distribution

The probability distribution of a continuous random variable.

The equation that describes the continuous probability distribution is called probability density function.

The criteria for a probability density function are as follows:

- The random variable Y is a function of X; that is, $y = f(x)$.
- The value of y is greater than or equal to zero for all values of x.
- The total area under the curve of the function is equal to one.

A continuous probability distribution differs from a discrete distribution in a few ways.

The probability distribution of a continuous random variable is called continuous probability distribution. The probability density function is the equation that describes the continuous probability distribution.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

The conditions that the probability density function should follow, are given above. A continuous probability distribution differs from a discrete probability distribution in a few ways.

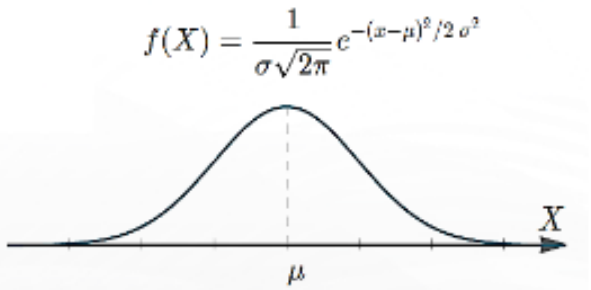
- The probability that a continuous random variable will assume a particular value is zero.
- Therefore, a continuous probability distribution cannot be expressed in tabular form.
- An equation or formula is used to describe a continuous probability distribution, which is the probability density function.

Some of the examples for continuous probability distribution are: normal distribution, exponential distribution, t-distribution, chi-square distribution, and so on. We'll see in detail, some of these distributions in the forthcoming sections.

4.1.1 Normal Distribution

- The Normal Probability Distribution is very common in the field of statistics.
- The random variable X is said to be normally distributed with mean μ and standard deviation σ if its probability distribution is given by

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\,\sigma^2}$$



A normal curve

(where $-\infty < \mu < \infty$ and $0 < \sigma^2 < 1$ are arbitrary parameters. If X has a normal distribution with parameters μ and σ^2, then we write $X \sim N(\mu, \sigma^2)$.)

Among continuous probability distributions, the normal distribution is very well known since it arises in many applications. Normal distribution was discovered by a French mathematician Abraham DeMoivre.

Whenever you measure things like people's height, weight, salary, opinions or votes, the graph of the results is very often a normal curve. The main importance of normal distribution lies on the central limit theorem which says that the sample mean has a normal distribution if the sample size is large.

4.1.1.1 Properties of Normal Distribution

The properties of normal distribution are as follows:

- The normal curve is symmetrical about the mean μ.
- The mean is at the middle and divides the area into halves.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- The total area under the curve is equal to 1.
- It is completely determined by its mean and standard deviation σ (or variance σ2)

In a normal distribution, only 2 parameters are needed, namely μ and σ2.

Go through the slide and understand the properties of normal distribution.

4.1.1.2 Area Under the Normal Curve

- The probability that a variable is within range in a normal distribution is calculated by finding the area under the normal curve.
- The area depends on the values of μ (mean) and σ (standard deviation).
- The z-score table is used to find the area under the normal curve.
- Z-score is the standardized value of observation x from a distribution that has mean μ and standard deviation σ.
- In a z-score table, the left most column means the number of standard deviations above the the mean to 1 decimal place, the top row gives the second decimal place, and the intersection of a row and column gives the probability.

Area under the normal curve helps us in finding the probability of a variable being in the range of a normal distribution. The z-score table is used to find the area under the normal curve. There are a lot of sources to access the z-score table. An example for such source is: http://users.stat.ufl.edu/~athienit/Tables/Ztable.pdf.

As given above, in a z-score table, the left most column means the number of standard deviations above the the mean to 1 decimal place, the top row gives the second decimal place, and the intersection of a row and column gives the probability.

For example, if we want to know the probability that a variable is no more than 0.51 standard deviations above the mean, we need to see the 6th row down in page 2 of the pdf, that starts with 0.0

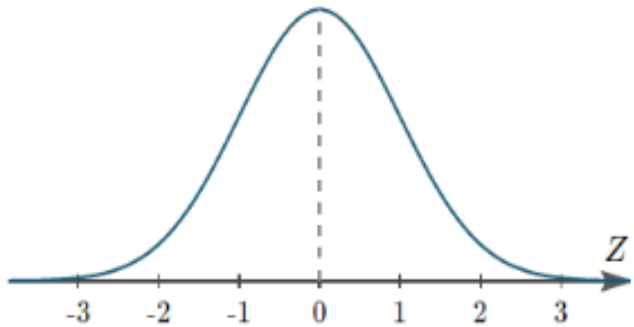
This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

(corresponding to 0.5) and the 2nd column (corresponding to 0.01). The intersection of the 6th row and 2nd column is 0.6950. This tells us that there is a 69.50% percent chance that a variable is less than 0.51 sigmas above the mean.

4.1.1.3 Standard Normal Distribution

- A random variable X is called a standard normal variate if its mean is zero and its standard deviation is unity, i.e. 1.
- We can transform all the observations of any normal random variable X with mean μ and variance σ to a new set of observations of another normal random variable Z with mean 0 and variance 1 using the following transformation:

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



Standard Normal Curve $\mu = 0, \sigma = 1$

A normal distribution with mean μ and standard deviation σ can be converted into standard normal distribution performing change of scale and origin. The new distribution of the normal random variable Z with mean 0 and variance 1 (or standard deviation 1) is called a standard normal distribution. Standardizing the distribution like this makes it much easier to calculate probabilities.

4.1.2 Continuous Uniform Distribution

The probability density function and cumulative distribution function for a continuous uniform distribution on the interval [a,b] are:

$$P(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{for } x > b \end{cases}$$

$$D(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \le x \le b \\ 1 & \text{for } x > b. \end{cases}$$

These can be written in terms of the Heaviside step function $H(x)$ as

$$P(x) = \frac{H(x-a) - H(x-b)}{b-a}$$

$$D(x) = \frac{(x-a)H(x-a) - (x-b)H(x-b)}{b-a},$$

the latter of which simplifies to the expected $D(x) = (x-a)/(b-a)$ for $a < x < b$.

The continuous distribution is implemented as `UniformDistribution`[a, b].

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

A uniform distribution is sometimes also known as a rectangular distribution, is a distribution that has constant probability.

4.1.2.1 Characteristic Function for a Uniform Distribution

The following formula explains the characteristic function:

For a *continuous* uniform distribution, the characteristic function is

$$\phi(t) = \frac{2}{(b-a)t} \sin\left[\frac{1}{2}(b-a)t\right] e^{i(a+b)t/2}.$$

If $a = 0$ and $b = 1$, the characteristic function simplifies to

$$\phi(t) = \frac{2\sin\left(\frac{1}{2}t\right)e^{it/2}}{t} = \frac{i - i\cos t + \sin t}{t}.$$

The formula for finding the characteristic function for a uniform distribution is given above.

4.1.2.2 Moment Generating Function of a Uniform Distribution

The following formula explains the moment generating function of a uniform distribution:

If we assume t ≠ 0, then:

$$\begin{aligned} M(t) &= E\left(e^{tX}\right) \\ &= \int_a^b e^{tx} \frac{1}{b-a}\, dx \\ &= \frac{1}{b-a}\left[\frac{e^{tx}}{t}\right]_a^b \\ &= \frac{e^{tb} - e^{ta}}{t(b-a)}. \end{aligned}$$

If t = 0, and M(0) = 1,

$$M(t) = \begin{cases} 1 & \text{if } t = 0 \\ \frac{e^{tb} - e^{ta}}{t(b-a)}, & \text{if } t \neq 0 \end{cases}$$

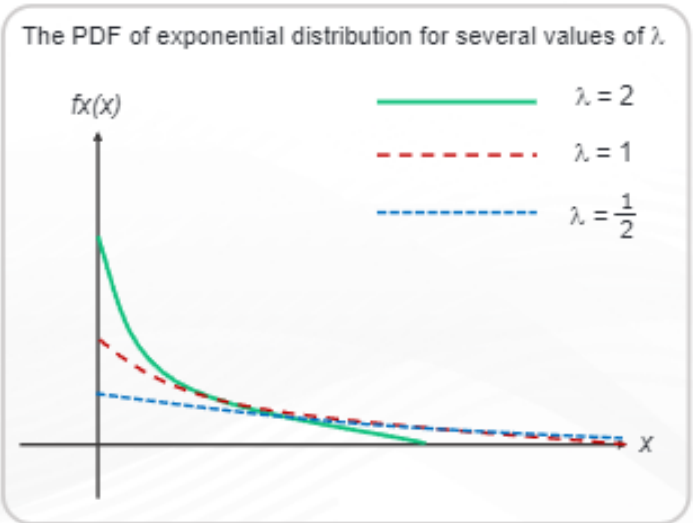
The formula for calculating the moment generating function for a uniform distribution is given above.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

4.1.3 Exponential Distribution

Exponential Distribution:

- Is one of the widely used continuous distributions. It is often used to model the time elapsed between events.
- Is first defined mathematically, and its mean and expected value are derived.
- The intuition for the distribution will then be developed.



An interesting property of the exponential distribution is that it can be viewed as a continuous analogue of the geometric distribution. You toss a coin (repeat a Bernoulli experiment) until you observe the first heads (success). Now, suppose that the coin tosses are Δ seconds apart and in each toss, the probability of success is p = Δλ. Also, suppose that Δ is very small, so the coin tosses are very close together in time and the probability of success in each trial is very low. Let X be the time you observe the first success.

To get some intuition for this interpretation of the exponential distribution, suppose you are waiting for an event to happen. For example, you are at a store and are waiting for the next customer. In each millisecond, the probability that a new customer enters the store is very small. You can imagine that, in each millisecond, a coin (with a very small P(H)) is tossed, and if it lands heads a new customer enters. If you toss a coin every millisecond, the time until a new customer arrives approximately follows an exponential distribution.

4.1.3.1 Formula for Exponential Distribution

A continuous random variable X is said to have an exponential distribution, with parameter λ > 0, shown as X ~ Exponential(λ), its probability density function (PDF) is given by:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Go through the slide and understand the formula for calculating exponential distribution.

4.1.3.2 CDF of an Exponential Distribution

It is convenient to use the unit step function defined as

$$u(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

so we can write the PDF of an $Exponential(\lambda)$ random variable as

$$f_X(x) = \lambda e^{-\lambda x} u(x).$$

Let us find its CDF, mean and variance. For $x > 0$, we have

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}.$$

So we can express the CDF as

$$F_X(x) = \left(1 - e^{-\lambda x}\right) u(x).$$

The steps involved in calculating the CDF of exponential distribution is given in the slide.

4.1.3.3 Variance of Exponential Distribution

Let $X \sim Exponential(\lambda)$. We can find its expected value as follows, using integration by parts:

$$\begin{aligned} EX &= \int_0^\infty x\lambda e^{-\lambda x} dx \\ &= \frac{1}{\lambda} \int_0^\infty y e^{-y} dy \qquad \text{choosing } y = \lambda x \\ &= \frac{1}{\lambda} \Big[- e^{-y} \\ &- y e^{-y} \Big]_0^\infty \\ &= \frac{1}{\lambda}. \end{aligned}$$

Now let's find Var(X). We have

$$\begin{aligned} EX^2 &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= \frac{1}{\lambda^2} \int_0^\infty y^2 e^{-y} dy \\ &= \frac{1}{\lambda^2} \Big[- 2e^{-y} - 2ye^{-y} - y^2 e^{-y} \Big]_0^\infty \\ &= \frac{2}{\lambda^2}. \end{aligned}$$

The steps involved in deriving the formula for calculating the variance of an exponential distribution is given in the slide.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

The formula for calculating the variance of an exponential distribution:

$$\mathrm{Var}(X) = EX^2 - (EX)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

If $X \sim Exponential(\lambda)$, then $EX = \frac{1}{\lambda}$ and $\mathrm{Var}(X) = \frac{1}{\lambda^2}$.

The formula for calculating the variance of an exponential distribution is given in the slide.

4.1.4 p-Value

- When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results. Hypothesis tests are used to test the validity of a claim that is made about a population. This claim that's on trial, in essence, is called the null hypothesis.
- The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it.
- All hypothesis tests ultimately use a p-value to weigh the strength of the evidence (what the data are telling you about the population)

A p-value is used in hypothesis testing to help you support or reject the null hypothesis. The p-value is the evidence against a null hypothesis. The smaller the p-value, the strong the evidence that you should reject the null hypothesis.

P values are expressed as decimals although it may be easier to understand what they are if you convert them to a percentage. For example, a p-value of 0.0254 is 2.54%. This means there is a 2.54% chance your results could be random (i.e. happened by chance). That's pretty tiny. On the other hand, a large p-value of 0.9 (90%) means your results have a 90% probability of being completely random and not due to anything in your experiment. Therefore, the smaller the p-value, the more important ("significant") your results.

4.1.4.1 p-Value Interpretation

- The p-value is a number between 0 and 1 and interpreted in the following way:
- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- The p-values very close to the cutoff (0.05) are considered to be marginal (could go either way).
- Always report the p-value so your readers can draw their own conclusions.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

The p-value is used to test the significance of the hypothesis.

For example, suppose a pizza place claims their delivery times are 30 minutes or less on average but you think it's more than that. You conduct a hypothesis test because you believe the null hypothesis, Ho, that the mean delivery time is 30 minutes max, is incorrect. Your alternative hypothesis (Ha) is that the mean time is greater than 30 minutes.

You randomly sample some delivery times and run the data through the hypothesis test, and your p-value turns out to be 0.001, which is much less than 0.05. In real terms, there is a probability of 0.001 that you will mistakenly reject the pizza place's claim that their delivery time is less than or equal to 30 minutes.

Since typically we are willing to reject the null hypothesis when this probability is less than 0.05, you conclude that the pizza place is wrong; their delivery times are in fact more than 30 minutes on average. (You could also be wrong by having sampled an unusually high number of late pizza deliveries just by chance.)

4.1.5 Confidence Interval

Confidence interval:

- Is an interval estimate combined with a probability statement.

  (This means that if we used the same sampling method to select different samples and computed an interval estimate for each sample, we would expect the true population parameter to fall within the interval estimates 95% of the time.)
- If repeated samples were taken and the 95% confidence interval was computed for each sample, 95% of the intervals would contain the population mean.
- A 95%confidence interval has a 0.95 probability of containing the population mean. 95% of the population distribution is contained in the confidence interval.

Statisticians use a confidence interval to describe the amount of uncertainty associated with a sample estimate of a population parameter. The confidence level describes the uncertainty associated with a sampling method.

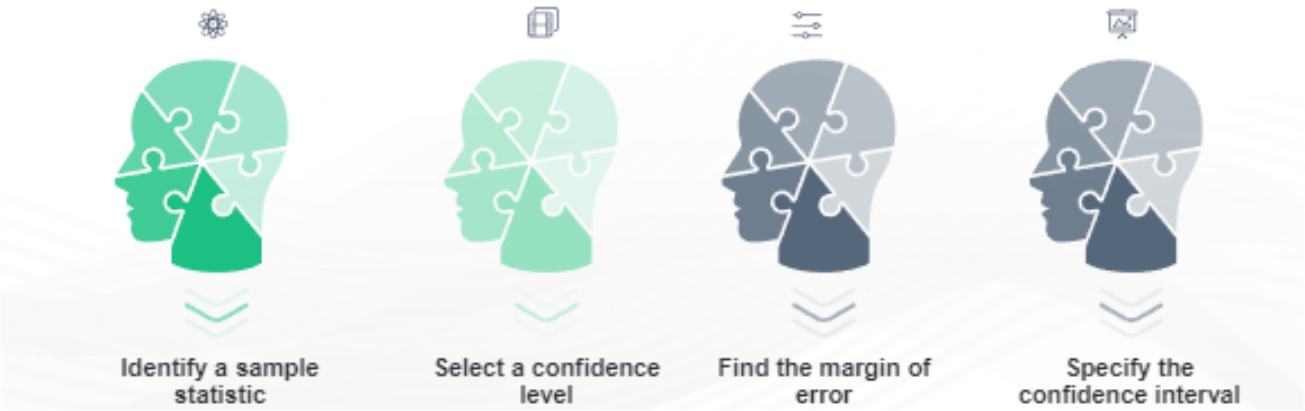
Suppose we used the same sampling method to select different samples and to compute a different interval estimate for each sample. Some interval estimates would include the true population parameter and some would not. A 90% confidence level means that we would expect 90% of the interval estimates to include the population parameter; a 95% confidence level means that 95% of the intervals would include the parameter; and so on.

The purpose of taking a random sample from a lot of population and computing a statistic, such as the mean from the data, is to approximate the mean of the population. A confidence interval addresses this issue because it provides a range of values which is likely to contain the population parameter of interest.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

4.1.5.1 Constructing a Confidence Interval

There are four steps involved in constructing a confidence interval:



To express a confidence interval, you need three pieces of information.

- Confidence level
- Statistic
- Margin of error

There are four steps to constructing a confidence interval.

1. Identify a sample statistic. Choose the statistic (e.g, sample mean, sample proportion) that you will use to estimate a population parameter.
2. Select a confidence level. As we noted in the previous section, the confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
3. Find the margin of error. If you are working on a homework problem or a test question, the margin of error may be given. Often, however, you will need to compute the margin of error, based on one of the following equations.
   Margin of error = Critical value * Standard deviation of statistic
   Margin of error = Critical value * Standard error of statistic
4. Specify the confidence interval. The uncertainty is denoted by the confidence level. And the range of the confidence interval is defined by the following equation.

   Confidence interval = sample statistic + Margin of error

4.1.6 t-Value

T-tests:

- Are statistical hypothesis tests that you use to analyze one or two sample means.
- Depending on the t-test that you use, you can compare a sample mean to a hypothesized value, the means of two independent samples, or the difference between paired samples.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- Are a type of test statistic. Hypothesis tests use the test statistic that is calculated from your sample to compare your sample to the null hypothesis.
- If the test statistic is extreme enough, this indicates that your data are so incompatible with the null hypothesis that you can reject the null.

The term "t-test" refers to the fact that these hypothesis tests use t-values to evaluate your sample data. T-values are a type of test statistic.

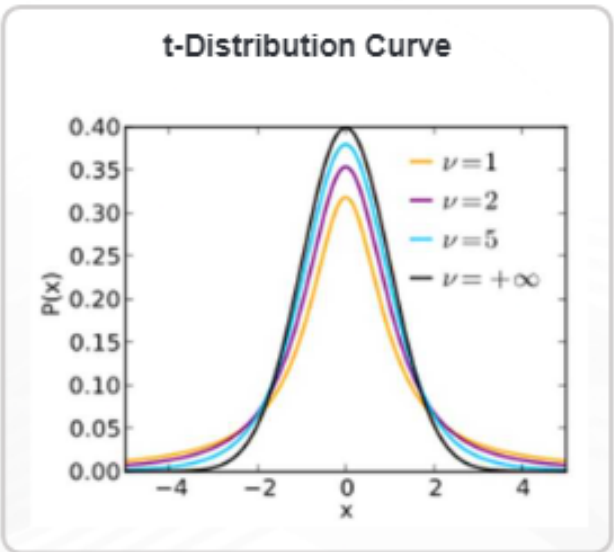
When you analyze your data with any t-test, the procedure reduces your entire sample to a single value, the t-value. These calculations factor in your sample size and the variation in your data. Then, the t-test compares your sample means(s) to the null hypothesis condition in the following manner:

- If the sample data equals the null hypothesis precisely, the t-test produces a t-value of 0.
- As the sample data become progressively dissimilar from the null hypothesis, the absolute value of the t-value increases.

To be able to interpret individual t-values, we have to place them in a larger context. T-distributions provide this broader context so we can determine the unusualness of an individual t-value.

4.1.7 t - Distribution

- The t-distribution (also called Student's t-distribution) is a family of distributions that look almost identical to the normal distribution curve, only a bit shorter and fatter.
- The t-distribution is used instead of the normal distribution when you have small samples.



The t distribution (Student's t-distribution) is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

There are actually many different t distributions. The particular form of the t distribution is determined by its degrees of freedom. The degrees of freedom refers to the number of independent observations in a set of data.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

The larger the sample size, the more the t distribution looks like the normal distribution. In fact, for sample sizes larger than 20 (e.g. more degrees of freedom), the distribution is almost exactly like the normal distribution.

For calculations, we use the t-distribution table, which is available here: http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf.

When you look at the t-distribution table, you'll see "df." This means "degrees of freedom" and is just the sample size minus one.

- Step 1: Subtract one from your sample size. This will be your degrees of freedom.
- Step 2: Look up the df in the left hand side of the t-distribution table. Locate the column under your alpha level (the alpha level is usually given to you in the question).

4.1.7.1 Properties of t-Distribution

The properties of t-distribution are as follows:

- The t distribution ranges from $-\infty$ to ∞.
- The t distribution is bell shaped and symmetric like the standard normal curve.
- The shape of the t distribution changes with the change in degrees of freedom.
- The variance of the t distribution is always greater than one and is defined only for 3 or more degrees of freedom.
- The t distribution has a greater dispersion than the standard normal distribution.

