Advanced Statistics

Erbil Technical Engineering College

PhD. Course of Technical Mechanical and Information Systems Engineering

2023-2024

Assist. prof. Dr. Paree khan A. Omer

Administration & Economics College / Statistics and Informatics Department

Lect. 1

Ch.1 Overview and Descriptive statistics

1.1 Concept of Engineering Statistics

- **Statistics** is the science of collecting data, presentation, analyzing, interpreting and drawing conclusions from information. Statistics is a science that helps us make decisions and draw conclusions in the presence of variability. Specifically, Statistical techniques can be powerful aids in designing new products and systems, improving existing designs, developing, and improving production processes. In generally, Statistics is a way to get information from data. Scientists and engineers approach the study of statistics by considering a four-step process called the **Scientific method** for interpreting and drawing conclusions from collected data:
- Defining the problem.
- Collecting the data.
- Summarizing the data.
- Analyzing the data, interpreting the analyses and utilizing the results.

Statistical methods are used to help us describe and understand **Variability.** variability, we mean that successive observations of a system or phenomenon do not produce exactly the same result. We all encounter variability in our everyday lives, and **Statistical Thinking** can give us a useful way to incorporate this variability into our decision-making processes.

1.2 Three Fundamental Components of Statistics

- In general, statistics provides methods for:

1. Planning and carrying out research studies (Design).

2. Summarizing and exploring data (Description).

3. Predicting and generalizing about phenomena represented by data (Inference).

There are two major types of statistics:

1. Descriptive Statistics

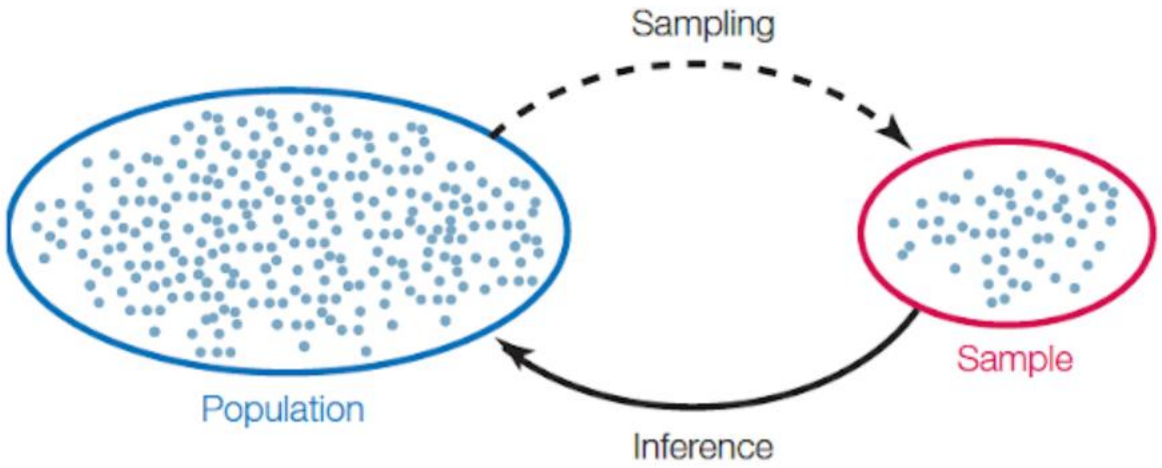
- Descriptive Statistics is the branch of statistics consist of methods for organizing, summarizing and description of data. This branch includes both numerical measures (e.g. mean or the median) and graphical displays of data (e.g. bar or pie charts).

