Sanjivani Rural Education Society’s

Sanjivani College of Engineering, Kopargaon-423 603

Department of Information Technology

Foundation of Data Science (IT305B)

Prepared by

Mr. Umesh B. Sangule

Assistant Professor

Department of Information Technology

Unit-II

STATISTICS AND PROBABILITY FOR DATA SCIENCE

Course Objectives : *To introduce the statistics and probability for data science*

Course Outcome(CO2) : *Understand the statistics and probability for data science.*

Content

- Statistics for data science,
- Types of Analysis
- Dependence and Independence
- Conditional probability, Bayes Theorem
- Random Variables

Statistic for Data Science:

WHAT IS STATISTICS?

- Statistics exists because of the prevalence of variability in the real world.
- Statistics is the science of collecting,organizing and analyzing data,

1) Descriptive Statistics:

- In its simplest form, known as descriptive statistics, statistics provides us with tools tables, graphs, averages, ranges, correlations for organizing and summarizing the inevitable

- Variability in collections of actual observations or scores.
- Example : A tabular listing, ranked from most to least, A graph showing the annual change in global temperature during the last 30 years

Statistic for Data Science:

2) *Inferential Statistics:*

- Statistics also provides tools a variety of tests and estimates for generalizing beyond the collections of actual observations.

- This more advanced area is known as inferential statistics

- Example :

  " An assertion about the relationship between job satisfaction and overall happiness"

Terminologies in Statistics for Data Science:

Types of Data and Variables**:**:

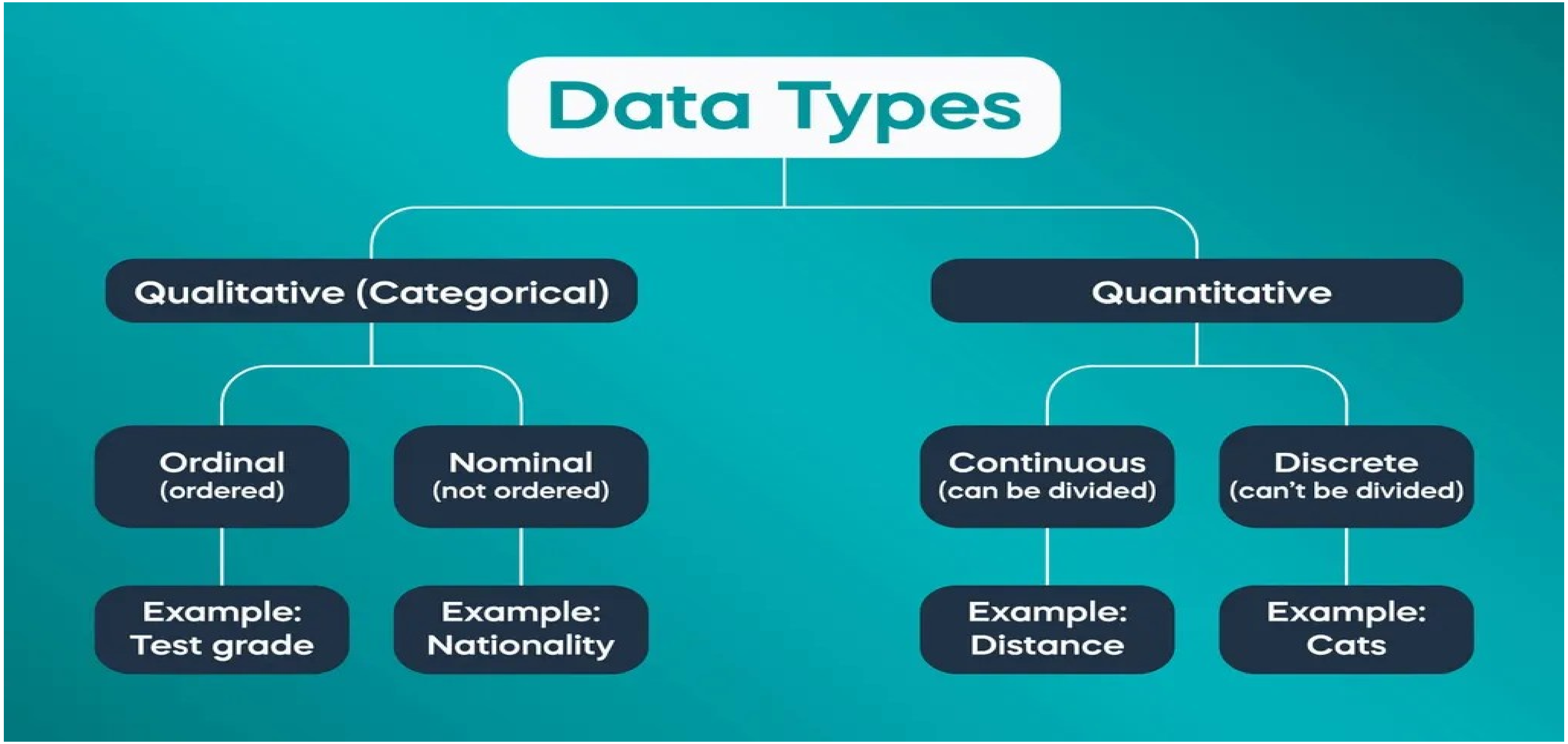
- Data can be descriptive (characteristics) or numerical (numbers).
- Let us take a look at some of the most prevalent types of data.
- There are two types of data:

1) *Qualitative Data*

2) Quantitative Data

Terminologies in Statistics for Data Science:

Types of Data:



Data Types

Qualitative (Categorical)

Ordinal (ordered)

Example: Test grade

Nominal (not ordered)

Example: Nationality

Quantitative

Continuous (can be divided)

Example: Distance

Discrete (can't be divided)

Example: Cats

Terminologies in Statistics for Data Science:

(A) Qualitative Data:

- There are **no numbers** in qualitative data, so it cannot be measured. It is also called "***Categorical Data***" because the data can be sorted by category rather than by number.

- Qualitative data is dealing with characteristics and descriptions that are difficult to measure but may be subjectively observed,

- Example:

  smells, tastes, textures, attractiveness, color, .

  They may include favourite foods, religions, pictures, symbols, colours, etc

<u>Terminologies in Statistics for Data Science</u>:

(A) <u>Qualitative Data</u>:

- These data are described by some characteristics, for example, gender, blood group. This data can provide answers to questions such as:

“How did it occur?” or **“Why did this occur?"**

- In general, qualitative data can be divided into two types:

1. Nominal data

2. Ordinal data

Terminologies in Statistics for Data Science:

1. Nominal data:

- This type of data is used for naming variables and has no numerical value

- Nominal data is a collection of values (non-numeric) that do not have a natural order.

- For example, it is not possible to state that 'Green' is greater than 'Blue', so we cannot compare one color to another, and so the color of a thing is a nominal data type

- Examples of Nominal Data:

  Colours: (Brown, Red, etc.)

  Taste: (Sour, Sweet, Salty, etc.)

  Languages: (Hindi, English, Marathi, Tamil, Telugu, etc.)

Terminologies in Statistics for Data Science:

2. Ordinal data:

- Ordinal data is defined as qualitative data whose values are ordered.

- In this type of data, a natural ordering occurs while maintaining class values. In other words, ordinal data is data that is sorted by its scale position.

- Ordinal numbers cannot be used for arithmetic because they only display sequence.

- Examples of Ordinal Data:

  Economic status: (low, medium, high)

  Letter grades: (A, B, C, D, E, etc.)

  Rank in a competition: (First, Second, Third)

Terminologies in Statistics for Data Science:

(B) Quantitative Data:

- Quantitative data are numbers.
- Numbers make up quantitative data. That is, the data represented in numbers, are quantitative data.
- Quantitative data is made up of numbers and things that can be measured objectively,
- Example:

  " e.g. area, volume, height, width, length, weight, speed, humidity, temperature, etc."

Terminologies in Statistics for Data Science:

(B) Quantitative Data:

- Quantitative data is always represented by numbers that indicate either

  “ **how much or how many.**”

- In general, quantitative data can be divided into two types:

  1. Discrete data

  2. Continuous data

Terminologies in Statistics for Data Science:

1. Discrete Data:

- Discrete data is counted, but it can only have certain values.

- Discrete data consists of finite, numeric, countable, and non-negative integers with discrete variables.

- Generally, it involves integers. The number of pupils, the number of children, the shoe size, and so on are all examples of discrete data.