The properties of t-distribution are listed above in the slide.

4.1.7.2 t-Distribution Formula

t-Distribution is given by the formula:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

where

$\bar{x}$ is the mean of the first sample;

μ is the mean of the second sample;

$\frac{s}{\sqrt{N}}$ is the estimate of the standard error of the difference between the means.

The formula for calculating t-distribution is given in the slide.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

4.1.7.3 Assumptions of t-Distribution

The following are the assumptions of student's t test:

Based on the assumptions given above, the test of hypothesis of population mean is given as follows:

When the population standard deviation σ is unknown then to test the population means we use the t statistic. The null hypothesis is there is no difference between the sample mean and the population mean and alternative hypothesis is there is difference between the sample mean and the population mean.

We find the t table value at α level of significance and n –1 degrees of freedom. If the calculated value is less than the critical value, then we accept the null hypothesis that the population mean and the sample mean are equal. Otherwise we reject the null hypothesis and conclude that the population mean and the sample mean are not equal.

4.1.7.4 Example for t-Distribution

Problem:

Acme Corporation manufactures light bulbs. The CEO claims that an average Acme light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

There are two ways to solve this problem, using the T Distribution Calculator. Both approaches are presented in the next slide.

Solution A is the traditional approach. It requires you to compute the t statistic, based on data presented in the problem description. Then, you use the T Distribution Calculator to find the probability.

Solution B is easier. You simply enter the problem data into the T Distribution Calculator. The calculator computes a t statistic "behind the scenes", and displays the probability. Both approaches come up with exactly the same answer.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

The first thing we need to do is compute the t statistic, based on the following equation:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

t = (290 - 300) / [50 / sqrt(15)]

t = -10 / 12.909945 = - 0.7745966

- Now, we can use the t distribution Calculator. Since we know the t statistic, we select "t score" from the random variable drop down. The degrees of freedom are equal to 15 - 1 = 14 and the t statistic is equal to - 0.7745966.
- The calculator displays the cumulative probability: 0.226. Hence, if the true bulb life were 300 days, there is a 22.6% chance that the average bulb life for 15 randomly selected bulbs would be less than or equal to 290 days.

We've seen one way of calculating the probability. Now, let's look at the other way. This time, we will work directly with the raw data from the problem. We will not compute the t statistic; the t Distribution Calculator will do the math for us.

Since we will work with the raw data, we select "Sample mean" from the Random Variable drop down. Then, we enter the following data:

- The degrees of freedom are equal to 15 - 1 = 14.
- The population mean equals 300.
- The sample mean equals 290.
- The standard deviation of the sample is 50.

The calculator displays the cumulative probability: 0.226. Hence, there is a 22.6% chance that the average sampled light bulb will burn out within 290 days.

4.1.8 Chi-square Distribution

The distribution of the chi-square statistic is called the chi-square distribution. The **chi-square distribution** is defined by the following probability density function:

$Y = Y_o * (X^2)^{(v/2 - 1)} * e^{-X2/2}$

- Y_o is a constant that depends on the number of degrees of freedom,
- X_2 is the chi-square statistic,
- v = n - 1 is the number of degrees of freedom,
- e is a constant equal to the base of the natural logarithm system (approximately 2.71828).
- Y_o is defined, so that the area under the chi-square curve is equal to one.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

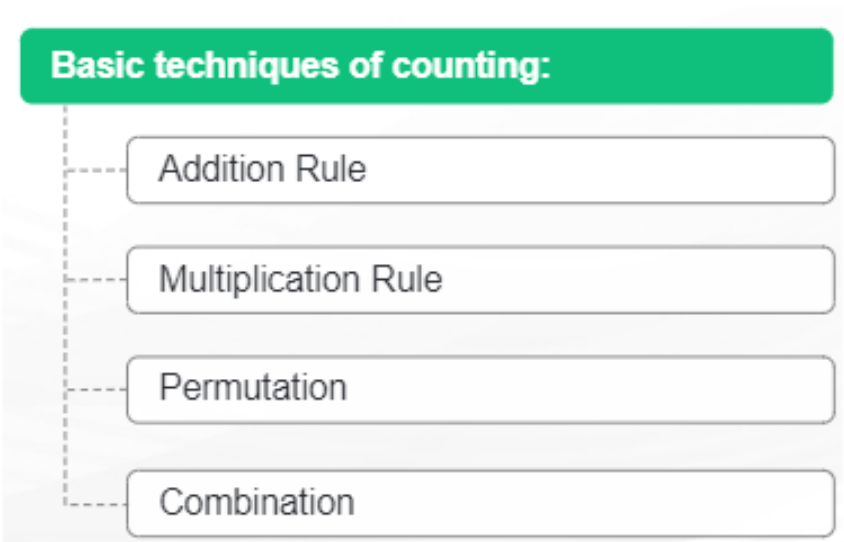
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

4. Bayes Theorem
5. Discrete Probability Distribution
6. Covariance and Correlation
7. Continuous Probability Distribution
8. Central Limit Theorem
9. Hypothesis Testing

1. Principles of Counting

Counting principles:

- Important to handle large scale data and to get a good understanding of probability.
- The branch of mathematics that deals with various counting techniques is called 'combinatorics'.



Counting techniques are important to handle large masses of data. Understanding of the Principles of counting is helpful for gaining a better understanding of probability theory. The following are the basic techniques of counting:

- **Addition rule:**

  The addition rule states that "If E1 with n1 outcomes and E2 with n2 outcomes are mutually exclusive events, and if E is the situation where either E1 or E2 will occur, then

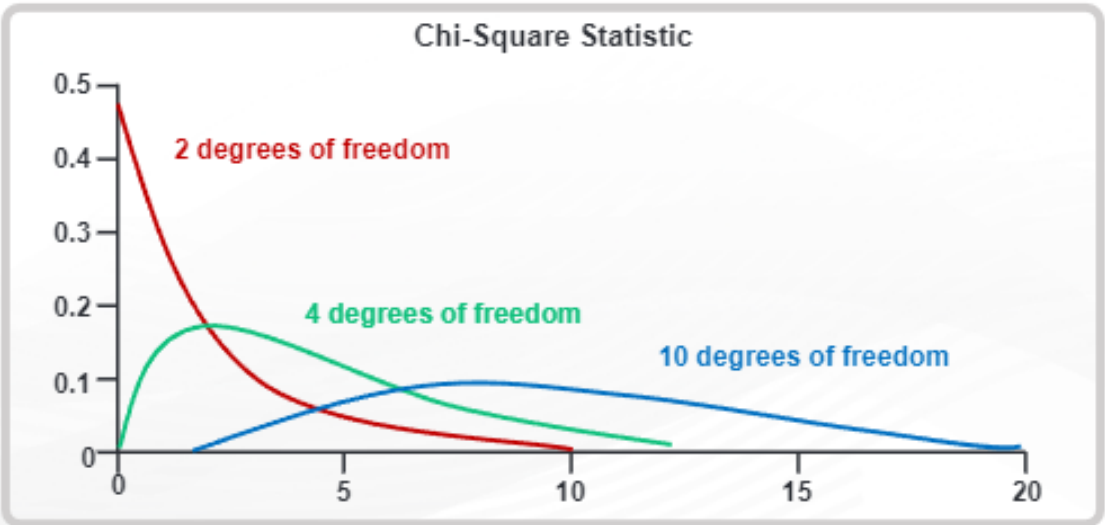
  n = n1 + n2

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

The chi-square distribution results when v independent variables with standard normal distributions are squared and summed. In a testing context, the chi-square distribution is treated as a "standardized distribution" (i.e., no location or scale parameters). However, in a distributional modeling context (as with other probability distributions), the chi-square distribution itself can be transformed with a location parameter, μ, and a scale parameter, σ.

4.1.8.1 Chi-square Statistics

The graph illustrates chi-square distribution of different sample sizes.



In the graph above, the red curve shows the distribution of chi-square values computed from all possible samples of size 3, where degrees of freedom is n - 1 = 3 - 1 = 2.

Similarly, the green curve shows the distribution for samples of size 5 (degrees of freedom equal to 4);

The blue curve, for samples of size 11 (degrees of freedom equal to 10).

4.1.8.2 Properties of Chi-square Distribution

The chi-square distribution has the following properties:

- The mean of the distribution is equal to the number of degrees of freedom: $\mu = v$.
- The variance is equal to two times the number of degrees of freedom: $\sigma^2 = 2*v$.
- When the degrees of freedom is greater than or equal to 2, the maximum value for Y occurs when $X^2 = v - 2$.
- As the degrees of freedom increase, the chi-square curve approaches a normal distribution.

The properties of chi-square distribution are listed in the slide.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

4.1.8.3 Example for Chi-square Distribution

Problem:

The Acme Battery Company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 4 minutes.

Suppose, the manufacturing department runs a quality control test. They randomly select 7 batteries. The standard deviation of the selected batteries is 6 minutes. What would be the chi-square statistic represented by this test?

Your facilitator will now guide you in solving the given problem.

4.1.8.3 Example for Chi-square Distribution (Contd.)

Solution:

- The standard deviation of the population is 4 minutes.
- The standard deviation of the sample is 6 minutes.
- The number of sample observations is 7.

To compute the chi-square statistic, we plug these data in the chi-square equation, as shown below.

$X^2 = [(n - 1) * s^2] / \sigma^2$

$X^2 = [(7 - 1) * 6^2] / 4^2 = 13.5$

where X^2 is the chi-square statistic, n is the sample size, s is the standard deviation of the sample, and σ is the standard deviation of the population.

Let's revisit the problem presented in the previous slide. The manufacturing department ran a quality control test, using 7 randomly selected batteries. In their test, the standard deviation was 6 minutes, which equated to a chi-square statistic of 13.5.

Suppose they repeated the test with a new random sample of 7 batteries. What is the probability that the standard deviation in the new test would be greater than 6 minutes?

Solution

We know the following:

The sample size n is equal to 7.

The degrees of freedom are equal to n - 1 = 7 - 1 = 6.

The chi-square statistic is equal to 13.5 (from the above example).

Given the degrees of freedom, we can determine the cumulative probability that the chi-square statistic will fall between 0 and any positive value. To find the cumulative probability that a chi-square statistic

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

falls between 0 and 13.5, we enter the degrees of freedom (6) and the chi-square statistic (13.5) into the Chi-Square Distribution Calculator. The calculator displays the cumulative probability: 0.96.

This tells us that the probability that a standard deviation would be less than or equal to 6 minutes is 0.96. This means that the probability that the standard deviation would be greater than 6 minutes is 1 - 0.96 or 0.04.

5. Central Limit Theorem

- The central limit theorem (CLT) is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.
- Let $X_1, X_2, ..., X_n$ be a random sample of size n and sample mean from a distribution with mean μ and variance $\sigma^2 < \infty$, then the limiting distribution of

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- is standard normal, that is Zn converges in distribution to Z where Z denotes a standard normal random variable.

The central limit theorem (also known as Lindeberg-Levy Theorem) states that even though the population distribution may be far from being normal, yet for large sample size n, the distribution of the standardized sample mean is approximately standard normal with better approximations obtained with the larger sample size.

All samples will follow an approximately normal distribution pattern, with all variances being approximately equal to the variance of the population divided by each sample size.

According to the central limit theorem, the mean of a sample of data will be closer to the mean of the overall population in question as the sample size increases, notwithstanding the actual distribution of the data, and whether it is normal or non-normal.

As a general rule, sample sizes equal to or greater than 30 are considered sufficient for the central limit theorem to hold, meaning the distribution of the sample means is fairly normally distributed.

5.1 Convergence in Distribution

- The type of convergence used in the Central Limit Theorem is called the convergence in distribution, which is defined as follows:
- Suppose X is a random variable with cumulative density function F(x) and the sequence $X_1, X_2, ...$ of random variables with cumulative density functions $F_1(x), F_2(x), ...$, respectively. The sequence X_n converges in distribution to X if

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

$$\lim_{n \to \infty} F_n(x) = F(x)$$

(for all values x at which F(x) is continuous. The distribution of X is called the limiting distribution of X_n)

Whenever a sequence of random variables X1, X2, ... converges in distribution to the random variable X, it will be denoted by X_n X.

5.2 Properties of Central Limit Theorem

Central Limit Theorem:

- Specifies a theoretical distribution.
- Is formulated by the selection of all possible random samples of a fixed size 'n'.
- Sample mean is calculated for each sample and the distribution of sample means is considered.

The properties of central limit theorem are listed in the slide.

5.3 Examples for Central Limit Theorem

- If an investor is looking to analyze the overall return for a stock index made up of 1,000 stocks, he can take random samples of stocks from the index to get an estimate for the return of the total index.
- The samples must be random, and at least 30 stocks must be evaluated in each sample for the central limit theorem to hold. Random samples ensure a broad range of stock across industries and sectors is represented in the sample.
- Stocks previously selected must also be replaced for selection in other samples to avoid bias. The average returns from these samples approximate the return for the whole index and are approximately normally distributed. The approximation holds even if the actual returns for the whole index are not normally distributed.

Your facilitator will explain the given example on Central Limit Theorem.

5.4 Central Limit Theorem for Sums

- Suppose, X is a random variable with a distribution that may be known or unknown (it can be any distribution), and suppose:

  Ax = the mean of X

  Cx = the standard deviation of x

- The central limit for sums states that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution), which approaches a normal distribution as the sample increases.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

The normal distribution has a mean equal to the original mean multiplied by the sample size

The standard deviation is equal to the original standard deviation multiplied by the square root of the sample size The Central Limit Theorem CLT for Sums Using the Central Limit Theorem

5.5 Example for Central Limit Theorem

Problem:

An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn from the population.

Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7500.

Solution:

Let X = one value from the original unknown population. The probability question asks you to find a probability for the sum (or total of) 80 values.

ÆX = the sum or total of 80 values. Since μx= 90, σx= 15, and n=80, ÆX ~N((80),(90),(80)(15)) – Mean of the sums = (n)(μx) = (80)(90) = 7,200 – Standard deviation of the sums = n σx = 80 15 – Sum of 80 values = Æx = 7,500

6. Hypothesis Testing

- There are two main methods used in inferential statistics:

  1. Estimation

  2. Hypothesis testing

- A Hypothesis is a statement or an assumption about relationships between variables.

  or

- A Hypothesis is a tentative explanation for certain behaviors, phenomenon or events that have occurred or will occur.

We already learned in the previous module about the basics and the types of statistics: descriptive and inferential statistics. There are two major methods used in inferential statistics:

- Estimation
- Hypothesis testing

A hypothesis is an educated guess about something in the world around you. It should be testable, either by experiment or observation.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Examples:

1. A new medicine you think might work.
2. A way of teaching you think might be better.
3. A possible location of new species.
4. A fairer way to administer standardized tests.

Some interesting hypotheses:

Bankers assumed high-income earners are more profitable than low-income earners. Old clients were more likely to diminish CD balances by large amounts compared to younger clients. This was non-intrusive because conventional wisdom suggested that older clients have a larger portfolio of assets and seek less risky investments.

6.1 Hypothesis Statement

- If you are going to propose a hypothesis, it's customary to write a statement. Your statement will look like this:
  - "If I…(do this to an independent variable)….then (this will happen to the dependent variable)."
- For example:
  - If I (decrease the amount of water given to herbs) then (the herbs will increase in size).
  - If I (give patients counseling in addition to medication) then (their overall depression scale will decrease).
  - If I (give exams at noon instead of 7) then (student test scores will improve).
  - If I (look in this certain location) then (I am more likely to find new species).

A good hypothesis statement should:

- Include an "if" and "then" statement.
- Include both the independent and dependent variables.
- Be testable by experiment, survey or other scientifically sound technique.
- Be based on information in prior research (either yours or someone else's).
- Have design criteria (for engineering or programming projects).

6.2 Criteria for Hypothesis Construction

A hypothesis should meet the following criteria:

- It should be empirically testable, whether it is right or wrong.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- It should be specific and precise.
- The statements in the hypothesis should not be contradictory.
- It should specify variables between which the relationship is to be established.
- It should describe one issue only.

The slide lists the criteria that a hypothesis statement should meet.

6.3 Steps in Hypothesis Testing

The following are the steps for constructing a hypothesis test.

1. Select the type of hypothesis
   a) Null Hypothesis (H0)
   b) Alternative Hypothesis (Ha or H1)
2. Establish Critical or Rejection region
3. Select the Suitable Test of significance or Test Statistic
4. Check whether the test involves one sample, two samples, or multiple samples?
5. Check whether two or more samples used are independent or related?
6. Is the measurement scale nominal, ordinal, interval, or ratio?
7. The choice of a probability distribution of a sample statistics is guided but the sample size n and the value of population standard deviation.
8. Formulate a Decision Rule to Accept Null Hypothesis.
   a) Accept H0 if the test statistic value falls within the area of acceptance.
9. Reject otherwise.

The steps involved in a hypothesis test are given in the slide. Hypothesis testing can be one of the most confusing aspects for students, mostly because before you can even perform a test, you have to know what your null hypothesis is. Often, those tricky word problems that you are faced with can be difficult to decipher. But it's easier than you think; all you need to do is:

- Figure out your null hypothesis,
- State your null hypothesis,
- Choose what kind of test you need to perform,
- Either support or reject the null hypothesis.

The first step is to specify the null hypothesis. For a two-tailed test, the null hypothesis is typical that

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

a parameter equals zero although there are exceptions. A typical null hypothesis is $\mu1 - \mu2 = 0$ which is equivalent to $\mu1 = \mu2$. For a one-tailed test, the null hypothesis is either that a parameter is greater than or equal to zero or that a parameter is less than or equal to zero. If the prediction is that $\mu1$ is larger than $\mu2$, then the null hypothesis (the reverse of the prediction) is $\mu2 - \mu1 \geq 0$. This is equivalent to $\mu1 \leq \mu2$.

The second step is to specify the α level which is also known as the significance level. Typical values are 0.05 and 0.01.

The third step is to compute the probability value (also known as the p-value). This is the probability of obtaining a sample statistic as different or more different from the parameter specified in the null hypothesis given that the null hypothesis is true.

Finally, compare the probability value with the α level. If the probability value is lower then you reject the null hypothesis. Rejecting the null hypothesis is not an all-or-none decision. The lower the probability value, the more confidence you can have that the null hypothesis is false. However, if your probability value is higher than the conventional α level of 0.05, most scientists will consider your findings inconclusive. Failure to reject the null hypothesis does not constitute support for the null hypothesis. It just means you do not have sufficiently strong data to reject it.

6.4 Null Hypothesis

- If you trace back the history of science, the null hypothesis is always the accepted fact.
- Simple examples of null hypotheses that are generally accepted as being true are:
  - DNA is shaped like a double helix.
  - There are 8 planets in the solar system (excluding Pluto).

The statistical procedure for testing a hypothesis requires some understanding of the null hypothesis. Think of the outcome (dependent variable). From a statistical (and sampling) perspective), the null hypothesis asserts that the samples being compared or contrasted are drawn from the same population with regard to the outcome variable. This means that

- any observed differences in the dependent variable (outcome) must be due to sampling error (chance)
- the independent (predictor) variable does not make a difference

The symbol H0 is the abbreviation for the null hypothesis, the small zero stands for null.

Stating a null hypothesis:

You won't be required to actually perform a real experiment or survey in elementary statistics (or even disprove a fact like "Pluto is a planet"!), so you'll be given word problems from real-life situations.

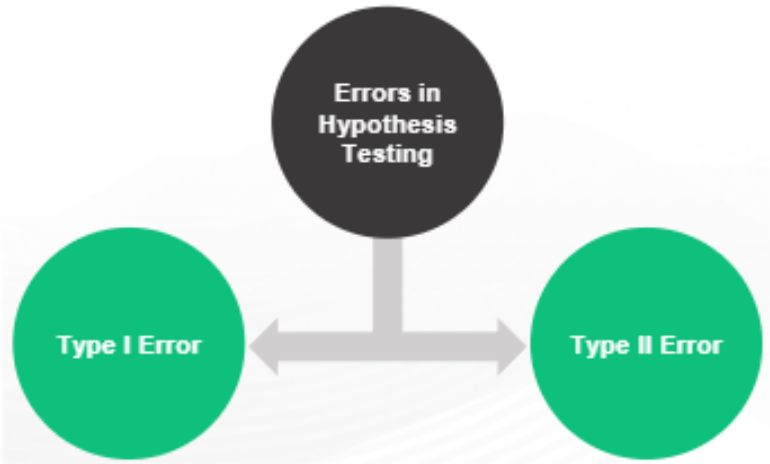
You'll need to figure out what your hypothesis is from the problem. This can be a little trickier than just figuring out what the accepted fact is. With word problems, you are looking to find a fact that is mollifiable (i.e. something you can reject).

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

6.5 Errors in Hypothesis Testing

Two types of hypothesis testing:



There are two types of errors in hypothesis testing.