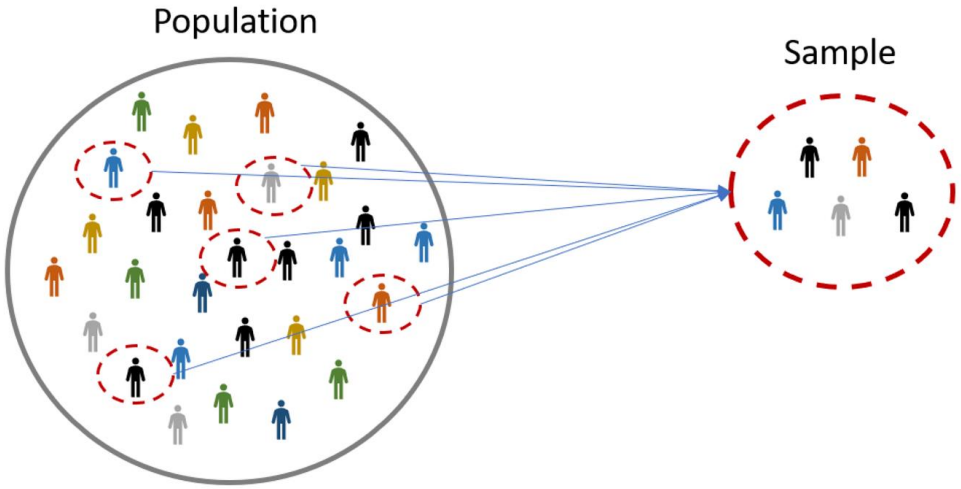
2. Inference (inferential) Statistics

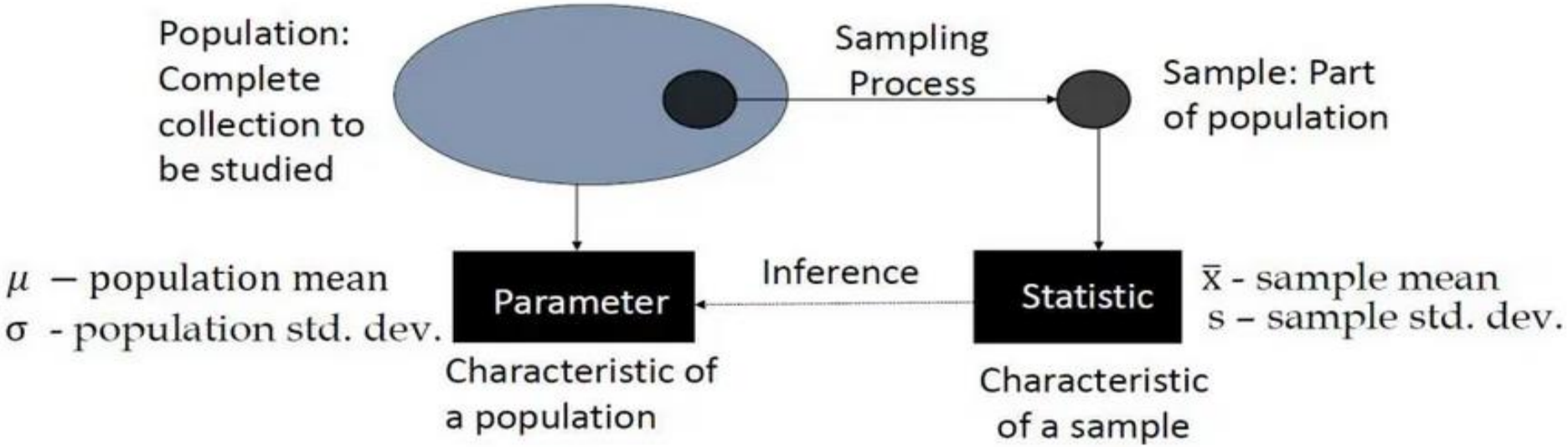
- Inferential Statistics consists of methods for drawing conclusions and measuring the reliability of conclusions about population on the basis of measurements or observations made on a sample of units from the population. it includes methods like:

1. Estimation (point estimation and interval estimation)
2. Hypothesis testing which are all based on probability theory.

- A **Population** is a data set (usually large, sometimes conceptual) that is our target of interest.
- A **Sample** is a subset of data selected from the target population. In some cases, the sample consists of the whole population, in which case it is termed a **census**.
- A **Parameter** is a characteristics or number represents some aspect of the population as a whole. It is an unknown numerical summary of the population or a descriptive measure of a population.
- **A Statistic** is a known numerical summary of the sample which can be used to make inference about parameters or a descriptive measure of a sample. It is a characteristics or number computed from the sample data.


