RANDOM VARIABLES

What are Algebraic Variables?

In Algebra a variable, like x, is an unknown value(x+5=10=>x=5)

What are Random Variables in Stats and Probability?

A Random Variable is a set of possible values from a random experiment.

Types of Random Variables?

1.discret random Variables 2. Continuous random Variables

2. PROBABILITY DISTRIBUTIONS

1.What are Probability Distributions?

A probability distribution is a list of all of the possible outcomes of a random variable along with their corresponding probability values.

COIN TOSS

COIN TOSS	1(H)	0(TAIL)
PROBABILITY	1/2	1/2

ONE DICE ROLL

DICE ROLL	1	2	3	4	5	6
PROBABILITY	1/6	1/6	1/6	1/6	1/6	1/6

Two dice roll

DICE	1	2	3	4	5	6		Probability			
1	2	3	4	5	6	7		2	1/36	8	5/36
2	3	4	5	6	7	8		3	2/36	9	4/36
3	4	5	6	7	8	9		4	3/36	10	3/36
4	5	6	7	8	9	10		5	4/36	11	2/36
5	6	7	8	9	10	11		6	5/36	12	1/36
6	7	8	9	10	11	12		7	4/36		

Problem with Distribution?

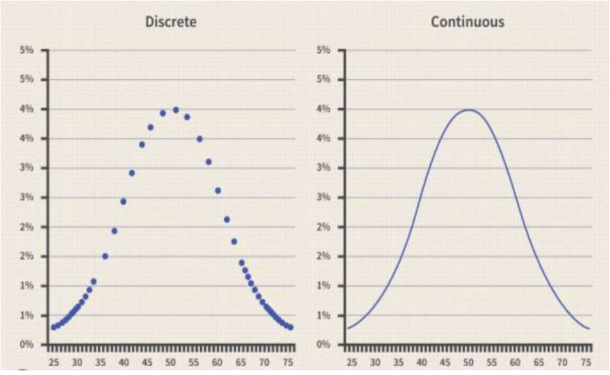
In many scenarios, the number of outcomes can be much larger and hence a table would be tedious to write down. Worse still, the number of possible outcomes could be infinite, in which case, good luck writing a table for that.

Example -Height of people, Rolling 10 dice together

Solution -Function?

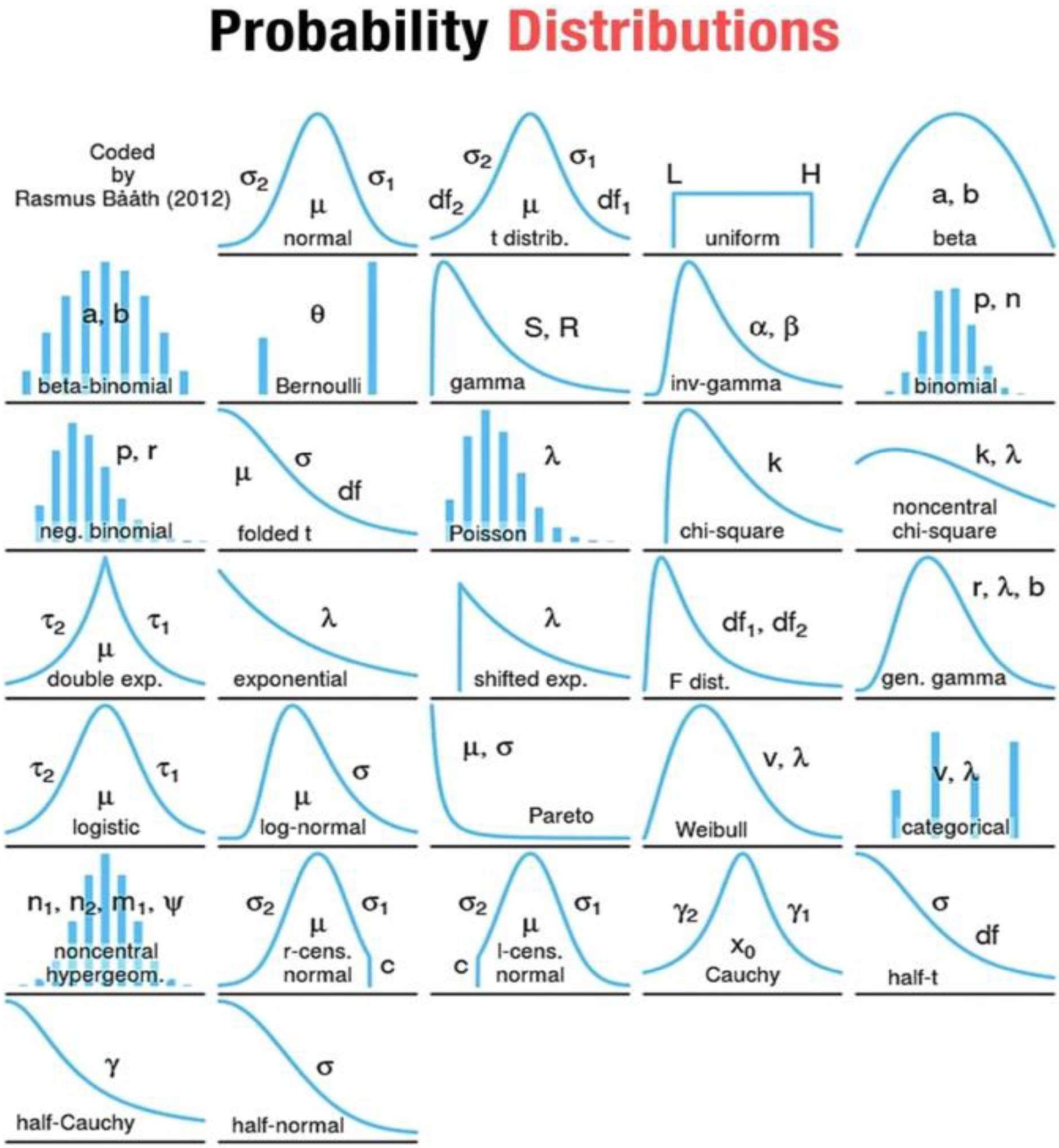
Note -A lot of time Probability Distribution and Probability Distribution Functions are used interchangeably.