1. Type I error
2. Type II error

6.5.1 Type I Error

Types I Error:

- It is also known as an error of the first kind, occurs when the null hypothesis (H0) is true, but is rejected.
- A type I error may be compared with a so called false positive.
- A Type I error occurs when we believe a falsehood.
- The rate of the type I error is called the size of the test and denoted by the Greek letter α (alpha).It usually equals the significance level of a test. If type I error is fixed at 5%, it means that there are about 5 chances in 100 that we will reject H0 when H0 is true.

More generally, a Type I error occurs when a significance test results in the rejection of a true null hypothesis.

By one common convention, if the probability value is below 0.05, then the null hypothesis is rejected. Another convention, although slightly less common, is to reject the null hypothesis if the probability value is below 0.01. The threshold for rejecting the null hypothesis is called the α (alpha) level or simply α. It is also called the significance level.

The Type I error rate is affected by the α level: the lower the α level, the lower the Type I error rate. It might seem that α is the probability of a Type I error. However, this is not correct. Instead, α is the probability of a Type I error given that the null hypothesis is true. If the null hypothesis is false, then it is impossible to make a Type I error.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

6.5.2 Type II Error

Types II Error:

- It is also known as an error of the second kind, occurs when the null hypothesis is false, but erroneously fails to be rejected.
- Type II error means accepting the hypothesis which should have been rejected. A type II error may be compared with a so-called False Negative.
- A Type II error is committed when we fail to believe a truth.
- A type II error occurs when one rejects the alternative hypothesis (fails to reject the null hypothesis) when the alternative hypothesis is true.
- The rate of the type II error is denoted by the Greek letter β (beta) and related to the power of a test (which equals 1-β).

The second type of error that can be made in significance testing is failing to reject a false null hypothesis. This kind of error is called a Type II error. Unlike a Type I error, a Type II error is not really an error. When a statistical test is not significant, it means that the data do not provide strong evidence that the null hypothesis is false. Lack of significance does not support the conclusion that the null hypothesis is true. Therefore, a researcher should not make the mistake of incorrectly concluding that the null hypothesis is true when a statistical test was not significant. Instead, the researcher should consider the test inconclusive. Contrast this with a Type I error in which the researcher erroneously concludes that the null hypothesis is false when, in fact, it is true.

A Type II error can only occur if the null hypothesis is false. If the null hypothesis is false, then the probability of a Type II error is called β (beta). The probability of correctly rejecting a false null hypothesis equals 1- β and is called power.

6.5.3 Error Scenarios in Hypothesis Testing

The following table summarizes the error scenarios in a hypothesis testing.

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative

Correct decisions:

- The null hypothesis is true and is accepted.
- The null hypothesis is false and is rejected.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Incorrect decisions:

Type I Error: The null hypothesis is true, but it is rejected. The probability of committing a type I error is denoted by α or p-value.

Type II Error: The null hypothesis is false, but it is accepted. The probability of committing type II error is denoted by β.

6.5.4 Type III Error

Types III Error:

- Many statisticians are now adopting a third type of error, type III, where the null hypothesis is rejected for the wrong reason.
- In an experiment, a researcher might assume a hypothesis and perform research. After analyzing the results, the null is rejected.
- The problem, that there may be some relationship between the variables, but it could be for a different reason than stated in the hypothesis.
- An unknown process may underlie the relationship.

A type III error is where you correctly reject the null hypothesis, but it's rejected for the wrong reason. This compares to a Type I error (incorrectly rejecting the null hypothesis) and a Type II error (not rejecting the null when you should). Type III errors are not considered serious, as they do mean you arrive at the correct decision. They usually happen because of random chance and are a rare occurrence.

You can also think of a Type III error as giving the right answer (i.e. correctly rejecting the null) to the wrong question. Either way, you're still arriving at the correct conclusion for the wrong reason. When we say the "wrong question", that normally means you've formulated your hypotheses incorrectly. In other words, both your null and an alternative hypothesis may be poorly worded or completely incorrect.

What did You Grasp?

1. State True or False.

   A type II error is one in which a false null hypothesis is not rejected.

   A) True
   B) False

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

2. Fill in the blank.

   The threshold for rejecting a null hypothesis is also called _____.

   A) Confidence level
   B) Significance level
   C) Hypothesis level
   D) None of the above

In a nutshell, we learnt:

1. The principles of counting including addition rule, product rule, permutation and combination.
2. Introduction and important definitions of Probability Theory.
3. The concept of Conditional Probability.
4. Bayes Theorem and its applications.
5. Different methods of discrete probability distribution like discrete uniform distribution, Poisson distribution, Bernoulli distribution and Binomial distribution.
6. The concepts of Covariance and Correlation.
7. Different methods of Continuous Probability Distribution, including normal distribution, continuous uniform distribution, exponential distribution, confidence intervals, t-distribution and chi-square distribution.
8. Properties of Central Limit Theorem with the help of examples.
9. Hypothesis Testing and the types of errors.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

MODULE 5

Linear Algebra

You will learn about the topic, 'Linear Algebra' in this module.

Module Learning Objectives

At the end of this module, you will be able to:

- Understand the introduction, history and the applications of linear algebra.
- Understand and apply the important notations in linear algebra.
- Describe the important concepts of linear algebra:
  - Scalars
  - Vectors
  - Vector spaces
  - Basis
  - Matrices
- Understand the important definitions on linear system of equations.

Module Topics

The following topics that will be covered in the module:

1. Introduction to Linear Algebra
2. Notations in Linear Algebra
3. Important Concepts of Linear Algebra
4. Definitions of Linear Algebra

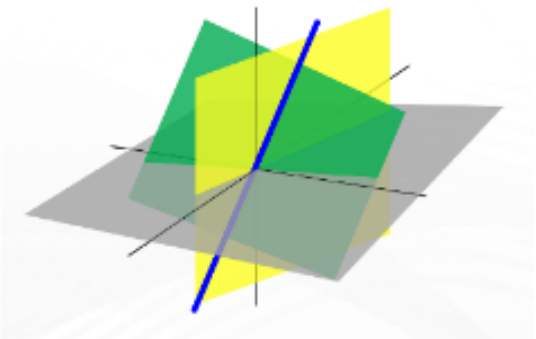
This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1 Introduction to Linear Algebra

1.1 Introduction to Linear Algebra

Linear Algebra:

- The branch of mathematics that involves the study of vector spaces and the linear transformations between them.
- It involves the study of linear equations and linear functions and their representation through matrices and vector spaces.
- To be precise, Linear algebra is the study of vectors and linear structures.
- Initial use cases were solving systems of linear equations.



Linear algebra, a branch of mathematics, is widely used in the field of science and engineering. Probability theory, linear algebra and matrices are the three important topics to be understood clearly for understanding and working with data processing and machine learning algorithms.

Linear algebra is very similar to the Algebra taught in the High school curriculum. The only difference is that linear algebra deals with vectors in place of ordinary single numbers, i.e., scalars. Many of the operations that you perform on scalars, such as addition, subtraction and multiplication, can be generalized to be performed on vectors.

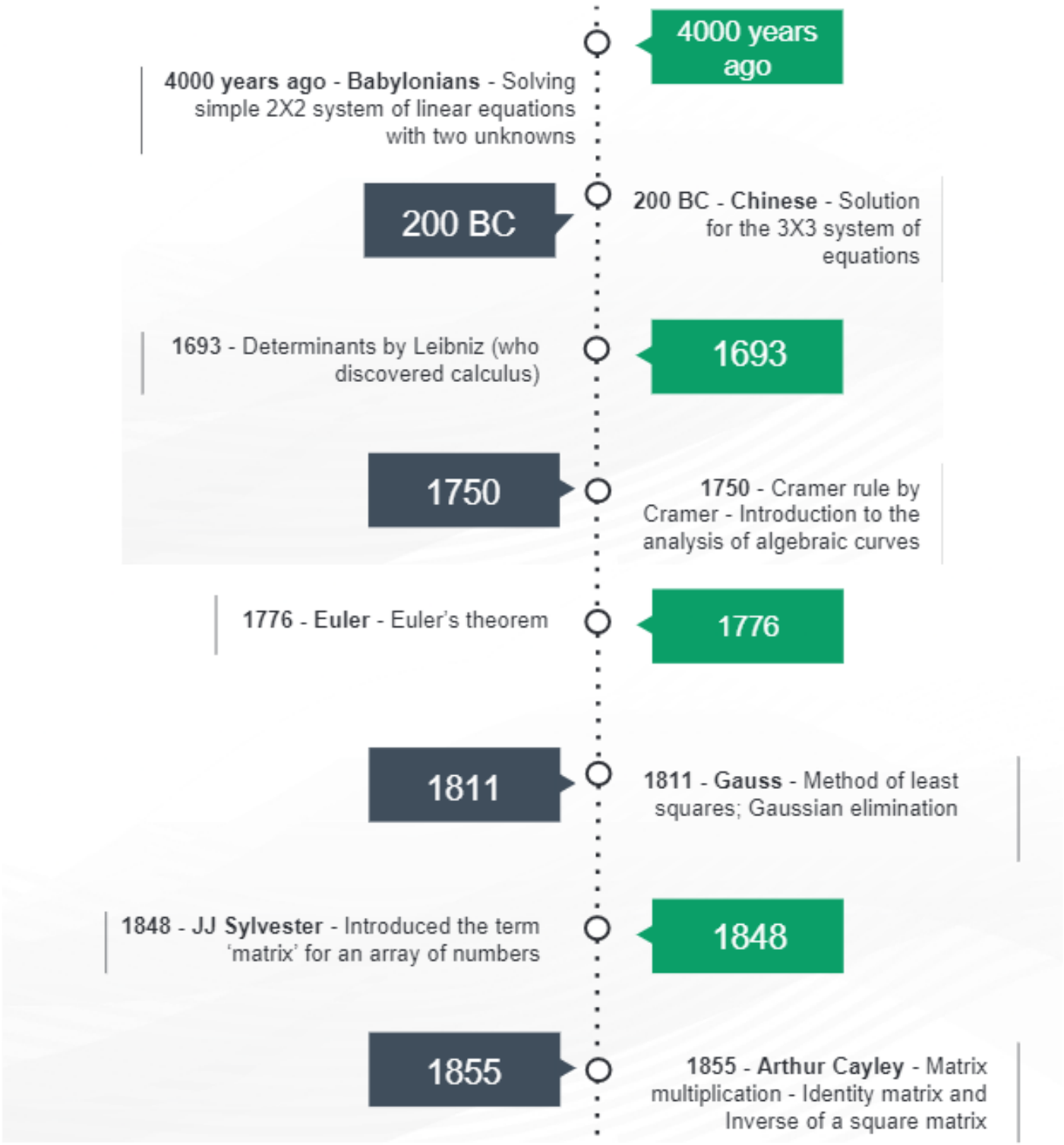
Thus Linear algebra is the study of vectors and the linear transformations between them.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1.2 History of Linear Algebra

The following timelines illustrate the history of Linear Algebra:



This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

The foundations of linear algebra were laid long back by Babylonians 4000 years ago, as they knew how to solve a simple 2X2 system of linear equations with two unknowns. Around 200 BC, the Chinese published that "Nine Chapters of the Mathematical Art," they displayed the ability to solve a 3X3 system of equations. The power and progress in linear algebra did not come to fruition until the late 17th century.

The modern study of systems of linear equations can be said to have originated with Leibniz, who in 1693 invented the notion of a determinant for this purpose. But his investigations remained unknown at the time. In 1750 Cramer published *Introduction to the Analysis of Algebraic Curves*, in this work he published a rule after him, Cramer's rule, for the solution of an nXn system, but he provided no proofs.

It was in the 1770s that Euler was perhaps the first to observe that a system of n equations in n unknowns does not necessarily have a unique solution, noting that to obtain uniqueness it is necessary to add conditions. He had in mind the idea of the dependence of one equation on the others, although he did not give precise conditions. In the eighteenth century the study of linear equations was usually subsumed under that of determinants, so no consideration was given to systems in which the number of equations differed from the number of unknowns. In 1776, Euler's theorem was proposed.

In 1811, Gauss, who invented the method of least squares, introduced a systematic procedure, now called Gaussian elimination, for the solution of systems of linear equations, though he did not use the matrix notation. Gaussian elimination uses the concepts of combining, swapping, or multiplying rows with each other in order to eliminate variables from certain equations. After variables are determined, the student is then to use back substitution to help find the remaining unknown variables.

In 1848, J.J. Sylvester introduced the term "matrix," the Latin word for womb, as a name for an array of numbers. Matrix multiplication or matrix algebra came from the work of Arthur Cayley in 1855. His work dealing with Matrix multiplication culminated in his theorem, the Cayley-Hamilton Theorem. Simply stated, a square matrix satisfies its characteristic equation. Cayley's efforts were published in two papers, one in 1850 and the other in 1858. His works introduced the idea of the identity matrix as well as the inverse of a square matrix.

1.3 Applications of Linear Algebra

Some of the disciplines where Linear algebra is applied:	Applications of Linear Algebra:
Mathematics	Oil exploration
Engineering	Electrical networks
Computer Science	Linear Programming
Physics	Aircraft design
Biology	Space flight and control systems
Economics	GPS navigation
Statistics	Multichannel image processing
	Computer graphics

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Some of the domains and the applications where linear algebra concepts are applied, are given in the slide. There is a saying, there is no domain where linear algebra concepts are not applied.

Linear algebra forms the centre of mathematics around which most of the mathematical methods revolve around. Linear algebra forms the basis of many science and engineering disciplines, as it is used in deriving the models of many naturally occurring phenomena and performing efficient computing with these models.

What did You Grasp?

1. Who introduced the method of least squares?

   A) Euler
   B) Cramer
   C) Gauss
   D) JJ Sylvester

2 Notations in Linear Algebra

2.1 Linear Algebra Notations

Important notations of linear algebra are listed in the table below.

$\mathbb{R}, \mathbb{R}^+, \mathbb{R}^n$	real numbers, reals greater than 0, ordered n-tuples of reals
$\mathbb{N}$	natural numbers: $\{0, 1, 2, \ldots\}$
$\mathbb{C}$	complex numbers
$\{\ldots \mid \ldots\}$	set of . . . such that . . .
$(a \ldots b), [a \ldots b]$	interval (open or closed) of reals between a and b
$\langle \ldots \rangle$	sequence; like a set but order matters
V, W, U	vector spaces
$\vec{v}, \vec{w}$	vectors

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

$M \oplus N$	direct sum of subspaces
$V \cong W$	isomorphic spaces
h, g	homomorphisms, linear maps
H, G	matrices
t, s	transformations; maps from a space to itself
T, S	square matrices

$\vec{0}, \vec{0}_V$	zero vector, zero vector of V
B, D	bases
$\mathcal{E}_n = \langle \vec{e}_1, \ldots, \vec{e}_n \rangle$	standard basis for $\mathbb{R}^n$
$\vec{\beta}, \vec{\delta}$	basis vectors
$\mathrm{Rep}_B(\vec{v})$	matrix representing the vector
$\mathcal{P}_n$	set of n-th degree polynomials
$\mathcal{M}_{n \times m}$	set of $n \times m$ matrices
$[S]$	span of the set S

$\mathrm{Rep}_{B,D}(h)$	matrix representing the map h
$h_{i,j}$	matrix entry from row i, column j
$\lvert T \rvert$	determinant of the matrix T
$\mathcal{R}(h), \mathcal{N}(h)$	rangespace and nullspace of the map h
$\mathcal{R}_\infty(h), \mathcal{N}_\infty(h)$	generalized rangespace and nullspace

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

The important notations and their explanation in Linear algebra are given in the table. The knowledge of these notations is important as you will come across the notations in many of the upcoming sections.

What did You Grasp?

1. Which of the following options is a correct notation for a vector space?

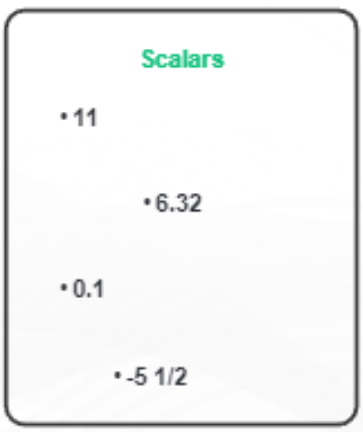
   A) V
   B) v
   C) V
   D) None of the above

3 Important Concepts of Linear Algebra

3.1 Scalars

Scalars:

- A single number or an element of a field, which forms the basis of a vector space.
- Refer to real numbers or other elements of a field.
- Have magnitude only, but not direction.
- Are denoted by lowercase italics.
- The kind or number of scalars is specified when scalars are introduced.

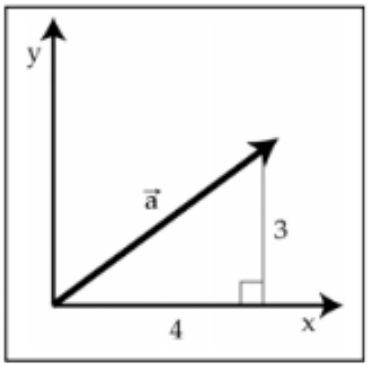


To be precise, a scalar is a number. Examples of scalars are temperature, distance, speed, or mass – all quantities that have a magnitude but no "direction", other than perhaps positive or negative.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

3.2 Vectors

Vectors:



- Are arrays of objects, which have magnitude and direction.
- Are not only stacks of numbers, but they also include, but not limited to polynomials, power series, functions, etc.
- Should be added if these are of the same kind.
- Are things that you can add and scalar multiply.

A very special vector can be produced from any vector of any kind by scalar multiplying any vector by the number 0. This is called the zero vector.

We saw that Linear algebra is a study of vectors and linear functions.
We'll now see what 'Vectors' are. Vectors are arrays of objects, but they are not constrained to numbers. Vectors are arrays or list of objects that have both magnitude and direction. Things that can be added and scalar multiplied, can be termed as vectors.

There are (at least) two ways to interpret vectors: One way to think of the vector as being a point in a space. Then vectors are a way of identifying that point in space, where each number represents the vectors component in that dimension. Another way to think of a vector is a magnitude and a direction, e.g. a quantity like 'velocity' ("the fighter jet's velocity is 250 mph north-by-northwest"). In this way, a vector is a directed arrow pointing from the origin to the endpoint given by the list of numbers

An object in R (refer to real numbers in notations section), or in any Rn, comprised of a magnitude and a direction is a vector.

The different kinds of vectors are:

- numbers
- n-vectors
- 2nd order polynomials
- polynomials
- power series
- functions with a certain domain

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

3.3 Vector Spaces

A vector space is a set that is closed under addition and scalar multiplication and is denoted as V.

A *vector space* (over $\mathbb{R}$) consists of a set V along with two operations '+' and '·' subject to the conditions that for all vectors $\vec{v}, \vec{w}, \vec{u} \in V$ and all *scalars* $r, s \in \mathbb{R}$:

(1) the set V is closed under vector addition, that is, $\vec{v} + \vec{w} \in V$
(2) vector addition is commutative, $\vec{v} + \vec{w} = \vec{w} + \vec{v}$
(3) vector addition is associative, $(\vec{v} + \vec{w}) + \vec{u} = \vec{v} + (\vec{w} + \vec{u})$
(4) there is a *zero vector* $\vec{0} \in V$ such that $\vec{v} + \vec{0} = \vec{v}$ for all $\vec{v} \in V$
(5) each $\vec{v} \in V$ has an *additive inverse* $\vec{w} \in V$ such that $\vec{w} + \vec{v} = \vec{0}$
(6) the set V is closed under scalar multiplication, that is, $r \cdot \vec{v} \in V$
(7) addition of scalars distributes over scalar multiplication, $(r+s)\cdot\vec{v} = r\cdot\vec{v}+s\cdot\vec{v}$
(8) scalar multiplication distributes over vector addition, $r\cdot(\vec{v}+\vec{w}) = r\cdot\vec{v}+r\cdot\vec{w}$
(9) ordinary multiplication of scalars associates with scalar multiplication, $(rs)\cdot\vec{v} = r\cdot(s\cdot\vec{v})$
(10) multiplication by the scalar 1 is the identity operation, $1\cdot\vec{v} = \vec{v}$.

A vector space is one that has a set of vectors, a set of scalars and a scalar multiplication operation that takes a scalar k and a vector **v** to another vector *k***v**.

The definition above involves two kinds of addition and two kinds of multiplication. In condition 7, the '+' on the left is addition of two real numbers while the '+' on the right is addition of two vectors in ***V***. r and s are real numbers so 'r + s' can only mean real number addition. In the same way, in condition 9, the left side 'rs' is ordinary real number multiplication, while its right side 's v' is the scalar multiplication defined for this vector space.

For any vector space, a ***subspace*** is a subset that is itself a vector space, under the inherited operations. The ***span*** (or linear closure) of a nonempty subset S of a vector space is the set of all linear combinations of vectors from S. The span of the empty subset of a vector space is its trivial subspace.