Sampling
Sample
Population
Inference


Population
Sample

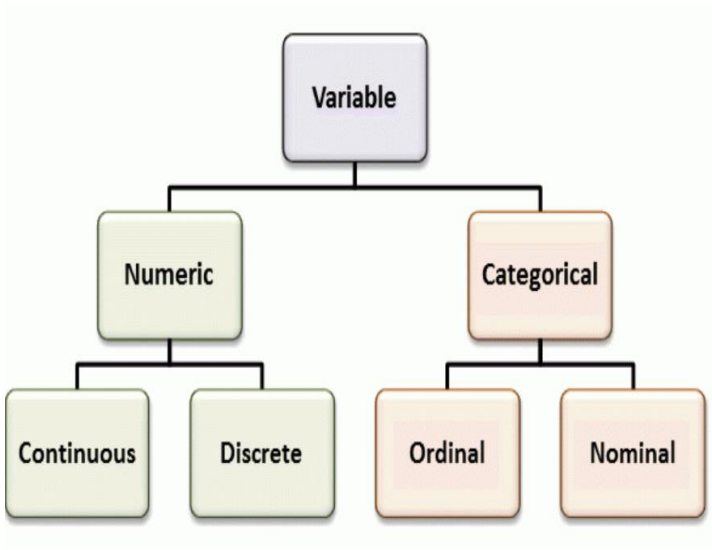

Population: Complete collection to be studied
Sampling Process
Sample: Part of population
μ – population mean
σ - population std. dev.
Parameter
Inference
Statistic
x̄ - sample mean
s – sample std. dev.
Characteristic of a population
Characteristic of a sample

1.3 Variable

- A variable is any characteristic, number, or quantity that can be measured or counted. A variable may also be called a data item. Like (Age, sex, business income and expenses, capital expenditure, class grades, eye colour and vehicle type) are examples of variables. It is called a variable because the value may vary between data units in a population, and may change in value over time. There are 2 basic types of variables: quantitative and qualitative.

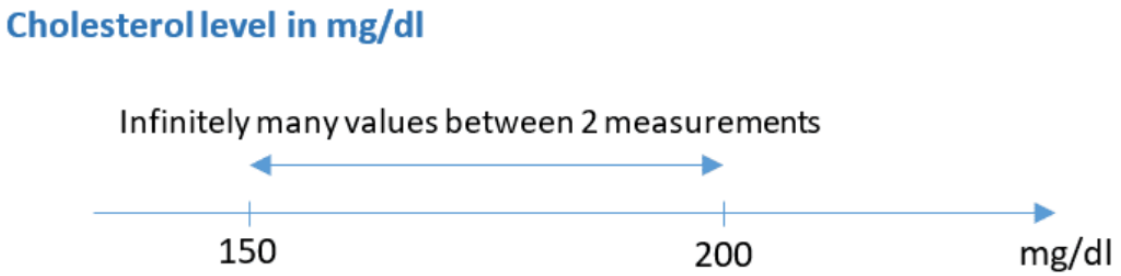
1. Quantitative or Numerical variable:

- A quantitative is a type of variable consisting of values that represent counts or measurements of a certain quantity. For instance, age, height, number of cigarettes smoked, etc.
- A quantitative variable can be either continuous or discrete.



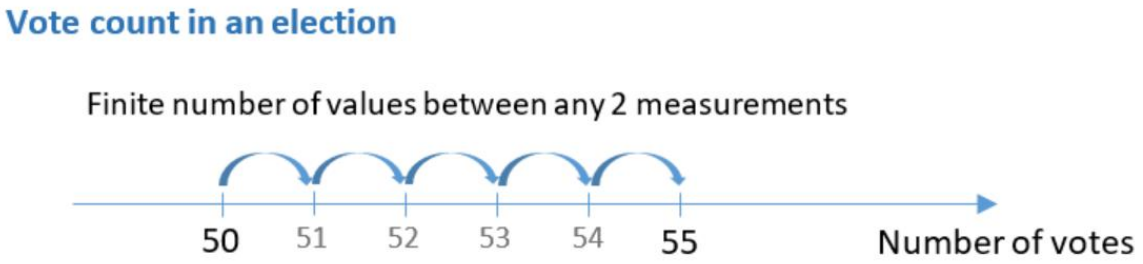
1.1. Continuous variable:

- A continuous variable is a type of quantitative variable consisting of numerical values that can be measured but not counted, because there are infinitely many values between 1 measurement and another. Example: Cholesterol level measured in mg/dl.