What if we use a mathematical function to model the relationship between outcome and probability?



1.Types of Probability Distributions

FAMOUS PROBABILITY DISTRIBUTIONS



Why are Probability Distributions important?

- Gives an idea about the shape/distribution of the data.
- And if our data follows a famous distribution then we automatically know a lot about the data.

A note on Parameters

- Parameters in probability distributions are numerical values that determine the shape, location, and scale of the distribution.
- Different probability distributions have different sets of parameters that determine their shape and characteristics, and understanding these parameters is essential in statistical analysis and inference.

Probability Mass Function (PMF)

PMF stands for Probability Mass Function. It is a mathematical function that describes the probability distribution of a **discrete random variable**.

The PMF of a discrete random variable assigns a probability to each possible value of the random variable.

The probabilities assigned by the PMF must satisfy **two conditions:**

a. The probability assigned to each value **must be non-negative** (i.e., greater than or equal to zero).

b. The sum of the probabilities assigned to all possible values **must equal 1**.

Examples of discrete data

- Nominal (e.g., gender, ethnic background, religious or political affiliation)
- Ordinal (e.g., extent of agreement, school letter grades)
- Quantitative variables with relatively few values (e.g., number of times married)

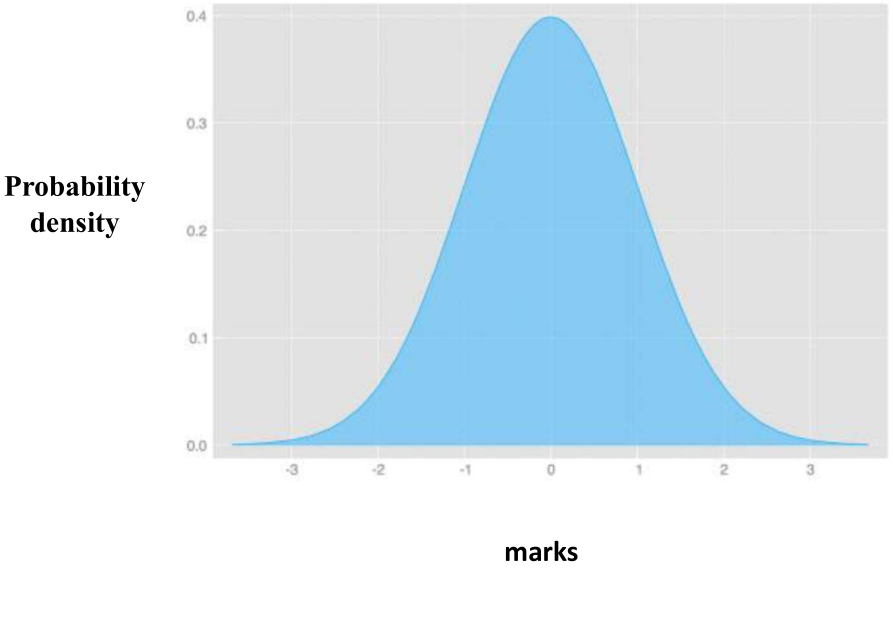
Cumulative Distribution Function(CDF) of PMF

The cumulative distribution function (CDF) F(x) describes the probability that a random variable X with a given probability distribution will be found at a value less than or equal to x **$F(x) = P(X \leq x)$**

PDF stands for Probability Density Function. It is a mathematical function that describes the probability distribution of a **continuous random variable.**

Probability Distribution Functions

PDF stands for Probability Density Function. It is a mathematical function that describes the probability distribution of a **continuous random variable**.