Some important definitions are:

Let **v***1*, **v**2, · · · **v***k* be vectors in a vector space ***V*** , then the span of these vectors is the set of all possible linear combinations of **v***1*, **v**2, · · · **v***k*. This set is denoted by Span{**v***1*, **v**2, · · · **v***k*}.

Let *S* be a subspace of a vector space ***V***. We say that **v***1*, **v**2, · · · vk form a spanning set for *S* if *S* = Span{**v***1*, **v**2, · · · **v***k*}.

A set of vectors {**v***1*, **v**2, · · · **v***k*} in a vector space ***V*** is said to be linearly dependent if there exist scalars, *c1*, *c2*, ... *ck* not all zero, such that *c1***v1** + *c2***v2** + · · · + *ck***vk** = **0**.

A set of vectors {**v***1*, **v**2, · · · **v***k*} in a vector space ***V*** is said to be linearly independent if *c1***v1** + *c2***v2** + · · · + *ck***vk** = **0** implies that all the coefficients *c1*, *c2*, ... *ck* are zero.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

3.4 Basis

- A basis for a vector space is a sequence of vectors that is linearly independent and that spans the space.
- For any

$$\mathcal{E}_n = \langle \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \rangle$$

is the standard (or natural) basis, and we denote these vectors, $\vec{e}_1, \ldots, \vec{e}_n$

- In any vector space, a subset is a basis, if and only if each vector in the space can be expressed as a linear combination of elements of the subset in one and only one way.

A set of vectors {**v***1*, **v**2, · · · **v***k*} in a vector space ***V*** is said to be a basis for a subspace *S* if it is a linearly independent spanning set for *S*.

A basis is a sequence, meaning that bases are different if they contain the same elements but in different orders, we denote it with angle brackets. A sequence is linearly independent if the multiset consisting of the elements of the sequence is independent. Similarly, a sequence spans the space if the set of elements of the sequence spans the space.

The dimension of a vector space *S* is equal to the number of vectors in a basis for *S*.

3.5 Matrices

- A matrix, like a vector, is an array of numbers, the difference is that matrix is an organized table of numbers instead of just being a list.
- Suppose that m and n are positive integers. An m x n matrix is a rectangular array of numbers with m rows and n columns. Matrices are usually denoted by capital letters.
- For example:

$$A = \begin{bmatrix} 8 & 10 & -6 & -8 \\ -5 & 7 & 6 & -6 \\ 0 & 5 & -10 & 1 \\ 1 & -3 & -10 & 5 \\ -4 & -8 & -1 & -10 \end{bmatrix}$$

is a 5 x 4 matrix.

Matrices are the result of organizing information related to linear functions. A Matrix is a 2-D array of numbers, so each element is identified by two indices instead of just one. We usually give matrices,

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

upper-case variable names with bold typeface, such as **A**. If a real-valued matrix **A** has a height of m (row) and a width of n (column), then we say that $A \in R^{m \times n}$.

What did You Grasp?

1. State True or False.

   Velocity of wind is a scalar.

   A) True
   B) False

2. State True or False.

   A basis has set of vectors that are linearly dependent.

   B) True
   C) False

4 Definitions of Linear Algebra

4.1 Definitions on Linear Equations

1. A linear equation in the variables x_1, x_2, ... xn is an equation of the form

$$a_1x_1 + a_2x_2 + \ldots\ldots + a_nx_n = b,$$

where a_1, a_2, ... an and b are real or complex numbers (that are often given in advance).

2. If m and n are positive integers, a system of linear equations in the variables x_1, x_2, ... x_n is a set of equations of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned}$$

where a_{ij} (i = 1, ... m, j = 1, ... n) and bi (i = 1, ... m) are real or complex numbers (that are often given in advance). This system of linear equations is said to consist of m equations with n unknowns.

Example for definition 1:

$4x + 4y - 7z + 8w = 8$ is a linear equation in the variables x, y, z and w. On the other hand, $3x + 5y^2 = 6$ is not a linear equation (because of the presence of y^2).

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Example for definition 2:

$2x_1 + 2x_2 - 5x_3 + 4x_4 - 2x_5 = 4$

$-4x_1 + x_2 - 5x_4 = -1$

$3x_1 - 2x_2 + 4x_4 - 2x_5 = -4$

is a system of three linear equations in five unknowns. (The "unknowns" or "variables" of the system are x_1, x_2, x_3, x_4, x_5).

3. The solution set of a system of linear equations is the set of all solutions of the system.
4. Two linear systems are said to be equivalent to each other if they have the same solution set.
5. A linear system is said to be consistent if it has at least one solution and is otherwise said to be inconsistent.

Definitions 3 and 4 are self-explanatory.

Example for definition 5:

The linear system

$2x_1 + 2x_2 - 5x_3 + 4x_4 - 2x_5 = 4$

$-4x_1 + x_2 - 5x_4 = -1$

$3x_1 - 2x_2 + 4x_4 - 2x_5 = -4$

is consistent because it has at least one solution. (It in fact has infinitely many solutions.)

However, the linear system

$x + 2y = 5$

$x + 2y = 9$

is obviously inconsistent.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

What did You Grasp?

1. When do two linear systems become equivalent?

   A) When they have unknowns.
   B) If they have the same solution set.
   C) If they have real numbers as solution.
   D) None of the above.

In a nutshell, we learnt:

1. The introduction, history and applications of Linear algebra.
2. The important notations of Linear algebra.
3. Important concepts in Linear algebra:
   - Scalars
   - Vectors
   - Vector spaces
   - Basis
   - Matrices
4. Important definitions on linear system of equations.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

MODULE 6

Matrices

You will learn about 'Matrices' in this module.

Module Learning Objectives

At the end of this module, you will be able to:

- Give an Introduction to Matrices along with the basic terminologies.
- Explain Matrix notations and types of matrices.
- Understand Matrix equality.
- Perform operations on matrices, like addition, subtraction and multiplication.
- Describe determinants.
- Understand the singularity of a Matrix.
- Explain orthogonal Matrix.
- Understand elementary transformations and elementary matrices and Matrix. inverse computation using elementary transformations.
- Describe echelon forms and echelon transformations.
- Find Matrix rank and reduce a Matrix to its normal form.
- Explain Vector Spaces and enumerate the axioms.
- Describe linear dependence and independence of vectors.
- Understand the consistency of linear system of equations.
- Describe eigenvalues and eigenvectors.
- State and prove Cayley Hamilton Theorem.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- Explain linear transformation and orthogonal transformation.
- Explain Matrix factorization and enumerate a few types such as LU, QR and SVD decomposition.

Module Topics

The following topics that will be covered in the module:

1. Introduction to Matrices
2. Matrix Notations and Types
3. Matrix Equality
4. Operations on Matrices
5. Determinants
6. Singularity of a Matrix
7. Orthogonal Matrix
8. Elementary Transformations and elementary matrices
9. Echelon forms and echelon transformations
10. Matrix Rank and Normal Form of a matrix
11. Vector Spaces and the axioms
12. Linear Dependence and Independence of vectors
13. Consistency of linear system of equations
14. Eigenvalues and eigenvectors
15. Cayley Hamilton Theorem
16. Linear Transformation and Orthogonal transformation
17. Matrix Factorization and Types

1. Introduction to Matrices

1.1 Introduction to Matrices

Matrix:

- A collection of numbers arranged into a fixed number of rows and columns.
- Usually the numbers are real numbers, but a Matrix can also contain complex numbers.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- A rectangular array of elements, where horizontal arrangements are called rows and vertical arrangements are columns.
- Order or dimension of a Matrix is defined as: Number of rows X Number of columns
- In the example, Matrix A has 2 rows and 3 columns, and the order of the Matrix is 2 X 3.

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix}$$

A Matrix is a rectangular array of numbers arranged in rows and columns. The order of a matrix also called the dimension of a Matrix refers to the number of rows and columns in a matrix. While specifying a matrix, the general way is to list the rows first and the columns next. Thus the order of a Matrix is represented as Number of rows X Number of columns.

The numbers present in the rows and columns of a Matrix are called elements of the matrix. In the example above, the element in the first column of the first row is 1; the element in the second column of the first row is 2, etc.

1.2 Matrix Notation

- Matrices are identified by numbers and symbols.
- Matrices are represented by boldface letters.
- Matrix $\mathbf{A}_{ij}$ refers to an i X j Matrix ('i' rows and 'j' columns)
- Matrix elements take the notation $\mathbf{A}_{ij}$, where 'i' is the row number and 'j' is the column number.

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \end{bmatrix}$$

- Boldface, capital letters are used to represent matrices, italic capital letters refer to Matrix elements, and subscripts reveal Matrix dimension. For example,
- **A** and **X** refer to matrices A and X, respectively.
- A_{ij} refers to the element in row i and column j of Matrix **A**.
- $\mathbf{A}_{ij}$ refers to an i x j Matrix **A**.

Notations for special matrices are listed below:

- **A'** refers to the transpose of Matrix **A**.
- **I** refers to an identity matrix.
- $\mathbf{I}_n$ refers to an n x n identity matrix.

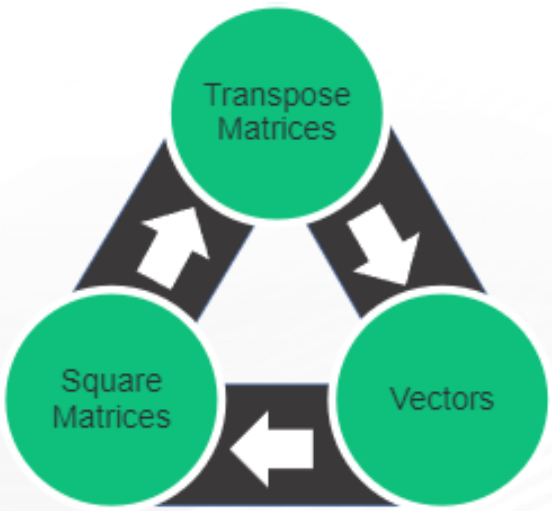
This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- **1** refers to the sum vector, a column vector having all of its elements equal to one.
- $\mathbf{1}_n$ is a 1 x n sum vector.
- |**A**| refers to the determinant of Matrix A.
- **x** refers to a Matrix of deviation scores derived from the raw scores of Matrix X.

2. Matrix Notations and Types

2.1 Types of Matrices

There are three important types of matrices:



There are three important types of matrices.

- Transpose matrices
- Vectors
- Square matrices

We will see each of these in detail in the upcoming sections.

2.2 Transpose Matrix

Transpose of a Matrix:

- The Matrix that is obtained by using the rows of the first Matrix as columns of the second matrix.
- Transpose of Matrix A is represented as A' (A-prime).

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Example:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix} \qquad A' = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{pmatrix}$$

As given in the slide, transpose of a Matrix is defined as the Matrix that is obtained by using the rows of the first Matrix as columns of the second matrix.

Transpose of Matrix **A** is represented as **A'**.

Row 1 of Matrix **A** becomes column 1 of **A'**; row 2 of A becomes column 2 of **A'**; and so on. The order of a Matrix will be reversed after it is transposed. For example, the transpose of a 2X3 Matrix will be a 3X2 matrix, since rows and columns are interchanged.

2.3 Vectors

Vectors:

- Matrices that have only one column or one row.
- Has two categories:
  - Column vectors
  - Row vectors

Example: $\boldsymbol{a}' = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$

Here a is a column vector and its transpose a' is a row vector.

Vector is a type of Matrix that has only one column or one row. There are two types of vectors: row vectors and column vectors.

Lower-case, boldface letters are used to represent column vectors. The transpose of a column vector is a row vector, we use lower-case, boldface letters plus a prime to represent row vectors. In the example above, vector **a** would be a column vector, and vector **a'** would be a row vector.

2.4 Square Matrices

- A Matrix is said to be a square matrix, if the number of rows and columns are equal.
- The order of a square Matrix in n x n. Example:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 5 & 4 \\ 4 & 5 & 8 \end{pmatrix}$$

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- Different types of square matrices:

Symmetric matrix	Diagonal matrix	Scalar matrix
$A = A' = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$	$A = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$	$A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$

A square Matrix is denoted by n x n, where the number of rows and columns are equal. Some of the types of square matrices are:

1. Symmetric matrix: a symmetric Matrix is one whose transpose is equal to itself.
2. Diagonal matrix: a diagonal Matrix is also a type of symmetric matrix, where the off-diagonal elements are zeroes (diagonal refers to the elements that run from the upper left corner to the lower right corner).
3. Scalar matrix: scalar Matrix is also a diagonal matrix, where the diagonals are equal valued numbers.

Examples of these types are given above.

What did You Grasp?

1. Which of the following options is the transpose of $A = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 6 & 7 \end{bmatrix}$?

A) $A' = \begin{bmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \end{bmatrix}$

B) $A' = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 6 & 7 \end{bmatrix}$

C) $A' = \begin{bmatrix} 5 & 6 & 7 \\ 1 & 2 & 3 \end{bmatrix}$

D) $A' = \begin{bmatrix} 1 & 6 \\ 5 & 2 \\ 3 & 7 \end{bmatrix}$

3. Matrix Equality

3.1 Matrix Equality

Two matrices are said to be equal if they satisfy the following conditions:

- Each Matrix has the same number of rows.
- Each Matrix has the same number of columns.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- Corresponding elements within each Matrix are equal.

Consider the three matrices **A**, B and C.

$$\mathbf{A} = \begin{bmatrix} 111 & x \\ y & 444 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 111 & 222 \\ 333 & 444 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} l & m & n \\ o & p & q \end{bmatrix}$$

The criteria for two matrices being equal is given above.

If **A** = B, then x = 222 and y = 333; since corresponding elements of equal matrices are also equal. Matrix C is not equal to **A** or B, since it has more columns than **A** or B.

3.2 Identity Matrices

- The identity Matrix is an n x n diagonal Matrix with 1's in the diagonal and zeros everywhere else.
- The identity Matrix is denoted by I or In. Two identity matrices are given below.

$$\mathbf{I_2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mathbf{I_3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Any Matrix that is pre or post-multiplied by I remains the same.

i.e , **AI = IA = A**

An identity Matrix has 1's in the diagonals and all other elements are 0's. Examples for identity Matrix is given above.

What did You Grasp?

1. What is the order of the following matrix?

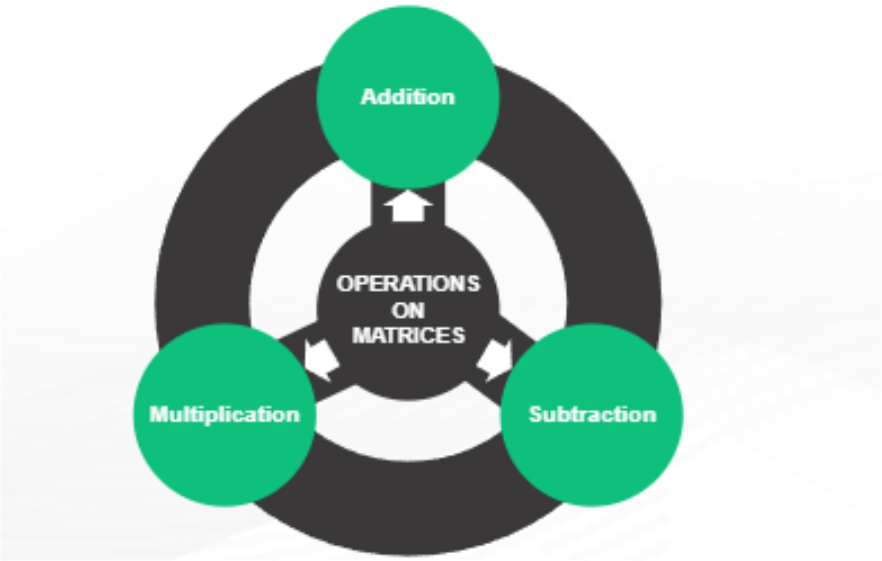
$$X = \begin{bmatrix} 1 & 2 & 4 & 4 \\ 3 & 4 & 8 & 0 \end{bmatrix}$$

A) 2x2
B) 2x3
C) 2x4
D) 1x4

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

3.3 Operations on Matrices

Three major operations can be done on matrices:



Three major operations can be performed on matrices.

- Addition and Subtraction
- Multiplication

We'll see how to perform these operations on matrices, in the upcoming sections.

3.3.1 Matrix Addition and Subtraction

- Matrices can be added or subtracted if they have the same order, i.e., the same number of rows and columns.
- Addition or subtraction can be performed by adding or subtracting the corresponding elements.

For example,

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 6 & 7 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 3 & 4 & 5 \\ 7 & 5 & 6 \end{bmatrix}$$

$$\boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} 1+3 & 2+4 & 3+5 \\ 5+7 & 6+5 & 7+6 \end{bmatrix} \qquad \boldsymbol{A} - \boldsymbol{B} = \begin{bmatrix} 1-3 & 2-4 & 3-5 \\ 5-7 & 6-5 & 7-6 \end{bmatrix}$$

$$\boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} 4 & 6 & 8 \\ 12 & 11 & 13 \end{bmatrix} \qquad \boldsymbol{A} - \boldsymbol{B} = \begin{bmatrix} -2 & -2 & -2 \\ -2 & 1 & 1 \end{bmatrix}$$

Like any addition or subtraction, matrices can also be added or subtracted. Two matrices can be added or subtracted only if they have the same number of rows and columns, i.e., if they have

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

the same order. Matrices can be added or subtracted by adding or subtracting the corresponding elements in each of the matrices.

An example of Matrix addition and subtraction is given above. From the example we can also understand that the order in which the elements are added is not important, i.e., A + B = B + A. On the other hand, it is also evident that A - B ≠ B - A.

3.3.2 Matrix Multiplication

There are two types of Matrix multiplication:

- Multiplication of a Matrix by a number (scalar)
- Multiplication of a Matrix by another matrix.

The result of both the operations is a Matrix by itself.

Like addition and subtraction, multiplication can also be performed on matrices. There are two types of Matrix multiplication:

- Multiplying a Matrix by a number or scalar.
- Multiplying a Matrix by another matrix.

3.3.2 (a) Multiplication of a Matrix by a Number

- Multiplication of a Matrix by a number involves multiplying each element of the Matrix by that number.
- The result of this multiplication is a matrix, which is called a scalar multiple.

Example:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 6 & 7 \end{bmatrix} \qquad 5\mathbf{A} = 5\begin{bmatrix} 1 & 2 & 3 \\ 5 & 6 & 7 \end{bmatrix}$$

$$5\mathbf{A} = \begin{bmatrix} 5*1 & 5*2 & 5*3 \\ 5*5 & 5*6 & 5*7 \end{bmatrix}$$

$$5\mathbf{A} = \begin{bmatrix} 5 & 10 & 15 \\ 25 & 30 & 35 \end{bmatrix}$$

As mentioned above, multiplying a Matrix by a number results in a scalar multiple. This multiplication is also referred to as multiplication of a Matrix by a scalar.

In the above example, every element of the Matrix A is multiplied by the scalar 5 to produce the scalar multiple, which is also a matrix.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

3.3.2 (b) Multiplication of Matrix by a Matrix

- While multiplying two matrices A and B, the product AB can be defined only if the number of columns in A is equal to the number of rows in B.
- Similarly, the Matrix product BA can be defined only if the number of columns in B is equal to the number of rows in A.
- If Matrix A has order m x n and Matrix B has order p x q,
  - AB exists only if n = p (order of AB will be m x q)
  - BA exists only if q = m (order of BA will be p x n)

Multiplication of Matrix by a Matrix should satisfy the conditions as stated above. The method of computing the Matrix product is explained in the next section.

Multiplication order:

Matrix multiplication is possible in both directions, results may be different. That is the product of two matrices **AB** is not always equal to **BA**. To describe the Matrix product **AB**, we can say **A** is post-multiplied by **B**; or we can say that **B** is premultiplied by **A**. Similarly, to describe the Matrix product **BA**, we can say **B** is post-multiplied by **A**, or we can say that A is premultiplied by **B**.

3.3.3 Computing the Product of Matrix Multiplication by a Matrix

- Suppose that **A** is an i x j matrix, and **B** is a j x k matrix. Then, the Matrix product **AB** results in a Matrix **C**, which has i rows and k columns; and each element in **C** can be computed according to the following formula:

$$C_{ik} = \Sigma_j A_{ij} B_{jk}$$

- *C_{ik} = the element in row i and column k from Matrix C*
- *A_{ij} = the element in row i and column j from Matrix A*
- *B_{jk} = the element in row j and column k from Matrix B*
- *Σ_j = indicates that the A_{ij} B_{jk} terms should be summed over j*

Example:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 6 & 7 \\ 8 & 9 \\ 10 & 11 \end{bmatrix} \quad \mathbf{AB} = \mathbf{C} = \begin{bmatrix} 28 & 31 \\ 100 & 112 \end{bmatrix}$$

Consider the example given above.

Matrix **A** and **B** are multiplied to get the product **C**.