1.2. Discrete variable:

- A discrete variable is a type of quantitative variable consisting of numerical values that can be measured and counted, because these values are separate or distinct. Example: Vote count in an election.



2. Qualitative or Categorical variable:

- A qualitative or categorical variable is a type of variable consisting of text characters or labels that describe groups of observations. For instance, gender, marital status, stages of a disease, etc. A qualitative variable can be either ordinal or nominal.

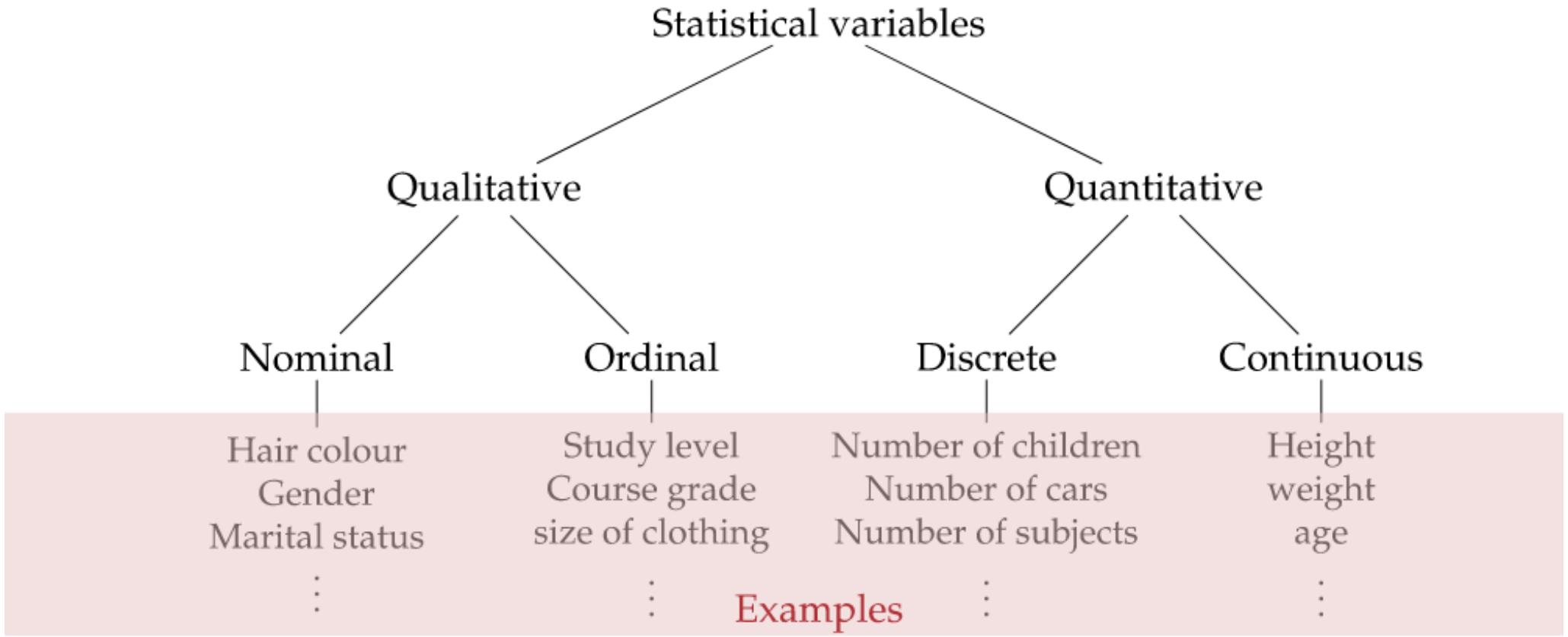
2.1. Ordinal variable:

- An ordinal variable is a type of qualitative variable consisting of text or labels that have a logical order, i.e. one category represents more or less of the other, but taking the difference between categories or their average is meaningless.

2.2. Nominal variable:

- A nominal variable is a type of qualitative variable consisting of text or labels that have no logical order.