- Examples of Discrete data:

  When we roll one die, we obtain 1, 2, 3, 4, 5, or 6 as discrete data.

  The total number of students enrolled in a class is discrete data

Terminologies in Statistics for Data Science:

2. Continuous Data:

- Continuous data is measured, and its value can be anything within a range.

- Continuous data is a set of numbers that can have any decimal or fractional value. Height, weight, length, time, temperature are all instances of continuous data.

- For example, The height of a person may be precisely 5.78 feet. We can measure someone's height in meters, centimetres, millimetres, and so on, so height is continuous data.

- Examples of continuous data:

  A freezer temperature

  Newborn babies' body weight

Terminologies in Statistics for Data Science:

2. Continuous Data:

- Continuous data can be further classified as measured on an interval scale or a ratio scale.

(i) Interval Scale:

Values that do not have a natural zero are referred to as the interval scale.

An interval scale has order and the difference between two values is significant.

Temperature, pH, and credit score are examples of interval variables.

Terminologies in Statistics for Data Science:

(ii) Ratio Scale:

- A ratio scale is a set of values that have a natural zero.
- A ratio variable contains all of the attributes of an interval variable, plus a distinct definition of 0.0.
- An example is a temperature measured in Kelvin. Below 0 degrees Kelvin, there is no value possible; it is absolute zero.
- Another example is weight; 0 kg indicates a notable absence of weight.

Types of Analysis:

Analysis:

> "Analysis is the process of inspecting, organizing, transforming and modelling collected data to identify patterns, connections and relationships."

From data science perspective Analysis can be grouped into four types:

1) Qualitative analysis
2) Quantitative analysis
3) Descriptive analysis
4) Predictive analysis

Types of Analysis:

1) Qualitative analysis:

- Qualitative data sets can be huge and diverse, analyzing them can provide valuable insights,
- Qualitative data analysis can help market researchers understand the mindset of their customers
- Analyzing purely numerical data (sales, revenue figures) without considering variable qualitative data (past experience, client feedback, customer reviews) may lead to incomplete findings or incorrect conclusions.

Types of Analysis:

2) *Quantitative analysis:*

- Quantitative data analysis simply means analysing data that is numbers based or data that can be easily “converted” into numbers without losing any meaning.

- For example, category-based variables like gender, ethnicity, or native language could all be “converted” into numbers without losing meaning

  e.g. English could equal 1, French 2, etc.

- This contrasts against qualitative data analysis, where the focus is on words, phrases and expressions that can’t be reduced to numbers.

Types of Analysis:

3) ***Descriptive analysis:***

- Descriptive analysis uses simple descriptive statistics, data visualization techniques, and business related queries to understand past data,

- Primary objective of descriptive analysis is innovative way of data summarization,

- Descriptive analysis is used for understanding the trends in past data,

- Most shoppers turn towards the right side when they enter a retail store Retailers keep products with higher profit on the right side of the store since most people turn right.

Types of Analysis:

3) ***Descriptive analysis:***

- ***Example:***

  “China Eastern Airline found that a man had booked a first class ticket more than 300 times in a year and cancelled it before its expiry for full refund so that he could eat free food at the airport's VIP lounge”

Types of Analysis:

*4) **Predictive analysis:***

- It aims to predict the probability of occurrence of a future event such as forecasting demand for products/services, customer churn, employee attrition, loan defaults, fraudulent transactions, insurance claim, and stock market fluctuations.

- Predictive analysis is used for predicting what is likely to happen in the future.

- How long a patient is likely to stay in the hospital, and so on will help organizations plan their future course of action.

Types of Analysis:

4) ***Predictive analysis:***

- Examples:

  "Amazon.com" uses predictive analytics to recommend products to their customers. It is reported that 35% of Amazon's sales is achieved through their recommender system (Siegel, 2013, MacKinzie 2013)

  "Hewlett Packard(HP)" Developed a flight risk score for its employees to predict who is likely to leave the company.

  "Netflix" Predicts which movie their customer is likely to watch next 75% of what customer watch at Netflix is from product recommendations.

Dependence and Independence:

- Dependent and Independent Events are the types of events that occur in probability.
- Suppose we have two events say Event A and Event B then if Event A and Event B are dependent events then the occurrence of one event is dependent on the occurrence of other events if they are independent events then the occurrence of one event does not affect the probability of other events.
- Examples such as the event of tossing two coins simultaneously the outcome of one coin does not affect the outcome of another coin then they are independent events.