Let **AB** = **C**. **A** has 2 rows, so **C** will have two rows; and **B** has 2 columns, so **C** will have 2 columns.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

To compute the value of every element in the 2 x 2 Matrix C, we use the formula $C_{ik} = \Sigma_j A_{ij} B_{jk}$, as shown below.

$C_{11} = \Sigma A_{1j} B_{j1}$ = 0*6 + 1*8 +2*10 = 0 + 8 + 20 = **28**

$C_{12} = \Sigma A_{1j} B_{j2}$ = 0*7 + 1*9 +2*11 = 0 + 9 + 22 = **31**

$C_{21} = \Sigma A_{2j} B_{j1}$ = 3*6 + 4*8 +5*10 = = 18 + 32 + 50 = **100**

$C_{22} = \Sigma A_{2j} B_{j2}$ = 3*7 + 4*9 +5*11 = 21 + 36 +55 = **112**

The result is what you see in the above example, i.e., the Matrix product C.

3.4 Vector Multiplication

- Multiplication of a vector by a vector results two kinds of products:

Vector inner product

$$\mathbf{a} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

a'b = 1*4 + 2*5 + 3*6 = **4** + 10 + 18 = 32

Vector outer product

$$\mathbf{a} = \begin{bmatrix} v \\ w \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\mathbf{C} = \mathbf{ab'} = \begin{bmatrix} v*x & v*y & v*z \\ w*x & w*y & w*z \end{bmatrix}$$

Explanation of both these concepts is given below.

Multiplication of a vector by a vector may result in vector inner product or vector outer product.

Vector inner product

Vector inner product is also termed as scalar product or dot product. If **a** and **b** are vectors, each with the same number of elements, then, the inner product of **a** and **b** is s.

a'b = **b'a** = *s*

where

a and **b** are column vectors, each having n elements,

a' is the transpose of **a**, which makes **a'** a row vector,

b' is the transpose of **b**, which makes **b'** a row vector, and

s is a scalar, i.e., a real number and not a matrix.

In the above example, **a'b** = 1*4 + 2*5 + 3*6 = 4 + 10 + 18 = 32, the inner product of **a'b** is equal to **32**.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Vector outer product

Outer product is also called cross product. If **a** and **b** are vectors. Then, the outer product of **a** and **b** is C.

ab' = C

where

a is a column vector, having *m* elements,

b is a column vector, having *n* elements,

b' is the transpose of **b**, which makes **b'** a row vector, and

C is a rectangular *m x n* matrix

The outer product of two vectors produces a rectangular matrix, not a scalar, which is illustrated in the example above. The elements of Matrix C consist of the product of elements from vector **a** crossed with elements from vector **b**. Thus, C is a Matrix of cross products from the two vectors.

What did You Grasp?

1. State True or False.

   Outer vector product is also called dot product.

   A) True
   B) False

2. State True or False.

   For two matrices A and B, A - B = B - A.

   A) True
   B) False

3. Find the product AB, given that $A = \begin{bmatrix} 2 & 4 & 6 \\ 1 & 3 & 5 \end{bmatrix}$ $B = \begin{bmatrix} 0 & 1 \\ 3 & 0 \\ 2 & 5 \end{bmatrix}$

   A) $\begin{bmatrix} 10 & 13 \\ 16 & 25 \end{bmatrix}$

   B) $\begin{bmatrix} 24 & 32 \\ 19 & 26 \end{bmatrix}$

   C) $\begin{bmatrix} 24 & 19 \\ 32 & 26 \end{bmatrix}$

   D) $\begin{bmatrix} 12 & 19 \\ 26 & 35 \end{bmatrix}$

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

4. What is 4A, if $A = \begin{bmatrix} 2 & 3 \\ 6 & 5 \end{bmatrix}$

A) $\begin{bmatrix} 10 & 3 \\ 6 & 25 \end{bmatrix}$

B) $\begin{bmatrix} 8 & 12 \\ 24 & 20 \end{bmatrix}$

C) $\begin{bmatrix} 5 & 3 \\ 6 & 5 \end{bmatrix}$

D) $\begin{bmatrix} 8 & 24 \\ 12 & 20 \end{bmatrix}$

4. Determinants

- The determinant is a value that can be computed from the elements of a square matrix. The determinant of a Matrix A is denoted det(A), det A, or |A|.
- For a 2x2 matrix, the determinant can be defined as:

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix}$$
$$= ad - bc$$

- For a 3x3 Matrix A, the determinant is:

$$|A| = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} - c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

$$= aei + bfg + cdh - ceg - bdi - afh$$

Determinant is a unique number that is associated with a square matrix. We have explained how to compute the determinant for 2x2 and 3x3 matrices.

The formula for computing the determinant of any nxn square Matrix is as follows:

$|A| = \Sigma(\pm) A_{1q}A_{2r}A_{3s} \ldots A_{nz}$

Points to note in this formula:

- The determinant is the sum of product terms made up of elements from the matrix.
- Each product term consists of *n* elements from the matrix.
- Each product term includes one element from each row and one element from each column.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- The number of product terms is equal to *n*! (where *n*! refers to n factorial).
- By convention, the elements of each product term are arranged in ascending order of the left-hand (or row-designating) subscript.
- To find the sign of each product term, we count the number of inversions needed to put the right-hand (or column-designating) subscripts in numerical order. If the number of inversions is even, the sign is positive; if odd, the sign is negative.

5. Singularity of a Matrix

- If a square Matrix is said to be a singular matrix if its determinant is zero.
- In such a case, the Matrix will have no inverse.

Example:

For a 2x2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

ad - bc = 0

For a 3x3 matrix

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \quad a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} - c\begin{vmatrix} d & e \\ g & h \end{vmatrix} = 0$$

$$aei + bfg + cdh - ceg - bdi - afh = 0$$

If the determinant of a Matrix is 0 then the Matrix has no inverse and it is called a Singular Matrix. Examples are given above.

6. Singularity of a Matrix

6.1 Orthogonal Matrix

- If the square Matrix with real elements, $A \in R^{m \times n}$ is the Gram Matrix forms an identity matrix, then the Matrix is said to be an orthogonal matrix.
- Conditions for an orthogonal matrix:

1. $A^T = A^{-1}$
2. $AA^T = A^TA = I$

where, the rows of Matrix **A** are orthogonal.

The definition for orthogonal Matrix is given above. Let's now look at an example to understand the concept.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

6.2 Example for Orthogonal Matrix

Example: Identify if the given Matrix is orthogonal.

$$A = \begin{bmatrix} cosx & sinx \\ -sinx & cosx \end{bmatrix}$$

Solution: Check the first condition of orthogonal matrix.

$$A^T = \begin{bmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{bmatrix} \quad .....(1)$$

$$A^{-1} = \frac{adj(A)}{|A|}$$

$$= \frac{\begin{bmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{bmatrix}}{(\cos^2 x + \sin^2 x)}$$

$$= \frac{\begin{bmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{bmatrix}}{(1)}$$

$$A^{-1} = \begin{bmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{bmatrix} \quad .....(2)$$

The first step is to verify that the given Matrix satisfies the first condition of orthogonal matrix.

Check for the second condition.

$$AA^T = \begin{bmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{bmatrix}\begin{bmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{bmatrix}$$

$$= \begin{bmatrix} \cos^2 + \sin^2 x & -\cos x \sin x + \sin x \cos x \\ -\sin x \cos x + \cos x \sin x & \sin^2 x + \cos^2 x \end{bmatrix}$$

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

$$= I$$

$$A^T A = \begin{bmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{bmatrix}\begin{bmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{bmatrix}$$

$$= \begin{bmatrix} \cos^2 + \sin^2 x & \cos x \sin x - \sin x \cos x \\ \sin x \cos x - \cos x \sin x & \sin^2 x + \cos^2 x \end{bmatrix}$$

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

$$= I$$

The second step proves that the given Matrix satisfies the second condition of orthogonal matrix.

Thus the example Matrix given above satisfies both the conditions of the orthogonal matrix.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

What did You Grasp?

1. Select the correct option.

   A = $\begin{bmatrix} 5 & 3 \\ 6 & 5 \end{bmatrix}$ then what is |A|.

   A) 8
   B) 10
   C) 15
   D) 7

7. Orthogonal Matrix

7.1 Elementary transformations of matrices

- Elementary transformations, also called elementary operations, of a Matrix, are:
  - Rearrangement of two rows (or columns).
  - Multiplication of all row (or column) elements of a Matrix to some number, not equal to zero.
  - Multiplication of a row (or column) by a number, not equal to zero, and adding the result to another row.
- Operations performed on rows are called elementary row operations and operations performed on columns are called elementary column operations.

There are three elementary operations that can be performed on matrices. There are certain notations that describe elementary operations. They are listed below:

Row operations:

1. Interchange rows *i* and *j*: R_i <--> R_j

2. Multiply row *i* by *s*, where $s \neq 0$: sR_i --> R_i

3. Add s times row *i* to row *j*: $sR_i + R_j$ --> R_j

Column operations:

1. Interchange columns *i* and *j*: C_i <--> C_j

2. Multiply column *i* by *s*, where $s \neq 0$: sC_i --> C_i

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

3. Add s times column *i* to column *j*: $sC_i + C_j \text{ --> } C_j$

7.2 Elementary Operators

- An Elementary operation may be performed by Matrix multiplication, using square matrices called elementary operators.
- If you want to interchange rows 1 and 2 of Matrix **A**. To accomplish this, you could premultiply **A** by E to produce B.

$$R_1 \text{<-->} R_2 = \underset{\mathbf{E}}{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}} \underset{\mathbf{A}}{\begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}}$$

$$R_1 \text{<-->} R_2 = \begin{bmatrix} 0+2 & 0+4 & 0+6 \\ 0+1 & 0+3 & 0+5 \end{bmatrix}$$

$$R_1 \text{<-->} R_2 = \begin{bmatrix} 2 & 4 & 6 \\ 1 & 3 & 5 \end{bmatrix} = \mathbf{B}$$

In the above example, E is an elementary operator. It operates on **A** to produce the desired interchanged rows in B. We'll now see how to find E.

7.3 Elementary Row Operations

- To perform an elementary row operation on a A, an r x c matrix, there are two steps that have to be followed.
  a) To find E, the elementary row operator, apply the operation to an r x r identity matrix.
  b) To carry out the elementary row operation, premultiply A by E.
- Three elementary row operations:
  - Interchange two rows
  - Multiply a row by a number
  - Multiply a row and add it to another row

We learnt about the steps involved in elementary row operations. We'll now learn how to perform the three types of elementary row operations.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

7.3.1 Interchange of Two Rows

Step 1:

$$\underset{I_3}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \Rightarrow \underset{E}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}}$$

Step 2:

$$R_2 <--> R_3 = \underset{E}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}} \underset{A}{\begin{bmatrix} 0 & 1 \\ 2 & 3 \\ 4 & 5 \end{bmatrix}}$$

$$R_2 <--> R_3 = \begin{bmatrix} 1*0 + 0*2 + 0*4 & 1*1 + 0*3 + 0*5 \\ 0*0 + 0*2 + 1*4 & 0*1 + 0*3 + 1*5 \\ 0*0 + 1*2 + 0*4 & 0*1 + 1*3 + 0*5 \end{bmatrix}$$

$$R_2 <--> R_3 = \begin{bmatrix} 0 & 1 \\ 4 & 5 \\ 2 & 3 \end{bmatrix}$$

In the above example, we want to interchange the second and third rows of **A**, a 3 x 2 matrix.

Step 1: To create the elementary row operator E, we interchange the second and third rows of the identity Matrix $\mathbf{I_3}$.

Step 2: To interchange the second and third rows of **A**, we premultiply **A** by E.

7.3.2 Multiply a Row by a Number

Step 1:

$$\underset{I_2}{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}} \Rightarrow \underset{E}{\begin{bmatrix} 1 & 0 \\ 0 & 7 \end{bmatrix}}$$

Step 2:

$$7R_2 --> R_2 = \underset{E}{\begin{bmatrix} 1 & 0 \\ 0 & 7 \end{bmatrix}} \underset{A}{\begin{bmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \end{bmatrix}}$$

$$7R_2 --> R_2 = \begin{bmatrix} 1*0 + 0*3 & 1*1 + 0*4 & 1*2 + 0*5 \\ 0*0 + 7*3 & 0*1 + 7*4 & 0*2 + 7*5 \end{bmatrix}$$

$$7R_2 --> R_2 = \begin{bmatrix} 0 & 1 & 2 \\ 21 & 28 & 35 \end{bmatrix}$$

Suppose we want to multiply each element in the second row of Matrix A by 7. Assume A is a 2 x 3 matrix.

- **Step 1:** To create the elementary row operator E, we multiply each element in the second row of the identity Matrix I2 by 7.
- **Step 2:** To multiply each element in the second row of A by 7, we premultiply A by E.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

7.3.3 Multiply a Row and Add it to Another Row

Step 1:

$$\underset{\mathbf{I_2}}{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}} \Rightarrow \begin{bmatrix} 1 & 0 \\ 0+3*1 & 1+3*0 \end{bmatrix} \Rightarrow \underset{\mathbf{E}}{\begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}}$$

Step 2:

$$3R_1 + R_2 \text{-->} R_2 = \underset{\mathbf{E}}{\begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}} \underset{\mathbf{A}}{\begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix}}$$

$$3R_1 + R_2 \text{-->} R_2 = \begin{bmatrix} 1*0+0*2 & 1*1+0*3 \\ 3*0+1*2 & 3*1+1*3 \end{bmatrix}$$

$$3R_1 + R_2 \text{-->} R_2 = \begin{bmatrix} 0 & 1 \\ 2 & 6 \end{bmatrix}$$

For the above example, assume **A** is a 2 x 2 matrix. Suppose we want to multiply each element in the first row of **A** by 3, and we want to add that result to the second row of **A**.

Step 1: For this operation, creating the elementary row operator is a two-step process. First, we multiply each element in the first row of the identity Matrix $\mathbf{I_2}$ by 3. Next, we add the result of that multiplication to the second row of $\mathbf{I_2}$ to produce E.

Step 2: To multiply each element in the first row of A by 3 and add that result to the second row, we premultiply **A** by E.

7.4 Elementary Column Operations

- To perform an elementary column operation on **A**, an r x c matrix, follow the steps.
  a) To find E, the elementary column operator, apply the operation to an c x c identity matrix.
  b) To carry out the elementary column operation, postmultiply **A** by E.
- The process for performing an elementary column operation on an *r* x *c* Matrix is very similar to the process for performing an elementary row operation. The main differences are:
  1. To operate on the *r* x *c* Matrix **A**, the row operator E is created from an *r* x *r* identity matrix; whereas the column operator, E is created from an *c* x *c* identity matrix.
  2. To perform a row operation, **A** is premultiplied by E; whereas to perform a column operation, **A** is *post-multiplied* by E.

The steps involved in performing an elementary column operation are given above. It is also important to note that the performing an elementary column operation on an *r* x *c* Matrix is very similar to the process for performing an elementary row operation, though there exist a few differences.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

7.4.1 Interchange of Two Columns

Step 1:

$$\underset{I_2}{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}} \Rightarrow \underset{E}{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}$$

Step 2:

$$C_1 \text{<-->} C_2 = \underset{A}{\begin{bmatrix} 0 & 1 \\ 2 & 3 \\ 4 & 5 \end{bmatrix}} \underset{E}{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}$$

$$C_1 \text{<-->} C_2 = \begin{bmatrix} 0*0 + 1*1 & 0*1 + 1*0 \\ 2*0 + 3*1 & 2*1 + 3*0 \\ 4*0 + 5*1 & 4*1 + 5*0 \end{bmatrix}$$

$$C_1 \text{<-->} C_2 = \begin{bmatrix} 1 & 0 \\ 3 & 2 \\ 5 & 4 \end{bmatrix}$$

In the above example, suppose we want to interchange the first and second columns of **A**, a 3 x 2 matrix.

Step 1: To create the elementary column operator E, we interchange the first and second columns of the identity Matrix I_2.

Step 2: To interchange the first and second columns of **A**, we postmultiply **A** by E.

7.5 Elementary Matrices

A square Matrix E of order n is called an elementary Matrix if it is obtained by applying exactly one elementary row operation to the identity matrix, I_n.

Three types of elementary matrices:

$$(E_{ij})_{(k,\ell)} = \begin{cases} 1 & \text{if } k = \ell \text{ and } \ell \neq i, j \\ 1 & \text{if } (k,\ell) = (i,j) \text{ or } (k,\ell) = (j,i) \\ 0 & \text{otherwise} \end{cases}$$

$$(E_k(c))_{(i,j)} = \begin{cases} 1 & \text{if } i = j \text{ and } i \neq k \\ c & \text{if } i = j = k \\ 0 & \text{otherwise} \end{cases}$$

$$(E_{ij})_{(k,\ell)} \begin{cases} 1 & \text{if } k = \ell \\ c & \text{if } (k,\ell) = (i,j) \\ 0 & \text{otherwise} \end{cases}$$

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

As mentioned above, there are three types of elementary matrices:

1. E_{ij}, which is obtained by the application of the elementary row operation R_{ij} to the identity matrix, I_n. The (k, l)th entry of E_{ij} is given above in type 1.
2. $E_k(c)$, which is obtained by the application of the elementary row operation $R_k(c)$ to the identity matrix, In. The (i,j)th entry of $E_k(c)$ is given above in type 2.
3. $E_{ij}(c)$, which is obtained by the application of the elementary row operation $R_{ij}(c)$ to the identity matrix, I_n. The (k,l)th entry of $E_{ij}(c)$ is given above in type 3.

7.6 Inverse of a Matrix Using Elementary Transformations

- If **A** is an n x n matrix, the inverse of **A** is another n x n matrix, denoted $\mathbf{A}^{-1}$, such that

  $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$

  where $\mathbf{I}_n$ is the identity matrix.
- Three types of transformations can be used to find the inverse of a matrix:
  1. Multiplying a row by a constant
  2. Adding a multiple of another row
  3. Swapping two rows

While using the elementary transformation method to find the inverse of a matrix, our goal is to convert the given Matrix into an identity matrix, using elementary transformations.

There are two ways to determine if inverse of a Matrix exists:

1. **Determine the rank of the matrix:** The rank of a Matrix is a unique number associated with a square matrix. If the rank of an n x n Matrix is less than n, the Matrix does not have an inverse.
2. **Compute the determinant of the matrix:** The determinant, a unique number associated with a square matrix. When the determinant for a square Matrix is equal to zero, the inverse for that Matrix does not exist.

A square Matrix that has an inverse is called nonsingular or invertible; a square Matrix that does not have an inverse is called singular.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

7.6.1 Example for Finding Matrix Inverse

Consider the Matrix A. Find A^{-1}.

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 0 & -1 \\ 2 & 3 & -1 \end{bmatrix}$$

Step 1:

$$\begin{array}{c} \begin{bmatrix} 1 & 2 & 0 & 1 & 0 & 0 \\ 2 & 0 & -1 & 0 & 1 & 0 \\ 2 & 3 & -1 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{c} a \\ b \\ c \end{array}$$

Solution

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

Step 2:

$$\begin{bmatrix} 1 & 2 & 0 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 & -1 & 1 \\ 2 & 3 & -1 & 0 & 0 & 1 \end{bmatrix} \begin{array}{c} a \to a \\ -b + c \to b \\ c \to c \end{array}$$

$$\begin{bmatrix} 1 & 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1/3 & 1/3 \\ 0 & 1 & 1 & 2 & 0 & -1 \end{bmatrix} \begin{array}{c} a \to a \\ b/3 \to b \\ -c + 2a \to c \end{array}$$

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 2/3 & -2/3 \\ 0 & 1 & 0 & 0 & -1/3 & 1/3 \\ 0 & 0 & 1 & 2 & 1/3 & -4/3 \end{bmatrix} \begin{array}{c} a - 2b \to a \\ b \to b \\ c - b \to c \end{array}$$

Step 1 is to write the identity Matrix on the right side of A.

Step 2 is to apply elementary transformations to the whole matrix. The point at which you get an identity Matrix on the left side, you will get the inverse of A on the right side.

8. Elementary Transformations and elementary matrices

8.1 Echelon Matrices

Echelon matrices come in two forms:

Row echelon form

$$\underset{A_{ref}}{\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}} \quad \underset{B_{ref}}{\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}} \quad \underset{C_{ref}}{\begin{bmatrix} 1 & 2 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}}$$

Reduced row echelon form

$$\underset{A_{rref}}{\begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}} \quad \underset{B_{rref}}{\begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}} \quad \underset{C_{rref}}{\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}}$$

Row Echelon Form: A Matrix is said to be in row echelon form (ref) when it satisfies the following conditions

- The first non-zero element in each row, called the leading entry, is generally 1, but according to some references, this need not be met.
- Each leading entry is in a column to the right of the leading entry in the previous row.
- Rows with all zero elements, if any, are below rows having a non-zero element.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Example for a row echelon form Matrix is given above.

Reduced Row Echelon Form

A Matrix is in reduced row echelon form (rref) when it satisfies the following conditions.