Density Estimation

Density estimation is a statistical technique used to estimate the probability density function (PDF) of a random variable based on a set of observations or data. In simpler terms, it involves estimating the underlying distribution of a set of data points.

Density estimation can be used for a variety of purposes, such as **hypothesis testing, data analysis, and data visualization**. It is particularly useful in areas such as **machine learning**, where it is often used to estimate the probability distribution of input data or to model the likelihood of certain events or outcomes.

There are various methods for density estimation, including **parametric** and **non-parametric approaches**. Parametric methods assume that the data follows a specific probability distribution (such as a normal distribution), while non-parametric methods do not make any assumptions about the distribution and instead estimate it directly from the data.

Commonly used techniques for density estimation include **kernel density estimation (KDE)**, **histogram estimation**, and **Gaussian mixture models (GMMs)**. The choice of method depends on the specific characteristics of the data and the intended use of the density estimate.

Parametric Density Estimation

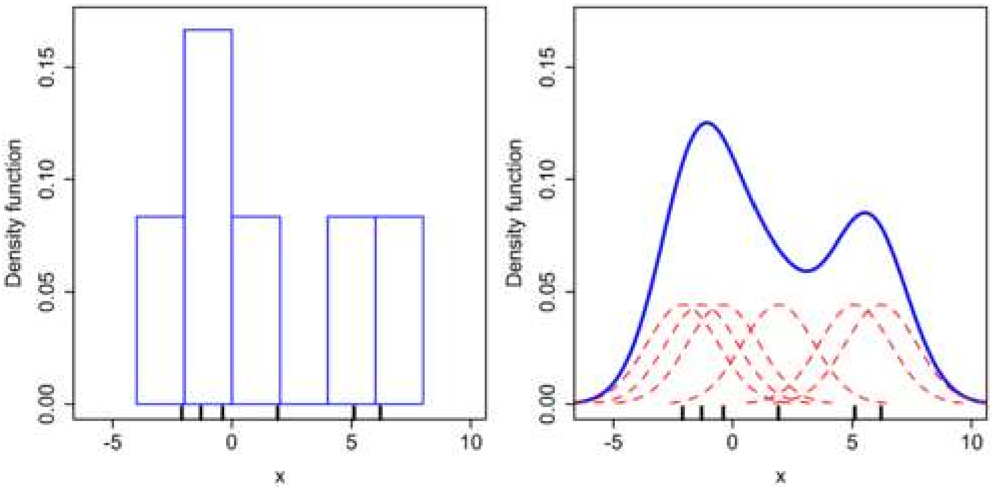
Parametric density estimation is a method of estimating the probability density function (PDF) of a random variable by assuming that the underlying distribution belongs to a specific parametric family of probability distributions, such as the normal, exponential, or Poisson distributions.

Non-Parametric Density Estimation (KDE)

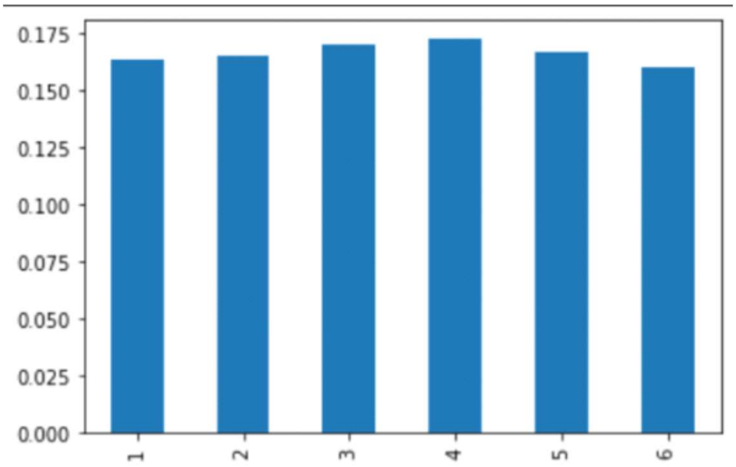
- But sometimes the distribution is not clear or it's not one of the famous distributions.
- Non-parametric density estimation is a statistical technique used to estimate the probability density function of a random variable without making any assumptions about the underlying distribution. It is also referred to as non-parametric density estimation because it does not require the use of a predefined probability distribution function, as opposed to parametric methods such as the Gaussian distribution.
- The non-parametric density estimation technique involves constructing an estimate of the probability density function using the available data. This is typically done by creating a **kernel density estimate**
- Non-parametric density estimation has several advantages over parametric density estimation. One of the main **advantages** is that it **does not require the assumption of a specific distribution**, which allows for more flexible and accurate estimation in situations where the underlying distribution is unknown or complex. However, non-parametric density estimation can be **computationally intensive and may require more data to achieve accurate estimates** compared to parametric methods.