Probability for Data Science:

Dependent Events:

- Dependent events are those events that are affected by the outcomes of events that had already occurred previously. i.e. Two or more events that depend on one another are known as dependent events.

- If one event is by chance changed, then another is likely to differ,

- For Example, let’s say three cards are to be drawn from a pack of cards. Then the probability of getting a king is highest when the first card is drawn, while the probability of getting a king would be less when the second card is drawn.

Probability for Data Science:

Independent Events:

- Independent events are those events whose occurrence is not dependent on any other event.
- If the probability of occurrence of an event A is not affected by the occurrence of another event B, then A and B are said to be independent events.
- Examples of Independent Events: “Tossing a Coin”, “Rolling a Die” etc.

Probability for Data Science:

Mutually exclusive events:

- Two or more events are said to be mutually exclusive, when the occurrence of any one event excludes the occurrence of the other event.

- Mutually exclusive events cannot occur simultaneously.

- For example when a coin is tossed, either the head or the tail will come up. Therefore the occurrence of the head completely excludes the occurrence of the tail. Thus getting head or tail in tossing of a coin is a mutually exclusive event.

Probability for Data Science:

Definitions of Probability:

There are two types of probability. They are Mathematical probability and Statistical probability.

- ***Mathematical Probability* (or a priori probability)**:

  “If the probability of an event can be calculated even before the actual happening of the event, that is, even before conducting the experiment, it is called ***Mathematical probability (or a priori probability)***.”

- Example:

  “If we keep using the examples of tossing of fair coin, dice etc., we can state the answer in advance (*prior*), without tossing of coins or without rolling the dice etc.,”

Probability for Data Science:

- ***Statistical Probability*** ***(or a posteriori probability)***:

  “If the probability of an event can be determined only after the actual happening of the event, it is called ***Statistical probability(or a posteriori probability***).”

- If a coin is tossed 10 times we may get 6 heads and 4 tails or 4 heads and 6 tails or any other result. In these cases the probability of getting a head is **not 0.5** as we consider in Mathematical probability.

- However, if the experiment is carried out a large number of times we should expect approximately equal number of heads and tails and we can see that the probability of getting head approaches 0.5.

Probability for Data Science:

- ***Conditional probability***:

- Let A be any event with p(A) >0.

- The probability that an event B occurs subject to the condition that A has already occurred is known as the conditional probability of occurrence of the event B on the assumption that the event A has already occurred.

- It is denoted by the symbol P(B/A) or P(B|A) and is read as the probability of B given A.

Probability for Data Science:

- ***Conditional probability***:

  The same definition can be given as follows also:

- Two events A and B are said to be dependent when A can occur only when B is known to have occurred (or vice versa).The probability attached to such an event is called the **conditional probability**

- It is denoted by P(B/A) or, in other words, probability of B given that A has occurred.

Probability for Data Science:

- ***Bayes Theorem:***

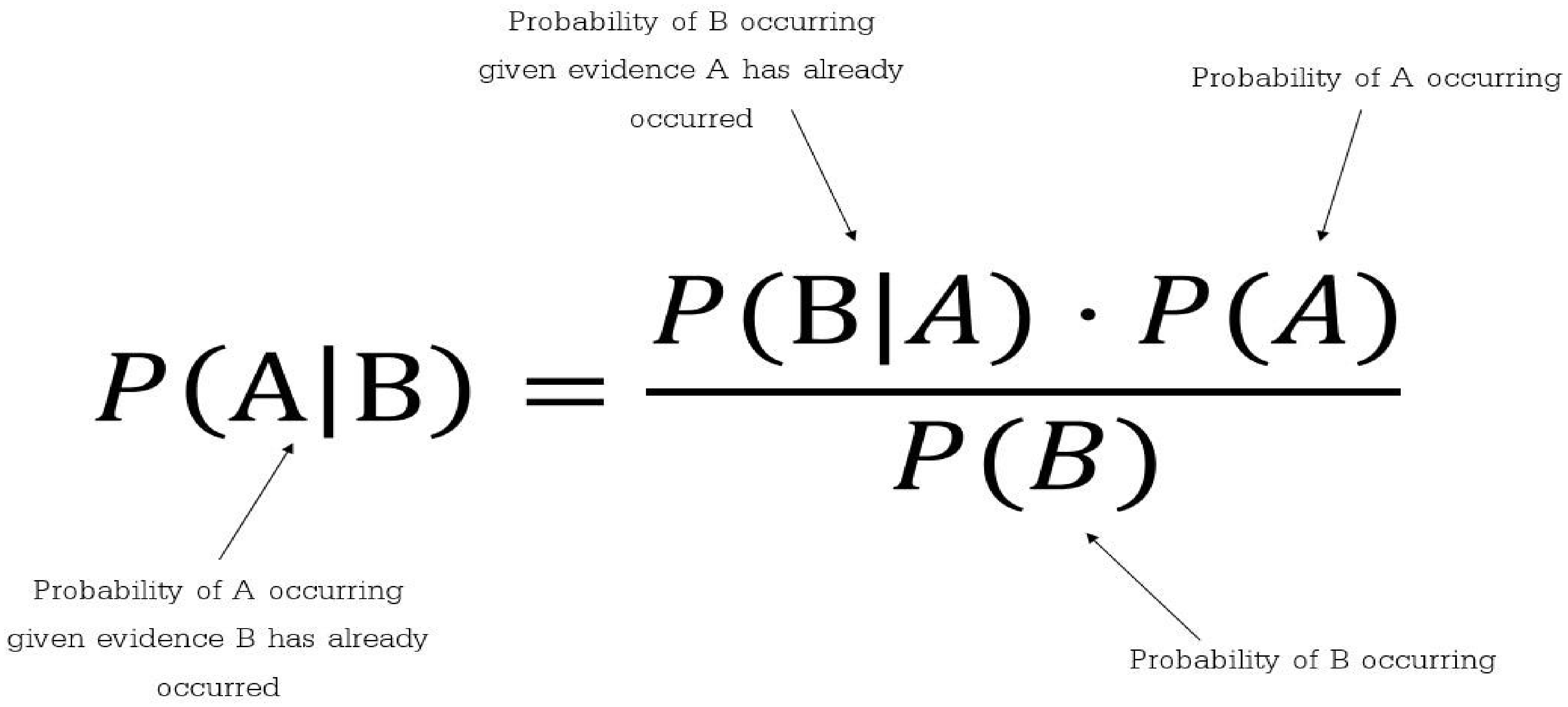
- The concept of conditional probability discussed earlier takes into account information about the occurrence of one event to predict the probability of another event.

- This concept can be extended to revise probabilities based on new information and to determine the probability that a particular effect was due to specific cause.

- The procedure for revising these probabilities is known as Bayes theorem.

Probability for Data Science:

➢ *Bayes Theorem:*



Probability of B occurring given evidence A has already occurred

Probability of A occurring

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

Probability of A occurring given evidence B has already occurred

Probability of B occurring

Random Variable:

- ***Random Variable:***

- Random variable is a function that maps every outcome in the sample space to real number,

- " A function that assigns a real number to each sample point in the sample space"

- Example: bank transactions

  {GGFF, GGGF,FFFFF}

  {1,2,3}

Random Variable:

- ***Random Variable:***

- Use of Random variable provides us flexibility required for modelling,

- Random Variable can be classified as-

  1. Discrete Random variables
  2. Continuous Random Variables

1. Discrete Random Variables:

- If the random variable X can assume only a finite or countably infinite set of values, then it is called a discrete random variable.

Random Variable:

1. Discrete Random Variables:

➢ There are very many situations where the random variable X can assume only finite or countably infinite set of values.

➢ Examples of discrete random variables are:

1. Credit rating (usually classified into different categories such as low, medium and high or using labels such as AAA, AA, A, BBB, etc.).
2. Number of orders received at an e-commerce retailer which can be countably infinite.
3. Customer churn [the random variables take binary values:

   (a) Churn and (b) Do not churn].

➢ In analytics, classification problems, an important class of problems, is an example of discrete random variable.