- The Matrix satisfies conditions for a row echelon form.
- The leading entry in each row is the only non-zero entry in its column.

Example for reduced row echelon form is given above.

8.2 Echelon Transformation

- Any Matrix can be transformed into its echelon form, using elementary operations.
- Steps are as follows:
  a) Find the pivot, the first non-zero entry in the first column of the matrix.
  b) To get the Matrix in row echelon form, repeat the pivot
  c) To get the Matrix in reduced row echelon form, process non-zero entries above each pivot.

$$\underset{\mathbf{A}}{\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 7 & 8 \end{bmatrix}} \Rightarrow \underset{\mathbf{A_1}}{\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \\ 2 & 7 & 8 \end{bmatrix}} \Rightarrow \underset{\mathbf{A_2}}{\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 3 & 6 \end{bmatrix}} \Rightarrow \underset{\mathbf{A_{ref}}}{\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix}} \Rightarrow \underset{\mathbf{A_{rref}}}{\begin{bmatrix} 1 & 0 & -3 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix}}$$

Elementary operations are used to transform a Matrix to its echelon forms. The steps are as follows:

1. Pivot the matrix
   - Find the pivot, the first non-zero entry in the first column of the matrix.
   - Interchange rows, moving the pivot row to the first row.
   - Multiply each element in the pivot row by the inverse of the pivot, so the pivot equals 1.
   - Add multiples of the pivot row to each of the lower rows, so every element in the pivot column of the lower rows equals 0.
2. To get the Matrix in row echelon form, repeat the pivot
   - Repeat the procedure from Step 1 above, ignoring previous pivot rows.
   - Continue until there are no more pivots to be processed.
3. To get the Matrix in reduced row echelon form, process non-zero entries above each pivot.
   - Identify the last row having a pivot equal to 1, and let this be the pivot row.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- Add multiples of the pivot row to each of the upper rows, until every element above the pivot equals 0.
- Moving up the matrix, repeat this process for each row.

In the example given above, the following steps were followed to transform Matrix **A** into its echelon forms using a series of elementary row operations.

1. The first non-zero entry in the first column of the Matrix in row 2; so rows 1 and 2 were interchanged to get Matrix $\mathbf{A}_1$.
2. Each element of row 1 in Matrix $\mathbf{A}_1$ was multiplied by -2 and the result was added with row 3. This resulted in $\mathbf{A}_2$.
3. Each element of row 2 in Matrix $\mathbf{A}_2$ by -3 and the result was added to row 3 to produce $\mathbf{A}_{ref}$. Now, Aref is in row echelon form, based on the conditions specified in the previous slide.
4. The second row of $\mathbf{A}_{ref}$ was multiplied by -2 and the result was added to the first row to produce $\mathbf{A}_{rref}$.

9. Echelon forms and echelon transformations

9.1 Rank of a Matrix

- Imagine an r x c Matrix as a set of r row vectors, each having c elements; it can also be thought of a set of c column vectors, each having r elements.
- The rank of a Matrix is defined as:
  a) the maximum number of linearly independent column vectors in the matrix, or
  b) the maximum number of linearly independent row vectors in the matrix. Both definitions are equivalent.

The definition for rank of a Matrix is given above.

For an r x c matrix,

- If r is less than c, then the maximum rank of the Matrix is r.
- If r is greater than c, then the maximum rank of the Matrix is c.

The rank of a Matrix would be zero, only if the Matrix had no elements. If a Matrix had even one element, its minimum rank would be one.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

9.2 Finding the rank of a matrix

- The maximum number of linearly independent vectors in a Matrix is equal to the number of non-zero rows in its row echelon matrix.
- To find the rank of a matrix, we need to transform the Matrix to its row echelon form and count the number of non-zero rows.
- Consider Matrix **A** and its row echelon matrix, $\mathbf{A}_{ref}$.

$$\underset{\mathbf{A}}{\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 7 & 8 \end{bmatrix}} \Rightarrow \underset{\mathbf{A_{ref}}}{\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix}}$$

In the above example, the row echelon form $\mathbf{A}_{ref}$ has two non-zero rows, and Matrix **A** has two independent row vectors; and the rank of Matrix **A** is 2.

Row 1 and Row 2 of Matrix **A** are linearly independent. However, Row 3 is a linear combination of Rows 1 and 2. Specifically, Row 3 = 3*(Row 1) + 2*(Row 2). Therefore, Matrix *A* has only two independent row vectors.

9.3 Full Rank Matrices

- When all of the vectors in a Matrix are linearly independent, the Matrix is said to be full rank. Consider the matrices **A** and **B** below:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix}$$

- Row 2 of Matrix **A** is a scalar multiple of row 1; i.e., row 2 is equal to twice row 1. Therefore, rows 1 and 2 are linearly dependent. Matrix **A** has only one linearly independent row, so its rank is 1. Hence, Matrix **A** is not full rank.
- Now, if we take Matrix **B**, all of its rows are linearly independent, so the rank of Matrix **B** is 3. Matrix **B** is full rank.

The definition of a full rank Matrix and examples are given above.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

What did You Grasp?

1. What is the rank of the Matrix X?

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

A) 1
B) 2
C) 3
D) 4

10. Matrix Rank and Normal Form of a matrix

10.1 Normal Form of a Matrix

Normal form of a Matrix is a Matrix satisfying following conditions:

- consist of only ones and zeros.
- every row has a maximum of single one and rest are all zeros (there can be rows with all zeros).

We can produce the normal form of a Matrix by using elementary operations

Rank of a Matrix can be found from its normal form by counting the number of rows with non zero elements. But using echelon forms is an easy alternative.

We saw about normal form of a matrix. We'll learn how to reduce a Matrix to its normal form, using an example.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

10.2 Reducing a Matrix to its Normal Form

Reduce the matrix:

$$A = \begin{bmatrix} 5 & 3 & 8 \\ 0 & 1 & 1 \\ 1 & -1 & 0 \end{bmatrix}$$

Solution: We can reduce the Matrix to its normal form using elementary operations.

Step 1:

Applying $R_1 \leftrightarrow R_3$, we have

$$A \sim \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & 1 \\ 5 & 3 & 8 \end{bmatrix}$$

Step 2:

Applying $R_3 \rightarrow R_3 - 5R_1$, we have

$$A \sim \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 0 & 8 & 8 \end{bmatrix}$$

Steps involved in reducing a Matrix to its normal form is explained above.

Step 3: Applying elementary row operations $R_1 \rightarrow R_1 + R_2$ and $R_3 \rightarrow R_3 - 8R_2$, we have

$$A \sim \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Step 4: Now, we apply elementary column operation $C_3 \rightarrow C_3 - C_2$, to get

$$A \sim \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Step 5: Again, applying $C_3 \rightarrow C_3 - C_1$, we have

$$A \sim \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Steps involved in reducing a Matrix to its normal form is explained above.

We'll now move on to learn about vectors, vector spaces and associated techniques.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

11. Vector Spaces and the axioms

11.1 Vector Spaces

Vector Space:

- A space in mathematics is a set in which the list of elements is defined by a collection of guidelines or axioms for how each element relates to another within the set.
- A space in which the elements are sets of numbers themselves. Each element in a vector space is a list of objects that have a specific length, which we call vectors.
- Often we refer to the elements of a vector space as n-tuples, with n as the specific length of each of the elements in the set.

To define a Vector Space, first we need a few basic definitions.

- A set is a collection of distinct objects called elements. The elements are usually real or complex numbers when we use them in mathematics, but the elements of a set can also be a list of things.
- We notate a set by encasing the elements within curly braces. Note that to be distinct, an element cannot be repeated within the same set.
- {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} is the set of single digit numbers that we use in mathematics.
- {a, b, c, d, ..., y, z} is the set of letters in the alphabet.

The definition of a vector space is given above. Each element of a vector space of length n can be represented as a matrix, or a collection of numbers within parentheses.

Field:

- We refer to any vector space, as a vector space defined over a given field F.
- A field is a space of individual numbers, usually real or complex numbers.
- The specific axioms to define a field are similar to those of a vector space.
- We'll define a field as a vector space whose elements are single numbers that adhere to the same set of axioms as listed in the next section.

Understand fields from the context of vector spaces, as given in the slide above.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1. For any $\vec{v}, \vec{w} \in V : \vec{v} + \vec{w} \in V$.
2. For any $\vec{v}, \vec{w} \in V : \vec{v} + \vec{w} = \vec{w} + \vec{v}$.
3. For any $\vec{u}, \vec{v}, \vec{w} \in V : (\vec{v} + \vec{w}) + \vec{u} = \vec{v} + (\vec{w} + \vec{u})$.
4. There is a **zero vector** $\vec{0} \in V$ such that $\vec{v} + \vec{0} = \vec{v}$ for all $\vec{v} \in V$.
5. Each $\vec{v} \in V$ has an **additive inverse** $\vec{w} \in V$ such that $\vec{w} + \vec{v} = \vec{0}$.
6. If r is a **scalar**, that is, a member of $\mathbb{R}$ and $\vec{v} \in V$ then the **scalar multiple** $r \cdot \vec{v}$ is in V.
7. If $r, s \in \mathbb{R}$ and $\vec{v} \in V$ then $(r + s) \cdot \vec{v} = r \cdot \vec{v} + s \cdot \vec{v}$.
8. If $r \in \mathbb{R}$ and $\vec{v}, \vec{w} \in V$, then $r \cdot (\vec{v} + \vec{w}) = r \cdot \vec{v} + r \cdot \vec{w}$.
9. If $r, s \in \mathbb{R}$ and $\vec{v} \in V$, then $(rs) \cdot \vec{v} = r \cdot (s \cdot \vec{v})$
10. For any $\vec{v} \in V$, $1 \cdot \vec{v} = \vec{v}$.

We've seen about these conditions in the previous module. This is a recap of the vector spaces section in the previous module. We'll see the verification for these conditions in the upcoming sections.

Verification of Conditions 1 to 3

ere are five conditions in item 1. For 1, closure of addition, note that for any $v_1, v_2, w_1, w_2 \in \mathbb{R}$ the result of the sum

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} v_1 + w_1 \\ v_2 + w_2 \end{pmatrix}$$

a column array with two real entries, and so is in $\mathbb{R}^2$. For 2, that addition of vectors commutes, take all entries to be real numbers and compute

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} v_1 + w_1 \\ v_2 + w_2 \end{pmatrix} = \begin{pmatrix} w_1 + v_1 \\ w_2 + v_2 \end{pmatrix} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

e second equality follows from the fact that the components of the vectors are real numbers, and the addition of real numbers is commutative).
ndition 3, associativity of vector addition, is similar.

$$\begin{aligned} \left(\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \right) + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} &= \begin{pmatrix} (v_1 + w_1) + u_1 \\ (v_2 + w_2) + u_2 \end{pmatrix} \\ &= \begin{pmatrix} v_1 + (w_1 + u_1) \\ v_2 + (w_2 + u_2) \end{pmatrix} \\ &= \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \left(\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \right) \end{aligned}$$

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Verification of Conditions 4 to 7

For the fourth condition we must produce a zero element — the vector of zeroes is it.

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

For 5, to produce an additive inverse, note that for any $v_1, v_2 \in \mathbb{R}$ we have

$$\begin{pmatrix} -v_1 \\ -v_2 \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

so the first vector is the desired additive inverse of the second.

The checks for the five conditions having to do with scalar multiplication are just as routine. For 6, closure under scalar multiplication, where $r, v_1, v_2 \in \mathbb{R}$,

$$r \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} rv_1 \\ rv_2 \end{pmatrix}$$

is a column array with two real entries, and so is in $\mathbb{R}^2$. Next, this checks 7.

$$(r+s) \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} (r+s)v_1 \\ (r+s)v_2 \end{pmatrix} = \begin{pmatrix} rv_1 + sv_1 \\ rv_2 + sv_2 \end{pmatrix} = r \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + s \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

Verification of Conditions 8 to 10

For 8, that scalar multiplication distributes from the left over vector addition, we have this.

$$r \cdot \left(\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \right) = \begin{pmatrix} r(v_1 + w_1) \\ r(v_2 + w_2) \end{pmatrix} = \begin{pmatrix} rv_1 + rw_1 \\ rv_2 + rw_2 \end{pmatrix} = r \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + r \cdot \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

The ninth

$$(rs) \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} (rs)v_1 \\ (rs)v_2 \end{pmatrix} = \begin{pmatrix} r(sv_1) \\ r(sv_2) \end{pmatrix} = r \cdot \left(s \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \right)$$

and tenth conditions are also straightforward.

$$1 \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1v_1 \\ 1v_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

11.2 Axioms for Vector Spaces

There are ten axioms that define a vector space. Let x, y, and z be elements of the vector space V. Let a and b be elements of the field F:

1. Closed under addition: For each element x and y in V, x + y is also in V.
2. Closed under scalar multiplication: For each element x in V and scalar a in F, ax is in V.
3. Commutativity of addition: For each element x and y in V, x + y = y + x.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

4. Associativity of addition: For each element x, y, and z in V, (x + y) + z = x + (y + z).
5. Existence of the additive identity: There exists an element in V denoted as 0 such that x + 0 = x, for all x in V.
6. Existence of the additive inverse: For each element x in V, there exists another element in V that we will call -x such that x + (-x) = 0.
7. Existence of the multiplicative identity: There exists an element in F notated as 1 such that for all x in V, 1x = x.
8. Associativity of scalar multiplication: For each element x in V, and for each pair of elements a and b in F, (ab)x = a(bx).
9. Distribution of elements to scalars: For each element a in F and each pair of elements x and y in V, a(x + y) = ax + ay.
10. Distribution of scalars to elements: For each element x in V, and each pair of elements a and b in F, (a + b)x = ax + bx.

The ten axioms of vector spaces are listed above.

What did You Grasp?

1. For any vector in the vector space, addition follows associativity.

   A) True
   B) False

12. Linear Dependence and Independence of vectors

12.1.1 Linear Dependence and Independence of Vectors

Definition 1:

- Let $A = \{ \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r \}$ be a collection of vectors from R^n .
- If r > 2 and at least one of the vectors in A can be written as a linear combination of the others, then **A** is said to be **linearly dependent**.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

- If none of the vectors can be expressed as a linear combination of the others, then the vectors are independent.

The motivation for this description is simple. At least one of the vectors depends (linearly) on the others. On the other hand, if no vector in A can be expressed as a linear combination of the other two, then the vectors are independent. It is also quite common to say that "the vectors are linearly dependent (or independent)" rather than "the set containing these vectors is linearly dependent (or independent)."

12.1.1 (a) Example 1

1. Are the vectors $\mathbf{v}_1 = (2, 5, 3)$, $\mathbf{v}_2 = (1, 1, 1)$, and $\mathbf{v}_3 = (4, -2, 0)$ linearly independent?

   From $k_1\mathbf{v}_1 + k_2\mathbf{v}_2 = \mathbf{v}_3$,

   $k_1(2,5,3) + k_2(1,1,1) = (4,-2,0)$

   which is $2k_1 + k_2 = 4$

   $5k_1 + k_2 = -2$

   $3k_1 + k_2 = 0$

From this, the vectors can be said to be linearly independent.

If none of these vectors can be expressed as a linear combination of the other two, then the vectors are independent; otherwise, they are dependent. If, for example, $\mathbf{v}_3$ were a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$, then there would exist scalars $k_1$and k_2 such that $k_1\mathbf{v}_1 + k_2\mathbf{v}_2 = \mathbf{v}_3$.

The example given above is an inconsistent system. For instance, subtracting the first equation from the third yields k 1 = −4, and substituting this value into either the first or third equation gives $k_2 = 12$. However, $(k_1, k_2) = (-4, 12)$ does not satisfy the second equation.

The conclusion is that $\mathbf{v}_3$ is not a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$. A similar argument would show that $\mathbf{v}_1$ is not a linear combination of $\mathbf{v}_2$ and $\mathbf{v}_3$ and that $\mathbf{v}_2$ is not a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_3$. Thus, these three vectors are indeed linearly independent.

12.1.2 Linear Dependence and Independence of Vectors

Definition 2:

- An alternative, but entirely equivalent and often simpler definition of linear independence reads as follows. A collection of vectors $v_1, v_2, \ldots, v_r$ from R^n is linearly independent if the only scalars that satisfy are
- $k_1 = k_2 = \ldots\ldots = k_r = 0$. This is called the trivial linear combination.
- If, on the other hand, there exists a *nontrivial* linear combination that gives the zero vector, then the vectors are dependent.

Read out and understand the alternative definition for linear dependence and independence of vectors.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

12.1.2 (a) Example 2

Using the second definition, find out if the vectors $\mathbf{v}_1 = (2, 5, 3)$, $\mathbf{v}_2 = (1, 1, 1)$, and $\mathbf{v}_3 = (4, -2, 0)$ linearly independent?

These vectors are linearly independent if the only scalars that satisfy are $k_1 = k_2 = k_3 = 0$. But (*) is equivalent to the homogeneous system. $k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + k_3\mathbf{v}_3 = \mathbf{0} \quad (*)$

$$\begin{bmatrix} | & | & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \\ | & | & | \end{bmatrix}\begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix} = \mathbf{0} \quad (**)$$

Row reduction of the coefficient Matrix yields:

$$\begin{bmatrix} | & | & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \\ | & | & | \end{bmatrix} = \begin{bmatrix} 2 & 1 & 4 \\ 5 & 1 & -2 \\ 3 & 1 & 0 \end{bmatrix} \xrightarrow{-2r_1 \text{ added to } r_2} \begin{bmatrix} 2 & 1 & 4 \\ 1 & -1 & -10 \\ 3 & 1 & 0 \end{bmatrix}$$

$$\xrightarrow{r_1 \leftrightarrow r_2} \begin{bmatrix} 1 & -1 & -10 \\ 2 & 1 & 4 \\ 3 & 1 & 0 \end{bmatrix}$$

$$\xrightarrow[-3r_1 \text{ added to } r_3]{-2r_1 \text{ added to } r_2} \begin{bmatrix} 1 & -1 & -10 \\ 0 & 3 & 24 \\ 0 & 4 & 30 \end{bmatrix}$$

$$\xrightarrow{(-4/3)r_2 \text{ added to } r_3} \begin{bmatrix} 1 & -1 & -10 \\ 0 & 3 & 24 \\ 0 & 0 & -2 \end{bmatrix}$$

This echelon form of the Matrix makes it easy to see that $k_3 = 0$, from which follow $k_2 = 0$ and $k_1 = 0$. Thus, equation (**), and therefore (*), is satisfied only by $k_1 = k_2 = k_3 = 0$, which proves that the given vectors are linearly independent.

13. Consistency of linear system of equations

13.1 Consistency of Linear System of Equations

- Solving a System of Linear Equations in Two Variables:
  - A **consistent system** is a system that has at least one solution.
  - An **inconsistent system** is a system that has no solution.
- The equations of a system are **dependent** if ALL the solutions of one equation are also solutions of the other two equations. In other words, they end up being **the same line**.
- The equations of a system are **independent** if they **do not share ALL solutions**. They can have one point in common (point of intersection), just not all of them.

We saw the definition for consistent and inconsistent systems. There are three possible solutions for a system.

No Solution (No point of Intersection):

If the three planes are parallel to each other, they will never intersect. This means they do not have any points in common. In this situation, you would have no solution.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Unique (one) solution (Only one point of intersection):

If the system in three variables has one solution, it is an ordered triple (*x, y, z*) that is a solution to ALL THREE equations.

Infinite (many) Solutions (Lying on the same line):

If the three planes end up lying on top of each other, then there is an infinite number of solutions. In this situation, they would end up being on the same plane, so any solution that would work in one equation will work in the other.

14. Eigenvalues and eigenvectors

14.1 Eigenvalues

- When you multiply a Matrix (A) by a vector (v) you get a new vector (x).

$$\mathbf{Av} = \mathbf{x}$$

- There's also a special case where instead of getting a completely new vector you get a scaled version of the same vector you started with. In other words, a Matrix times a vector equals a scalar (lambda) times that same vector.

$$\mathbf{Av} = \lambda\mathbf{v}$$

- When this happens we call the scalar, lambda, an eigenvalue of Matrix A.
- The number eigenvalues a Matrix has will depend on the size of the matrix. An nxn Matrix will have n eigenvalues.

We saw about eigenvalues of a matrix. Let's now learn about the general solution for finding the eigenvalue of a matrix.

We know that $Av = \lambda v$, i.e. $Av - \lambda v = 0$.

The important point to note is that zero is not a scalar but the zero vector. Also, we are searching for a solution to the above equation under the condition that v is not equal to zero. When v equals zero lambda's value becomes trivial because any scalar or Matrix multiplied by a zero vector equals another zero vector.

We now have to multiply λv by an identity Matrix (I). Multiplying by an identity Matrix is like multiplying by one for scalar equations. $Av - \lambda Iv = 0$.