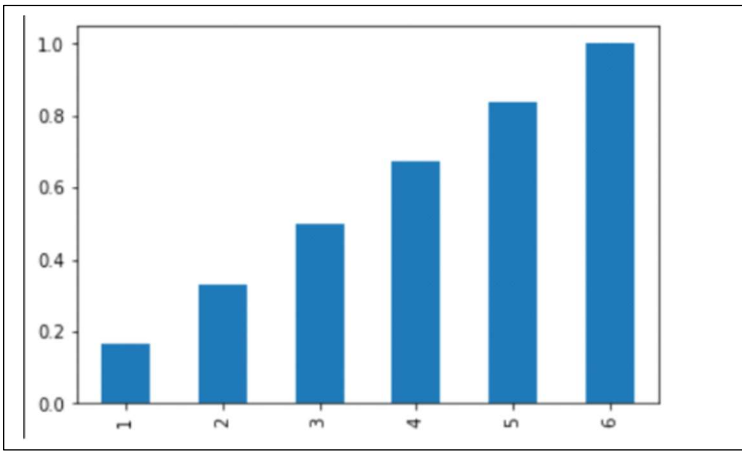
Kernel Density Estimate(KDE)

The KDE technique involves using a kernel function to smooth out the data and create a continuous estimate of the underlying density function.

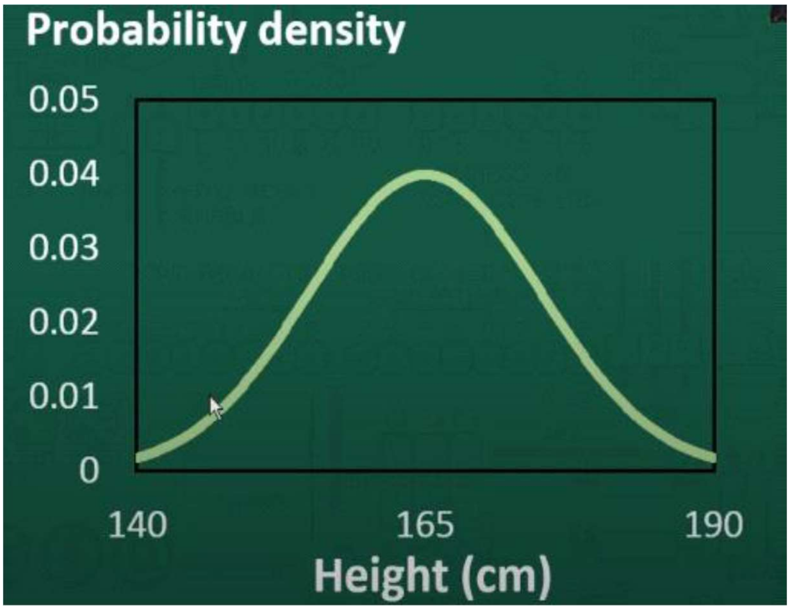


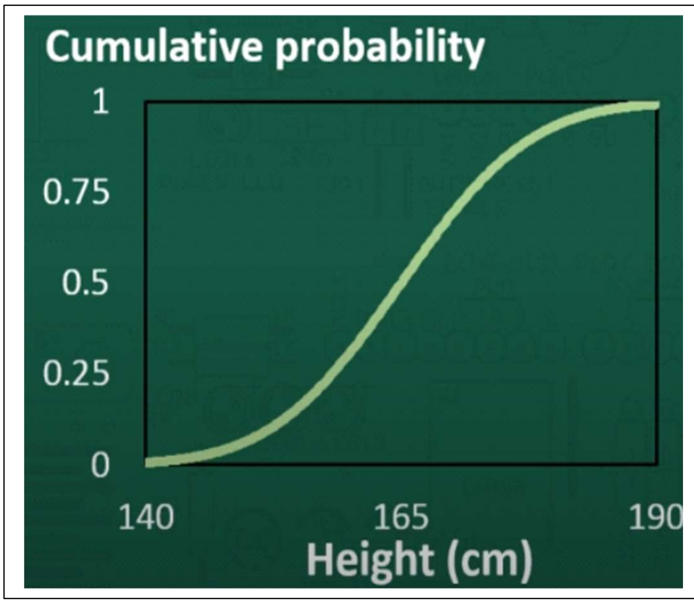
Cumulative Distribution Function(CDF) of PDF





How to use PDF and CDF in Data Analysis


Probability density
Height (cm)


Cumulative probability
Height (cm)