Random Variable:

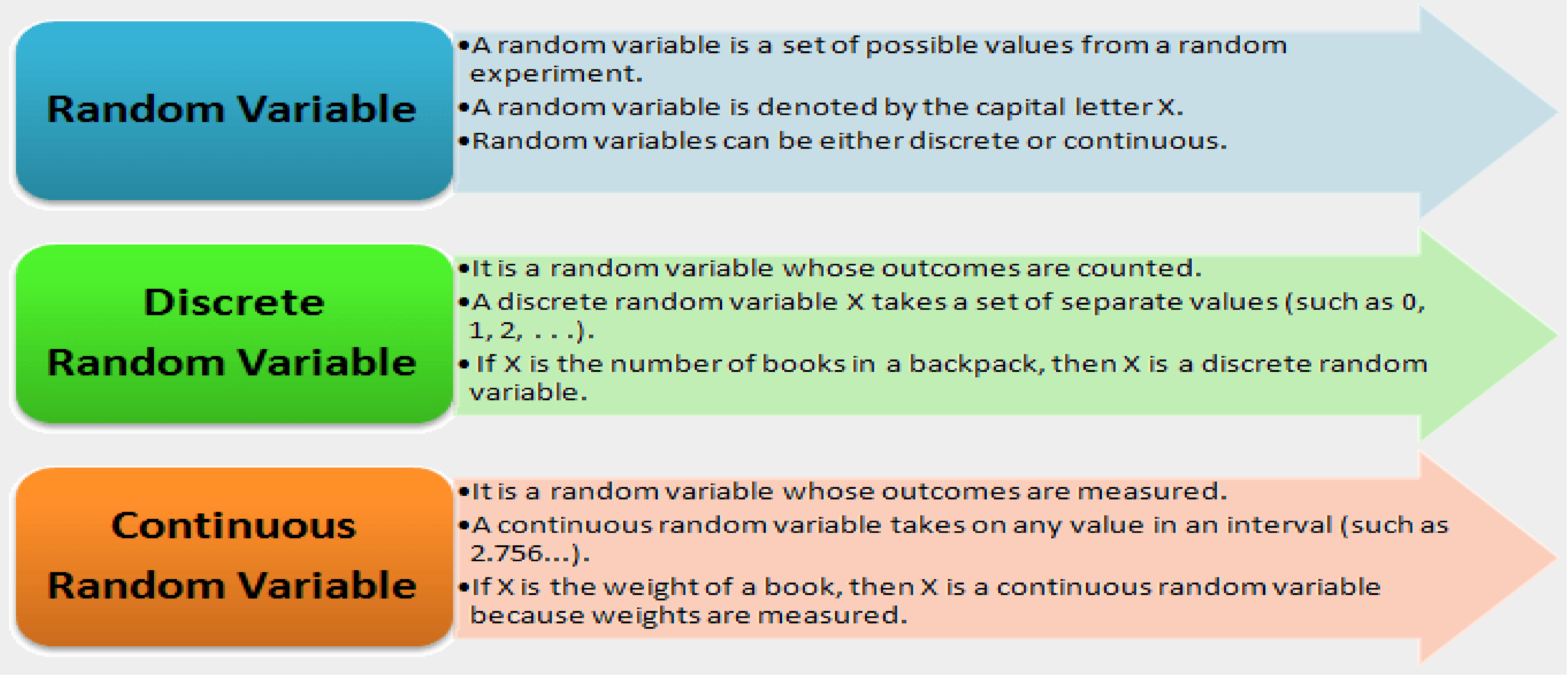
2. Continuous Random Variables:

- A random variable X which can take a value from an infinite set of values is called a continuous random variable.
- Examples of continuous random variables are listed below:

1. Market share of a company (which take any value from an infinite set of values between 0 and 100%).
2. Percentage of attrition among employees of an organization.
3. Time to failure of engineering systems.
4. Time taken to complete an order placed at an e-commerce portal.
5. Time taken to resolve a customer complaint at call and service centers.

- In many situations, a continuous variable may be converted to a discrete random variable for modelling purpose

Random Variable:



Random Variable

- A random variable is a set of possible values from a random experiment.
- A random variable is denoted by the capital letter X.
- Random variables can be either discrete or continuous.

Discrete Random Variable

- It is a random variable whose outcomes are counted.
- A discrete random variable X takes a set of separate values (such as 0, 1, 2, . . .).
- If X is the number of books in a backpack, then X is a discrete random variable.

Continuous Random Variable

- It is a random variable whose outcomes are measured.
- A continuous random variable takes on any value in an interval (such as 2.756...).
- If X is the weight of a book, then X is a continuous random variable because weights are measured.

THANK YOU