Since both A and lambda*I are multiplied by v we can factor it out. So, $(A - \lambda I)v = 0$

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

When v is not equal to zero this equation is true only if the Matrix we multiply v by is noninvertible. This means there must not exist a Matrix B such that C*B = B*C = I, where C = A - λI, here. A Matrix is noninvertible only when its determinant equals zero.

|A - λI| = 0

When we solve for the determinant we will get a polynomial with eigenvalues as its roots. We call this polynomial characteristic polynomial of the matrix. We can then figure out what the eigenvalues of the Matrix are by solving for the roots of the characteristic polynomial.

14.2 Finding Eigenvalue

Find the eigenvalue of the Matrix **A** = $\begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}$

Solution: By the equation, |A - λI| = 0.

$$\left| \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

Step 2:

$$\left| \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

$$\begin{vmatrix} 4 - \lambda & 2 \\ 1 & 3 - \lambda \end{vmatrix} = 0$$

Earlier we stated that an nxn Matrix has n eigenvalues. So a 2x2 Matrix should have 2 eigenvalues.

From the previous section, |A - λI| = 0, which is expanded as given above. First we insert our Matrix in for ***A***, and write out the identity matrix. In general an identity Matrix is written as an nxn Matrix with ones on the diagonal starting at the top left and zeroes everywhere else. We'll use a 2x2 identity Matrix here because we want it to be the same size as **A**.

The next step is to simplify everything inside the determinant to get a single matrix, which is given at step 2.

Now, we need to solve for the determinant and find the characteristic polynomial of the matrix.

(4 - λ)(3 - λ) - (2) (1) = 0

12 - 4λ - 3λ + λ2 = 0

λ2 - 7λ + 10 = 0

If we solve this polynomial, we'll get the solution as **λ = 5, λ = 2**

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

14.3 Properties of Eigenvalues

Properties f Eigenvalues are:

- **Property 1:**

  A square Matrix A and its transpose have the same eigenvalues.

- **Property 2:**

  The eigenvalues of a diagonal or triangular Matrix are its diagonal elements.

- **Property 3:**

  An n x n Matrix is invertible if and only if it doesn't have 0 as an eigenvalue.

The first three properties of eigenvalues are listed above. The proof can be given as follows:

Property 1:

We have that any solution of $\boldsymbol{det}(\mathbf{A} - \lambda\mathbf{I}) = 0$ is a solution of $\boldsymbol{det}(\mathbf{A} - \lambda\mathbf{I})^T = 0$ and vice versa. Thus **A** and $\mathbf{A}^T$ have the same eigenvalues.

$\boldsymbol{det}(\mathbf{A}^T - \lambda\mathbf{I})$

$= \boldsymbol{det}(\mathbf{A}^T - \lambda\mathbf{I}^T)$

$= \boldsymbol{det}(\mathbf{A} - \lambda\mathbf{I})^T$

$= \boldsymbol{det}(\mathbf{A} - \lambda\mathbf{I})$

The matrices **A** and $\mathbf{A}^T$ will usually have different eigen***vectors***.

Property 2:

Suppose the Matrix **A** is diagonal or triangular. If you subtract λ's from its diagonal elements, the result $\mathbf{A} - \lambda\mathbf{I}$ is still diagonal or triangular. Its determinant is the product of its diagonal elements, so it is just the product of factors of the form (diagonal element – λ). The roots of the characteristic equation must then be the diagonal elements.

Property 3:

An n x n Matrix **A** has an eigenvalue 0 if and only if $\boldsymbol{det}(\mathbf{A} - 0\mathbf{I}) = 0$, i.e. if and only if $\boldsymbol{det}(\mathbf{A}) = 0$. Since A is invertible if and only if $\boldsymbol{det}\mathbf{A} \neq 0$, **A** is invertible if and only if 0 is not an eigenvalue of **A**.

- **Property 4:**

  If a Matrix **A** has eigenvalue λ with corresponding eigenvector **x**, then for any k = 1, 2, ... , $\mathbf{A}^k$ has eigenvalue λ^k corresponding to the same eigenvector **x**.

- **Property 5:**

  If **A** is an invertible Matrix with eigenvalue λ corresponding to eigenvector **x**, then $\mathbf{A}^{-1}$ has eigenvalue λ^{-1} corresponding to the same eigenvector **x**.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

The fourth and fifth properties of eigenvalues are listed above. The proof can be given as follows:

Property 4:

Suppose the Matrix **A** has eigenvalue λ with eigenvector **x**, i.e. suppose that $\mathbf{Ax} = \lambda\mathbf{x}$. Then $\mathbf{A}^2\mathbf{x} = \mathbf{A}(\mathbf{Ax}) = \mathbf{A}(\lambda\mathbf{x}) = \lambda(\mathbf{Ax}) = \lambda(\lambda\mathbf{x}) = \lambda^2\mathbf{x}$. Multiply by more **A**'s to get $\mathbf{A}^3\mathbf{x} = \lambda^3\mathbf{x}$, $\mathbf{A}^4\mathbf{x} = \lambda^4\mathbf{x}$ and so on.

Property 5:

Multiply the equation $\mathbf{Ax} = \lambda\mathbf{x}$ by $\lambda^{-1}\mathbf{A}^{-1}$:

$$\lambda^{-1}\mathbf{A}^{-1}(\mathbf{Ax}) = \lambda^{-1}\mathbf{A}^{-1}\lambda\mathbf{x},$$

i.e.

$$\lambda^{-1}\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}.$$

Thus $\mathbf{A}^{-1}$ has eigenvalue λ^{-1} corresponding to the same eigenvector **x**.

What did You Grasp?

1. Find the eigenvalues of

$$\mathbf{A} = \begin{pmatrix} 2 & 7 \\ 7 & 2 \end{pmatrix}$$

A) -5, 9
B) 2, 7
C) 3, 8
D) -3, 9

Explanation for previous slide

Find all scalars, such that: λ has a nontrivial solution. That matrix equation has nontrivial solutions only if the matrix is not invertible or equivalently its determinant is zero. This is the characteristic equation.

$$(\mathbf{A} - \lambda\mathbf{I})\cdot\mathbf{x} = 0$$

$$|\mathbf{A} - \lambda\mathbf{I}| = \left|\begin{pmatrix} 2-\lambda & 7 \\ 7 & 2-\lambda \end{pmatrix}\right| = 0$$

$$(\mathbf{A} - \lambda\mathbf{I})\cdot\mathbf{x} = 0$$

$$\mathbf{P}(\lambda) = (2-\lambda)^2 - 49 = 0$$

$$\lambda_{1,2} = -5, 9$$

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

15. Cayley Hamilton Theorem

15.1 Cayley Hamilton Theorem

- Cayley-Hamilton theorem states that: "A square Matrix satisfies its own characteristic equation."

Explanation:

Let us consider that **A** be an nxn square Matrix and if its characteristic polynomial is defined as:

$P(\lambda) = |A - \lambda I_n|$

where, I_n is the identity Matrix of same order as **A**.

According to Cayley-Hamilton Theorem:P(A) = O, where, O represents the zero Matrix of same order as A.

We can say that if we replace λ by Matrix A, then the relation would be equal to zero. Hence Matrix A annihilates its own characteristic equation

15.2 Example for Cayley Hamilton Theorem

Prove Cayley Hamilton Theorem for the matrix $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

Solution:

Let A = $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

Characteristic polynomial of the Matrix is:

$$A - \lambda I = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$A - \lambda I = \begin{bmatrix} 1-\lambda & 2 \\ 3 & 4-\lambda \end{bmatrix}$$

$$P(\lambda) = |A - \lambda I| = (1-\lambda)(4-\lambda) - 6$$

$$= 4 - 5\lambda + \lambda^2 - 6$$

$$= \lambda^2 - 5\lambda - 2$$

For the given Matrix we have derived the characteristic equation. We'll now see how the Matrix satisfies Cayley Hamilton Theorem.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

15.3 Proof for Cayley Hamilton Theorem

In order to prove the statement of Cayley Hamilton theorem for A, we need to show that P(A) = O

$$A^2 - 5A - 2 = O$$

$$A^2 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$$A^2 = \begin{bmatrix} 7 & 10 \\ 15 & 22 \end{bmatrix}$$

$$A^2 - 5A - 2$$
$$= \begin{bmatrix} 7 & 10 \\ 15 & 22 \end{bmatrix} - 5\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} - 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 7 & 9 \\ 15 & 22 \end{bmatrix} - \begin{bmatrix} 5 & 10 \\ 15 & 20 \end{bmatrix} - \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

We have proved that P(A) = O, hence the Cayley Hamilton theorem for the given Matrix is proved.

15.4 Applications of Cayley Hamilton Theorem

- Cayley Hamilton theorem is widely applicable in many fields not only related to mathematics, but in other scientific fields as well.
- This theorem is used all over in linear algebra. One can easily find inverse of a Matrix using Cayley Hamilton theorem.
- It also plays an important role in solving ordinary differential equations. This theorem is quite useful in physics also.
- Cayley Hamilton theorem plays a vital role in computer programming and coding. In a newer subject - Rheology, where behavior of material is studied, this theorem is used to determine the equations that illustrate nature of materials.

Some of the applications of Cayley Hamilton theorem are listed above. In short, there is a vast use of Cayley Hamilton theorem in many areas where **linear equations** and **matrices** are needed to be used.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

16. Linear Transformation and Orthogonal transformation

16.1 Linear Transformation

- The main example of a linear transformation is given by Matrix multiplication. Given an nxm Matrix A, define T(v)=Av, where v is written as a column vector (with m coordinates).

For example, consider $A = \begin{bmatrix} 0 & 1 \\ -2 & 2 \\ 1 & 0 \end{bmatrix}$.

- Then, T is a linear transformation from $\mathbf{R}^2$ to $\mathbf{R}^3$ is defined by

$$T(x, y) = (y, -2x + 2y, x).$$

The definition for linear transformation is given above.

If T is a linear transformation, then T0 must be 0. (So if you find T0 ≠ 0, that means your T is not a linear transformation).

16.1 (a) Examples for Linear Transformations

1. $T : \mathbb{R}^5 \to \mathbb{R}^2$ defined by

$$T\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 2x_2 - 5x_3 + 7x_4 + 6x_5 \\ -3x_1 + 4x_2 + 8x_3 - x_4 + x_5 \end{bmatrix},$$

or equivalently,

$$T\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 0 & 2 & -5 & 7 & 6 \\ -3 & 4 & 8 & -1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}.$$

2. Scaling (expansion by factor 5): $T : \mathbb{R}^3 \to \mathbb{R}^3$ with matrix $\begin{bmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}$

A couple of examples for linear transformations are given above.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

16.2 Orthogonal Transformation

A linear transformation T from R^n to R^n is called orthogonal if it preserves the length of vectors:

$$\|T(\vec{x})\| = \|\vec{x}\|, \text{ for all } \vec{x} \text{ in } R^n$$

If $T(\vec{x}) = A\vec{x}$ is an orthogonal transformation, we say that A is an orthogonal matrix. For example, the rotation:

$$T(\vec{x}) = \begin{bmatrix} cos\phi & -sin\phi \\ sin\phi & cos\phi \end{bmatrix} \vec{x}$$ is an orthogonal transformation from R^2 to R^2, and

$$A = \begin{bmatrix} cos\phi & -sin\phi \\ sin\phi & cos\phi \end{bmatrix} \vec{x}$$ is an orthogonal Matrix for all angles ϕ.

The definition of orthogonal transformation along with an example is given above. We also need to know that:

- An orthogonal transformation is an isomorphism.
- The inverse of an orthogonal transformation is also orthogonal.
- The composition of orthogonal transformations is orthogonal.

An n x n Matrix is orthogonal if its columns are orthonormal.

17. Matrix Factorization and Types

17.1 Matrix Factorization

Matrix Factorization:

- Matrix decomposition methods, also called Matrix factorization methods.
- Are a foundation of linear algebra in computers, even for basic operations such as solving systems of linear equations, calculating the inverse, and calculating the determinant of a matrix.
- Just as its name suggests, Matrix factorization is to, obviously, factorize a matrix, i.e. to find out two (or more) matrices such that when you multiply them you will get back the original matrix.

Many complex Matrix operations cannot be solved efficiently or with stability using the limited precision of computers. Matrix decompositions are methods that reduce a Matrix into constituent parts that make it easier to calculate more complex Matrix operations.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

17.2 LU Decomposition

- A mxn Matrix is said to have a LU-decomposition if there exists matrices L and U with the following properties:
  - L is a mxn lower triangular Matrix with all diagonal entries being 1.
  - U is a mxn Matrix in some echelon form.
  - A = LU
- Let A be a mxn Matrix and E_1, E_2 .... E_k be elementary matrices such that

  U = E_k ..... E_1A is in non-echelon form. If none of the E_i's corresponds to the operation of row interchange, then

  C = E_k ..... E_1 is a lower triangular invertible matrix. Further L = C^{-1} is also a lower triangular Matrix with A = LU .

Suppose we want to solve a mxn system A**X** = b. If we can find a LU-decomposition for A , then to solve A**X** =b, it is enough to solve the systems, LY = b and UX = Y

Thus the system LY = b can be solved by the method of forward substitution and the system UX = Y can be solved by the method of backward substitution.

17.3 QR Decomposition

- Qr Decomposition is applicable to: mxn Matrix **A**.
- Decomposition: A=QR, where Q is a unitary Matrix of size mxm, and R is an upper triangular Matrix of size mxn
- Comment: The QR decomposition provides an alternative way of solving the system of equations Ax=b without inverting the Matrix A.
- The fact that Q is orthogonal means that Q^TQ =I, so that Ax=b is equivalent to $Rx=Q^Tb$, which is easier to solve since R is triangular.

The QR decomposition (also called the QR factorization) of a Matrix is a decomposition of the Matrix into an orthogonal Matrix and a triangular matrix. A QR decomposition of a real square Matrix A is a decomposition of A as A = QR, where Q is an orthogonal Matrix (i.e. Q^TQ = I) and R is an upper triangular matrix. If A is nonsingular, then this factorization is unique.

In general QR decomposition is not unique, but if A is of full rank, then there exists a single R that has all positive diagonal elements. If A is square, also Q is unique.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

17.4 Single Value Decomposition (SVD)

- The singular value decomposition of a Matrix A is the factorization of A into the product of three matrices $A = UDV^T$,
  - *A is an m × n matrix*
  - *U is an m × n orthogonal matrix*
  - *D is an n × n diagonal matrix*
  - *V is an n × n orthogonal matrix*
- For any m x n real Matrix A, the SVD consists of matrices U,D,V which are always real – this is unlike eigenvectors and eigenvalues of A which may be complex even if A is real.

SVD is applicable to: mxn Matrix A.

Decomposition: $A=UDV^T$, where D is a nonnegative diagonal matrix, and U and V are unitary matrices, and V^T is the conjugate transpose of V (or simply the transpose, if V contains real numbers only).

The singular values are always non-negative, even though the eigenvalues may be negative.

The diagonal elements of D are called the singular values of A. The singular value decomposition involves finding basis directions along which Matrix multiplication is equivalent to scalar multiplication, but it has greater generality since the Matrix under consideration need not be square.

Uniqueness: the singular values of A are always uniquely determined. U and V need not to be unique in general.

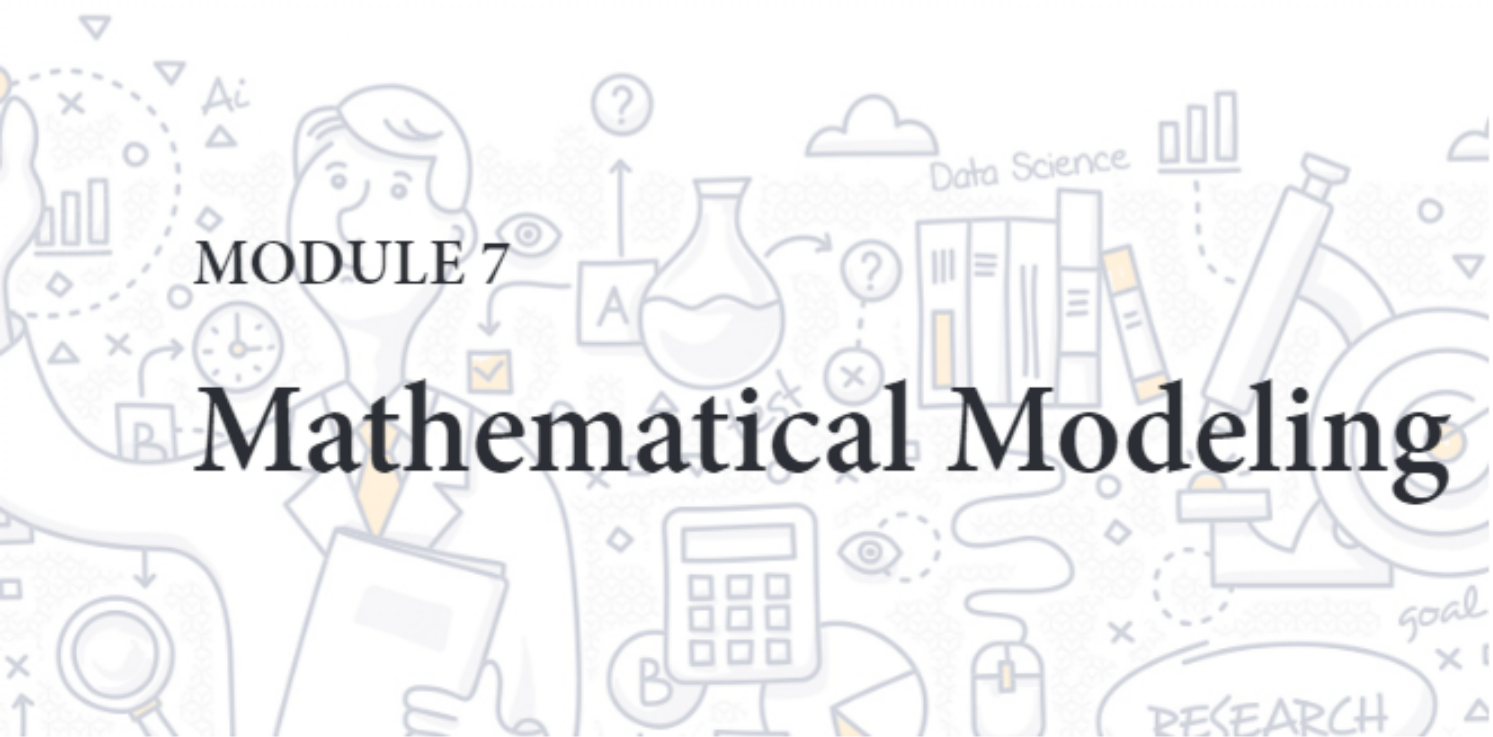
In a nutshell, we learnt:

1. Introduction to Matrices along with the basic terminologies
2. Matrix notations and types of matrices
3. Matrix equality
4. Operations on matrices like addition, subtraction and multiplication
5. Basics of Determinants
6. Singularity of a matrix
7. Orthogonal Matrix
8. Elementary transformations and elementary matrices and Matrix inverse computation using elementary transformations
9. Echelon forms and echelon transformations

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

10. How to find Matrix rank and reduce a Matrix to its normal form
11. Vector Spaces and enumerate the axioms
12. Linear dependence and independence of vectors
13. Consistency of linear system of equations
14. Eigenvalues and eigenvectors
15. Cayley Hamilton Theorem and its applications
16. Linear transformation and orthogonal transformation
17. Matrix factorization and enumerate a few types such as LU, QR and SVD decomposition

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna



MODULE 7

Mathematical Modeling

You will learn about the 'Mathematical Modeling' in this module.

Module Learning Objectives

At the end of this module, you will be able to:

- Understand mathematical modeling.
- Explain the reason for developing mathematical models along with applications.
- Describe the principles of mathematical modeling.
- Enumerate the different stages in developing a mathematical model.
- Enumerate the different ways to classify mathematical models.
- Describe the methodology for conceptualizing a mathematical model.
- Explain the concept of boundary conditions.