- **A Decision Tree for Identifying Variable Types**



1.4 Probability

Probability as a specific term is a measure of the likelihood that a particular event will occur. Just how likely is it that the outcome of a trial will meet a particular requirement? If we are certain that an event will occur, its probability is 1 or 100%. If it certainly will not occur, its probability is zero.

Probability formula is: $$P(A) = \frac{\text{Number of favorable outcomes to A}}{\text{Total number of possible outcomes}}$$

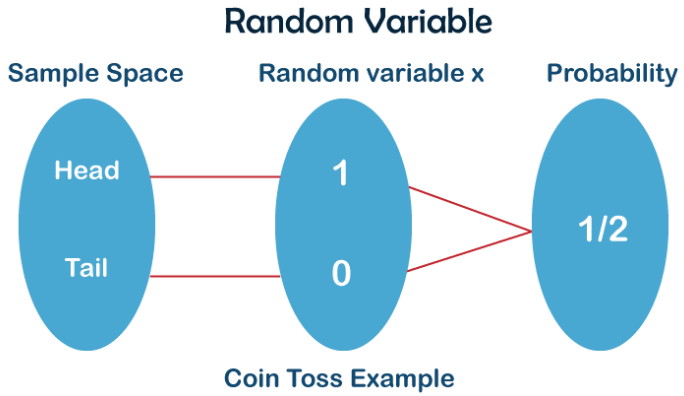
Ex. // Two fair coins are tossed. What is the probability of getting one heads and one tails?

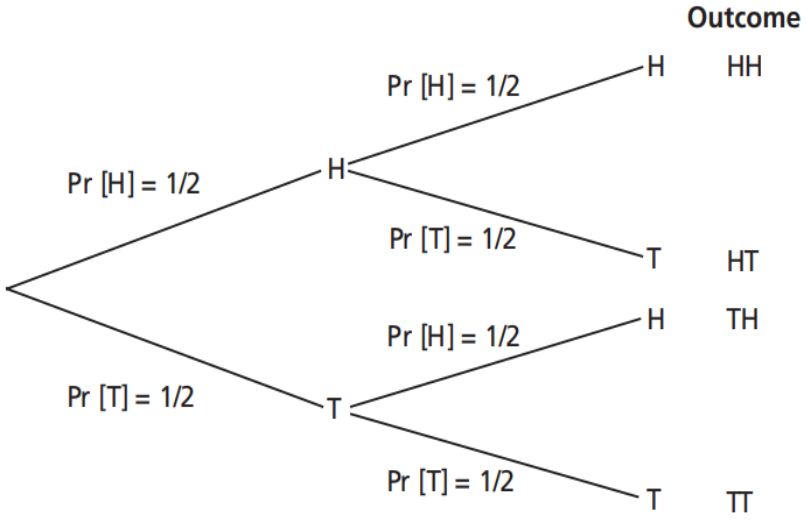
Answer: For a fair or unbiased coin, for each toss of each coin

Pr [head] = Pr [tail] =1/2

The possible outcomes of tossing the two coins are: HH, HT, TH, and TT

S=(HH.HT,TH,TT)





1.5 Probability theory

- Probability theory is a branch of mathematics that investigates the probabilities associated with a random phenomenon. A random phenomenon can have several outcomes. Probability theory describes the chance of occurrence of a particular outcome by using certain formal concepts.

- Example // Suppose the probability of obtaining a number 4 on rolling a fair dice needs to be established. The number of favorable outcomes is 1. The possible outcomes of the dice are {1, 2, 3, 4, 5, 6}. This implies that there are a total of 6 outcomes. Thus, the probability of obtaining 4 on a dice roll, using probability theory, can be computed as 1 / 6 = 0.167.