Module Topics

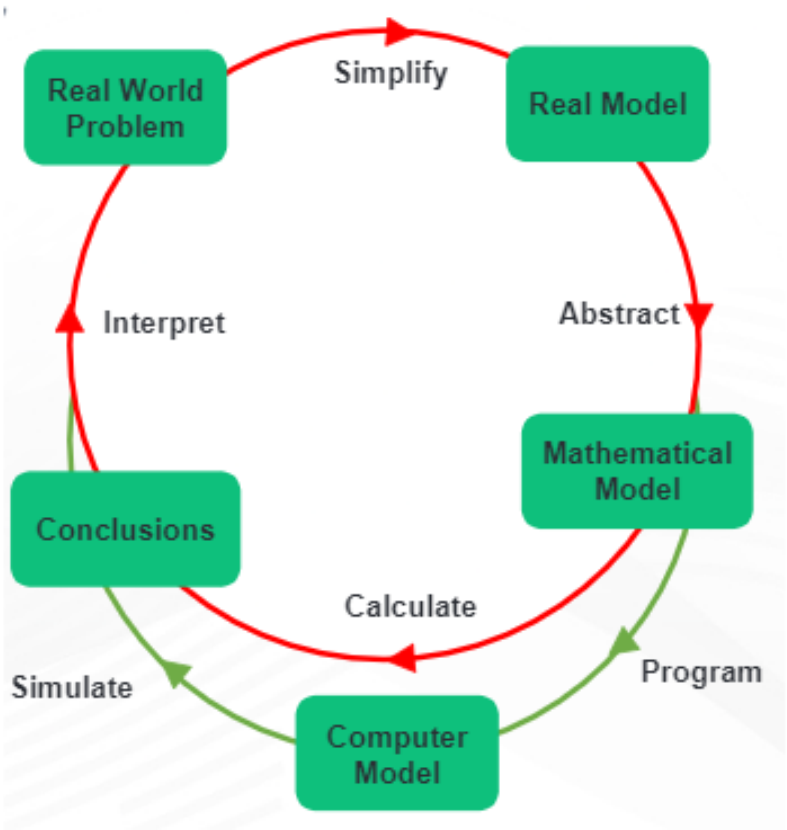
The following topics that will be covered in the module:

1. Introduction to mathematical modeling
2. Applications of mathematical modeling
3. Principles and stages involved in developing a mathematical model
4. Classification of mathematical modeling
5. Conceptualizing a mathematical model
6. Concept of boundary conditions

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

1. Introduction to Mathematical Modeling

- Mathematical model is a mathematical relation that describes a real-life situation.
- For a problem, a mathematical model is created and the model is used to analyse and derive the solution.
- The solution is then interpreted to the real-life problem being studied.
- Mathematical modeling is thus defined as the "conversion of physical situation into mathematics with some suitable conditions."



Machine Learning theory is a field that intersects statistical, probabilistic, computer science and algorithmic aspects arising from learning iteratively from data and finding hidden insights which can be used to build intelligent applications. Despite the immense possibilities of Machine and Deep Learning, a thorough mathematical understanding of many of these techniques is necessary for a good grasp of the inner workings of the algorithms and getting good results.

In mathematical modeling, we take a real-world problem and write it as an equivalent mathematical problem. We then solve the mathematical problem and interpret its solution in terms of the real-world problem. After this, we see to what extent the solution is valid in the context of the real-world problem.

Some of the common definitions of mathematical modeling are:

"A mathematical model is a representation, in mathematical terms, of certain aspects of a non-mathematical system." - Aris, 1999

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

"A mathematical model is a set of mathematical equations that are intended to capture the effect of certain system variables on certain other system variables." - G. C. Goodwin, et al., 2001

2. Why Mathematical Modeling

Mathematical Modeling:

- An essential tool for understanding the world and the processes around us.
- Provide a better representation of the real-life scenarios, which help in the development of better solutions.
- Can generate accurate mathematical models with the help of powerful computers and computing methods.
- Has countless applications in a number of industries.
- Some examples include:
  - Satellite launch
  - Monsoon prediction
  - Pollution control

Recollect that mathematical modeling is the process of describing a real-life situation mathematically and finding a solution for it so that it can be interpreted to the situation being studied.

Some of the examples where mathematical modeling is of great importance are:

- Finding the width and depth of a river at an unreachable place.
- Estimating the mass of the Earth and other planets.
- Estimating the distance between Earth and any other planet.
- Predicting the arrival of the monsoon in a country.
- Predicting the trend of the stock market.
- Estimating the volume of blood inside the body of a person.
- Predicting the population of a city after 10 years.
- Estimating the number of leaves in a tree.
- Estimating the ppm of different pollutants in the atmosphere of a city.
- Estimating the effect of pollutants on the environment.
- Estimating the temperature on the Sun's surface.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

3. Principles of Mathematical Modeling

Following are the principles of Mathematical Modeling:

1. Identify the need for the model. (for what we are looking for)
2. List the parameters/variables which are required for the model.
3. Identify the available relevant data. (what is given?)
4. Identify the circumstances that can be applied. (assumptions)
5. Identify the governing physical principles.
6. Identify:
   a. the equations that will be used.
   b. the calculations that will be made.
   c. the solution which will follow.
7. Identify tests that can check the
   a. consistency of the model.
   b. utility of the model.
8. Identify the parameter values that can improve the model.

Mathematical modeling has a few principles behind it, which are listed above. Based on these principles, there are five stages in mathematical modeling, which we will see in the next section.

What did You Grasp?

1. Fill in the blanks.

   While developing a mathematical model, tests are needed to check the ___ and ___ of the model.

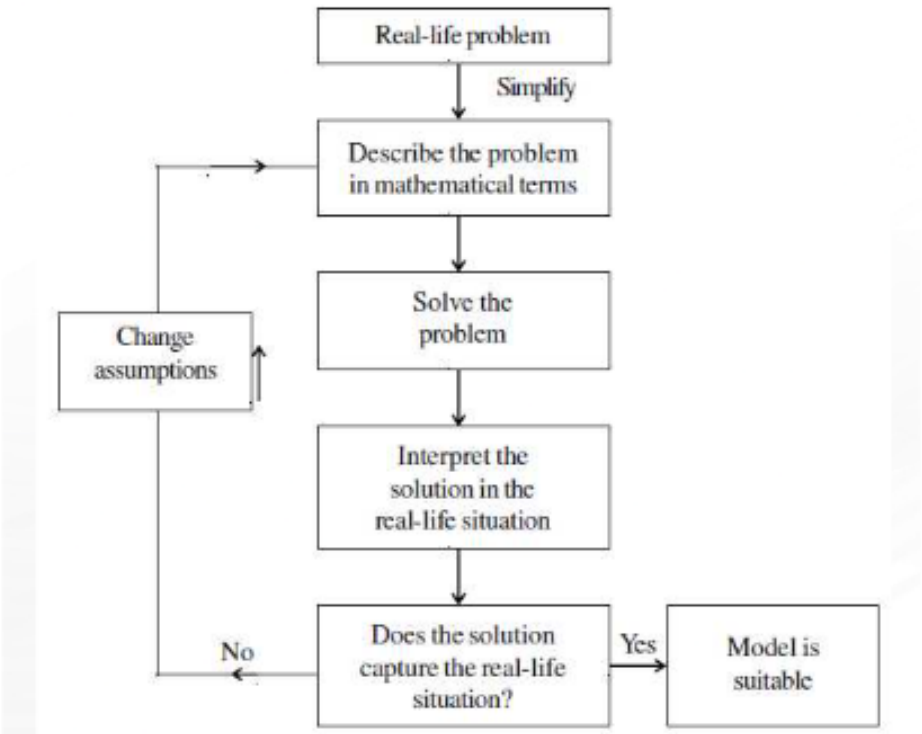
   A) Performance, Utility
   B) Consistency, Utility
   C) Efficiency, Robustness
   D) Parameters, Consistency

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

4. Stages in Mathematical Modeling

The stages in mathematical modeling are depicted in the illustration.



There are five important stages in developing a mathematical model as given in the above illustration.

1. **Understanding the problem:** The primary step is to understand and define the real problem. If you're working in a team, discuss the problem and the points that need to be understood clearly.

2. **Mathematical description and formulation:** To formulate, it is necessary to describe the different aspects of the problem mathematically. The following ways are used to describe the problem mathematically:

   - Define variables
   - Write equations or inequalities
   - Gather data and organise into tables
   - Make graphs
   - Calculate probabilities

3. **Solving the mathematical problem:** By means of mathematical description, the problem in hand is simplified. It is then solved using various mathematical techniques.

4. **Interpreting the solution:** The solution obtained in the previous stage is now interpreted to the real-life situation that is being studied.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

5. **Validating the model:** The final and the most important step is model validation. Here we go back to the problem being studied and check if the developed solution works and if it makes sense. If the model works, the model will be used until new information becomes available or the assumptions change.

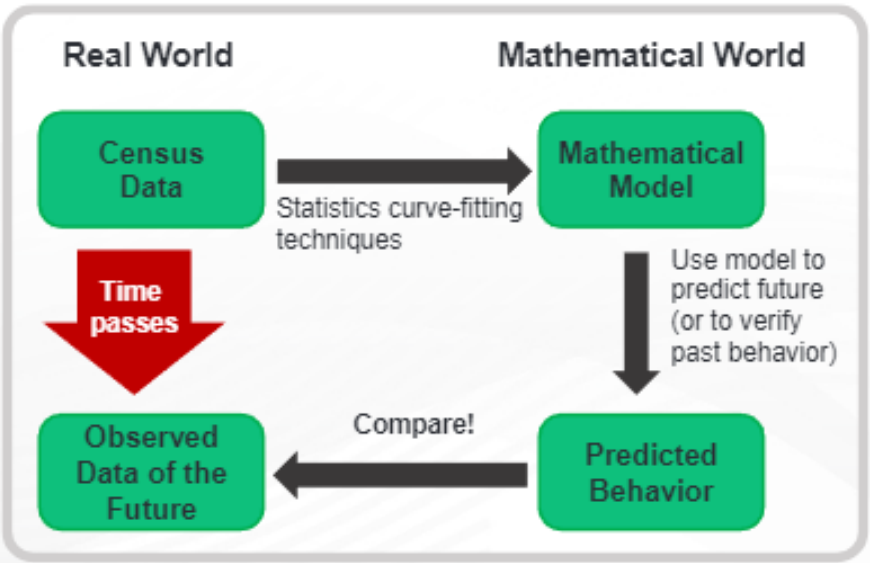
In some cases, because of the simplified assumptions, essential aspects of the real problem may be lost, while giving its mathematical description. In such cases, the solution could very often be off the mark, and may not make sense in the real situation. If this happens, assumptions made in stage 1 need to be reconsidered and revised to be more realistic, possibly by including some factors which were not considered earlier.

4.1 An Example for Mathematical Modeling

Construction of a mathematical model to predict the population of a country after 10 years.

Steps:

1. Formulation
2. Solution
3. Interpretation and Validation



Imagine a population control board of a certain country wants to predict the population of that country after 10 years. Let's look at the steps involved in constructing this model.

Step 1: Formulation

We know that population changes with time, and it increases with births and decreases with deaths. Suppose we want to find the population at a particular time 't', which is time in years. t takes the value of 1, 2, 3, … n. If t = 0 denotes the current year, t = 1, will be the next year. For any time t, let p(t) denote the population in that particular year.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

If we want to find the population of the country in a particular year, say, t_0 = 2018. Add the number of births and subtract the number of deaths in that year. If B(t) represents the number of births in one year between t and t + 1 and D(t) represents the number of deaths between t and t + 1. Then, we get the relation,

P(t + 1) = P(t) + B(t) - D(t) (Equation 1)

Let's make some assumptions and definitions now.

1. For the time interval t to t + 1, birth rate would be B(t)/P(t).
2. For the time interval t to t + 1, death rate would be D(t)/P(t).

Assumptions:

1. The birth rate and the death rates are the same for all intervals. So, there are two constants now, b and d, the birth and death rates, respectively, for all t >/= 0.
2. There is no migration into or out of the population, so the sources of change in population could be birth or death only. Based on assumptions 1 and 2, we deduce that, for t >/= 0,

P(t + 1) = P(t) + B(t) - D(t)

= P(t) + bP(t) - dP(t)

= (1 + b - d) P(t) (Equation 2)

Setting t = 0 in equation 2 gives,

P(1) = (1 + b - d) P(0) (Equation 3)

Setting t = 1 in equation 2 gives,

P(2) = (1 + b - d) P(1)

= (1 + b - d) (1 + b - d) P(0) (Using equation 3)

= $(1 + b - d)^2$ P(0)

Continuing this way, we get

P(t) = $(1 + b - d)^t$ P(0), for t = 0, 1, 2, ... (Equation 4)

The constant 1 + b - d is often abbreviated by r and called the growth rate, which is also known as Malthusian parameter. In terms of r, the equation 4 becomes

P(t) = P(0) r^t, for t = 0, 1, 2.... (Equation 5).

P(t) is an example of an exponential function. Any function of the form cr^t, where c and r are constants, is an exponential function. Equation 5 gives the mathematical formulation of the problem.

Step 2: Solution

Suppose the current population is 250,000,000 and the rates are b = 0.02 and d = 0.01. What will the population be in 10 years? Using the formula, we calculate P(10).

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

$$P(10) = (1.01)^{10}\,(250{,}000{,}000)$$
$$= (1.104622125)\,(250{,}000{,}000)$$
$$= 276{,}155{,}531.25$$

Step 3: Interpretation and Validation

The result we arrived in cannot be valid, as there cannot be 0.25 person. So we round it off and conclude that population would be 276,155,531 (approx). Here, we are not getting the exact answer, because of the assumptions that we have in our mathematical model. The above example shows how modeling is done to predict things that can happen a few years later, in a quantitative way.

What did You Grasp?

1. In which of the following stages of mathematical modeling, graphs are generated?

   A) Problem definition
   B) Mathematical description
   C) Solution interpretation
   D) Model validation

5. Classification of Mathematical Modeling

Different categories of classification:

1. Linear and Nonlinear
2. Static and Dynamic
3. Explicit and Implicit
4. Discrete and Continuous
5. Deterministic and Probabilistic (Stochastic)
6. Deductive, Inductive and Floating

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

There are different ways to classify mathematical models:

1. **Linear and Nonlinear:** The model is said to be linear, if all the operators in it exhibit linearity, otherwise the model is said to be nonlinear. The definition of linearity and nonlinearity is dependent on context, and there is a possibility of linear models containing nonlinear expressions in them. For example, in a statistical linear model, it is assumed that a relationship is linear in the parameters, but it may be nonlinear in the predictor variables. Similarly, a differential equation is said to be linear if it can be written with linear differential operators, but it can still have nonlinear expressions in it. In a mathematical programming model, if the objective functions and constraints are represented entirely by linear equations, then the model is regarded as a linear model. If one or more of the objective functions or constraints are represented with a nonlinear equation, then the model is known as a nonlinear model.

   Nonlinearity is often associated with phenomena such as chaos and irreversibility. Although there are exceptions, nonlinear systems and models tend to be more difficult to study than linear ones. A common approach to nonlinear problems is linearization, but this can be problematic if one is trying to study aspects such as irreversibility, which is strongly tied to nonlinearity.

2. **Static and Dynamic:** A dynamic model accounts for time-dependent changes in the state of the system, while a static (or steady-state) model calculates the system in equilibrium, and thus is time-invariant. Dynamic models typically are represented by differential equations or difference equations.

3. **Explicit and Implicit:** If all of the input parameters of the overall model are known, and the output parameters can be calculated by a finite series of computations, the model is said to be explicit. But sometimes it is the output parameters which are known, and the corresponding inputs must be solved for by an iterative procedure, such as Newton's method (if the model is linear) or Broyden's method (if non-linear). In such a case the model is said to be implicit. For example, a jet engine's physical properties such as turbine and nozzle throat areas can be explicitly calculated given a design thermodynamic cycle (air and fuel flow rates, pressures, and temperatures) at a specific flight condition and power setting, but the engine's operating cycles at other flight conditions and power settings cannot be explicitly calculated from the constant physical properties.

4. **Discrete and continuous:** A discrete model treats objects as discrete, such as the particles in a molecular model or the states in a statistical model; while a continuous model represents the objects in a continuous manner, such as the velocity field of fluid in pipe flows, temperatures and stresses in a solid, and electric field that applies continuously over the entire model due to a point charge.

5. **Deterministic and Probabilistic (stochastic):** A deterministic model is one in which every set of variable states is uniquely determined by parameters in the model and by sets of previous states of these variables. Therefore, a deterministic model always performs the same way for a given set of initial conditions. On the other hand, in a stochastic model randomness is present, and variable states are not described by unique values, but rather by probability distributions.

6. **Deductive, Inductive, and Floating:** A deductive model is a logical structure based on a theory. An inductive model arises from empirical findings and generalization from them. The floating model rests on neither theory nor observation but is merely the invocation of expected structure.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Application of mathematics in social sciences outside of economics has been criticized for unfounded models. Application of catastrophe theory in science has been characterized as a floating model.

What did You Grasp?

1. State True or False.
   A deterministic model varies with the state of the variables.

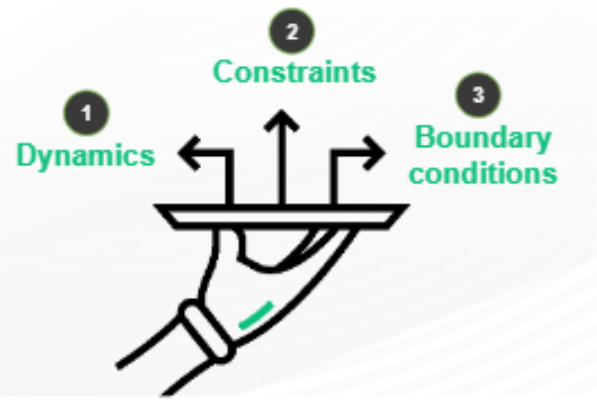
   A) True
   B) False

2. Fill in the blank.
   The model in which output parameters are known, but the corresponding input parameters need to be calculated, is called _____.

   A) Deterministic
   B) Explicit
   C) Implicit
   D) Deductive

6. Conceptualizing a Mathematical Model

- Building a mathematical model involves observing and understanding the problem, including the key unknowns.
- The phenomenology includes notes, quotes, questions, little pictures, mind maps, fragments of equations, fragments of pseudo-code, made-up graphs, etc.
- Three types of model building blocks, based on the phenomenology:



To conceptualize a mathematical model it is important to understand the problem deeply, including the unknowns. We can then sort out three types of model building blocks:

- **Dynamics:** Dynamics refers to the patterns in change and the laws that govern those changes. It forms the center of any mathematical thought. It is very easy to get insights with respect to the

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

dynamics and the sudden changes will become easily visible. Dynamics can be better explained using things like a swinging pendulum.

- **Constraints:** At the time of constraint creation, destruction, loosening or tightening, the changes are usually harder to notice, and the effects are often delayed or obscured. In case of a swinging pendulum, if the string of the pendulum is suddenly disturbed, the oscillation changes. But, if we notice only the dynamics, we might miss out the cause of it, that is, the introduction of a constraint. If not properly understood, valuable time will be lost in analysing the dynamics, instead of looking at the constraint.
- **Boundary conditions:** Most of the raw, primitive, numerical data in a mathematical modeling problem lives in the description of boundary conditions. The initial kick you might give a pendulum is an example. The fact that the rim of a vibrating drum skin cannot move is a boundary condition. When boundary conditions change, the effects can be extremely weird, and hard to sort out, if you aren't looking at the right boundaries.

7. Boundary Conditions

Boundary Conditions:

- Are a set of constraints. A boundary value problem comprises of a differential equation along with these additional constraints, termed as boundary conditions.
- If the solution to the differential equation is derived, which satisfies these boundary conditions, it serves as the solution of the boundary value problem as well.
- For every input to the problem, there is a unique solution, and it continuously depends on the input.

A boundary value problem has conditions specified at the extremes ("boundaries") of the independent variable in the equation whereas an initial value problem has all of the conditions specified at the same value of the independent variable (and that value is at the lower boundary of the domain, thus the term "initial" value).

For example, if the independent variable is time over the domain [0,1], a boundary value problem would specify values for y(t) at both t=0 and t=1, whereas an initial value problem would specify a value of y(t) and y'(t) at time t=0.

Finding the temperature at all points of an iron bar with one end kept at absolute zero and the other end at the freezing point of water would be a boundary value problem.

If the problem is dependent on both space and time, one could specify the value of the problem at a given point for all time or at a given time for all space.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.
Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna

Concretely, an example of a boundary value (in one spatial dimension) is the problem $y''(x)+y(x)=0$ to be solved for the unknown function $y(x)$ with the boundary conditions $y(0)=0$, $y(\Pi/2)=2$.

Without the boundary conditions, the general solution to this equation is $y(x)=A \sin(x)+B \cos(x)$

From the boundary condition $y(0) = 0$ we get $0 = A.0 +B.1$, which implies that $B = 0$. From the boundary condition $y(\Pi/2)=2$ we get $2 = A.1$ and so $A = 2$.

We see that imposing boundary conditions allows one to determine a unique solution, which in this case is $y(x) = 2 \sin(x)$.

What did You Grasp?

1. State True or False.

   Boundary conditions refer to the constraints specified at the upper extreme of the independent variable.

   A) True
   B) False

In a nutshell, we learnt:

1. Introduction to mathematical modeling.
2. The reason for developing mathematical models along with the applications.
3. The principles and the different stages in developing a mathematical model.
4. Different ways to classify mathematical models.
5. How to conceptualize a mathematical model.
6. The concept of boundary conditions.

This book is licensed to ABHAY GOYAL (XM2018073006) of Manan Rachna University, B.Tech CSE-DSML on 30 July 2018.
– Not to be copied, distributed or reproduced without prior written approval from MRU.

 Copyright © 2018, Xebia Group. All rights reserved. This course is licensed to Manav Rachna