1.6 Random Variable

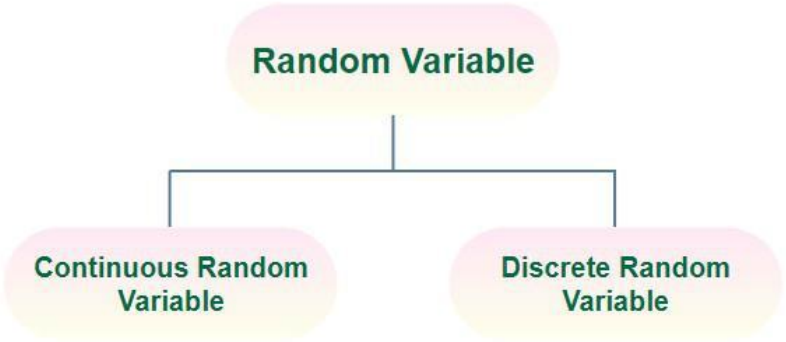
- A random variable is a mathematical function that assigns a numerical value to every outcome which is possible in a random experiment. It is called a random variable because its value depends on the output of a random experiment. Random variables are usually symbolized by capital letters, such as X, Y, or Z.

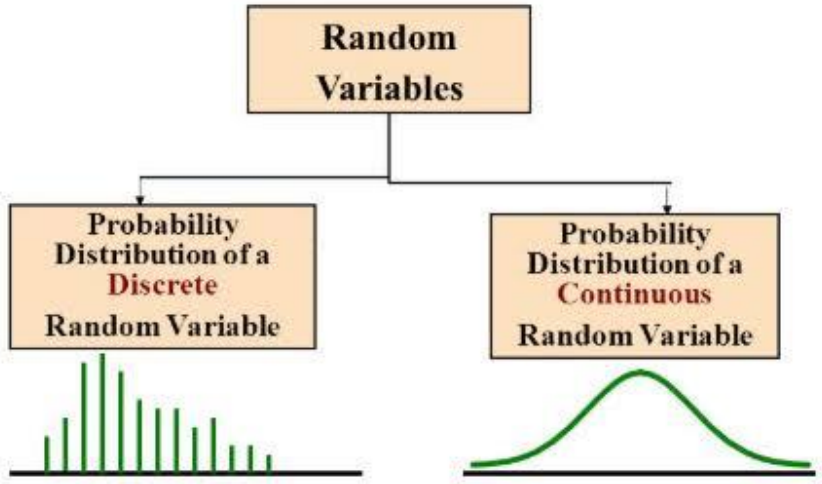
Types of variables

- There are two types of random variables as given below.

1. Discrete Random Variable.

2. Continuous Random Variable.





Discrete variable	Continuous variable
Countable support	Uncountable support
Probability mass function	Probability density function
Probabilities assigned to single values	Probabilities assigned to intervals of values
Each possible value has strictly positive probability	Each possible value has zero probability

1. Discrete Random Variables: Discrete random variables take on a finite or accountably infinite set of possible values. In other words, the range of a discrete random variable is a discrete set of numbers. For example, the number of heads obtained when flipping a coin five times is a discrete random variable that can take on values 0, 1, 2, 3, 4, or 5.

- The probability distribution for a discrete random variable is called a probability mass function (PMF). The PMF gives the probability that the random variable takes on a particular value.

- Example // consider the random variable X which represents the number of heads obtained when flipping a fair coin three times. $S = (TTT, HTT, THT, TTH, HHT, HTH, THH, HHH)$
- P(X = 0) = 1/8 (since there is only one way to get no heads: TTT)
- P(X = 1) = 3/8 (since there are three ways to get one head: HTT, THT, TTH)
- P(X = 2) = 3/8 (since there are three ways to get two heads: HHT, HTH, THH)
- P(X = 3) = 1/8 (since there is only one way to get three heads: HHH)

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

2. Continuous Random Variables: The Continuous random variables take on any value in a continuous range of possible values. In other words, the range of a continuous random variable is an uncountable set of numbers. For example, the height of a person can be referred to as a continuous random variable that can take on any value in a continuous range from zero to infinity. The probability distribution for a continuous random variable is called a probability density function (PDF). Unlike the PMF, the PDF does not give the probability that the random variable takes on a particular value, but rather the probability density at each point in the range of possible values. The probability density can be thought of as the height of a curve at a particular point on the x-axis, and the total area under the curve must equal 1.

- Definition: Probability Density Function
- A function $f(x)$ is a probability density function if
- $f(x) \geq 0$ for each x in its domain,
- $\int_{-\infty}^{\infty} f(x)dx = 1$