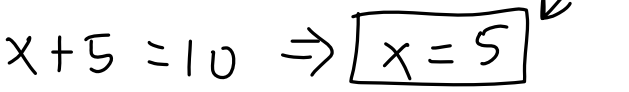
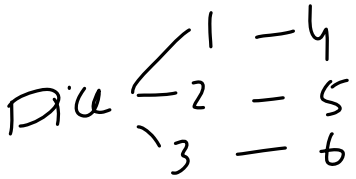
Random Variables

15 March 2023 11:43

- What are Algebraic Variables?

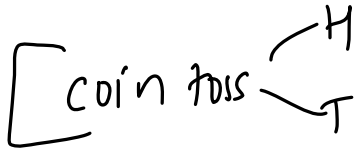
  In Algebra a variable, like x, is an unknown value

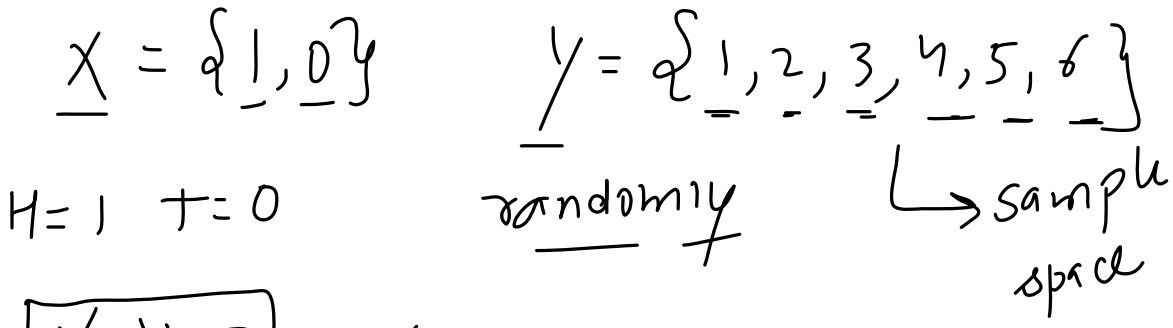




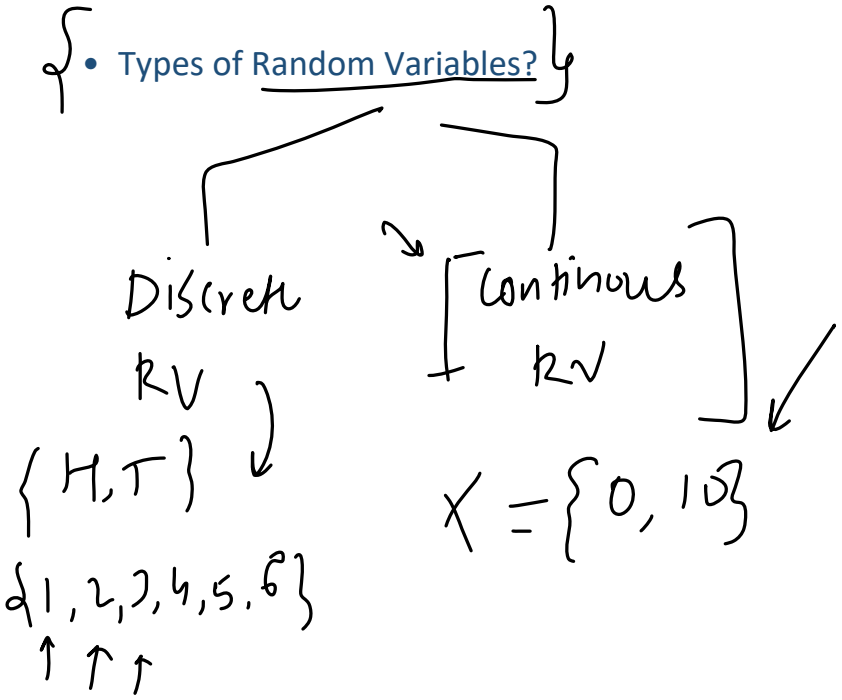
- What are Random Variables in Stats and Probability?

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





- Types of Random Variables?



Probability Distributions

15 March 2023 11:53

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	1 (H)	0 (T)
probab	$\frac{1}{2}$	$\frac{1}{2}$

2 dice →

2 3 4 5 6 7 8 9

dice

1	2	3	4	5	6
1/6	1/6	1/6	1/6	1/6	1/6

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

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

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?

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

X → outcome
Y → prob

1	1.5	2	3	4	5	6
1/6		1/6	1/6	1/6	1/6	1/6

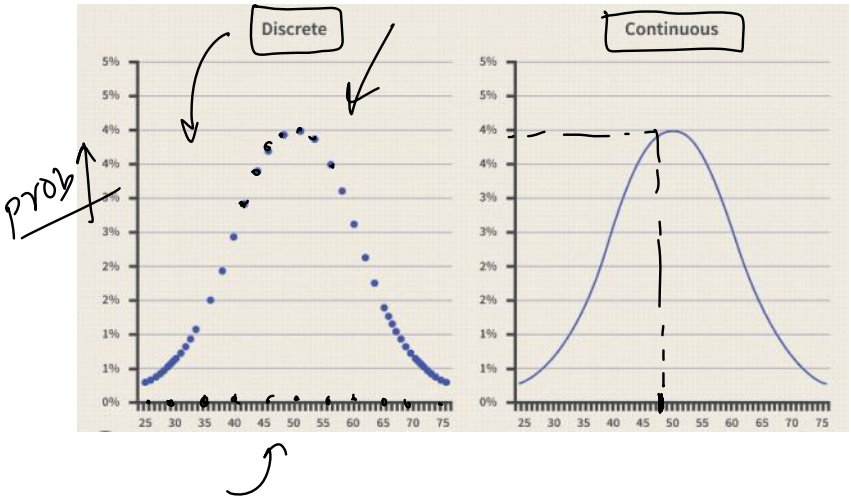
$Y = f(X)$

PDF → table → f(X)

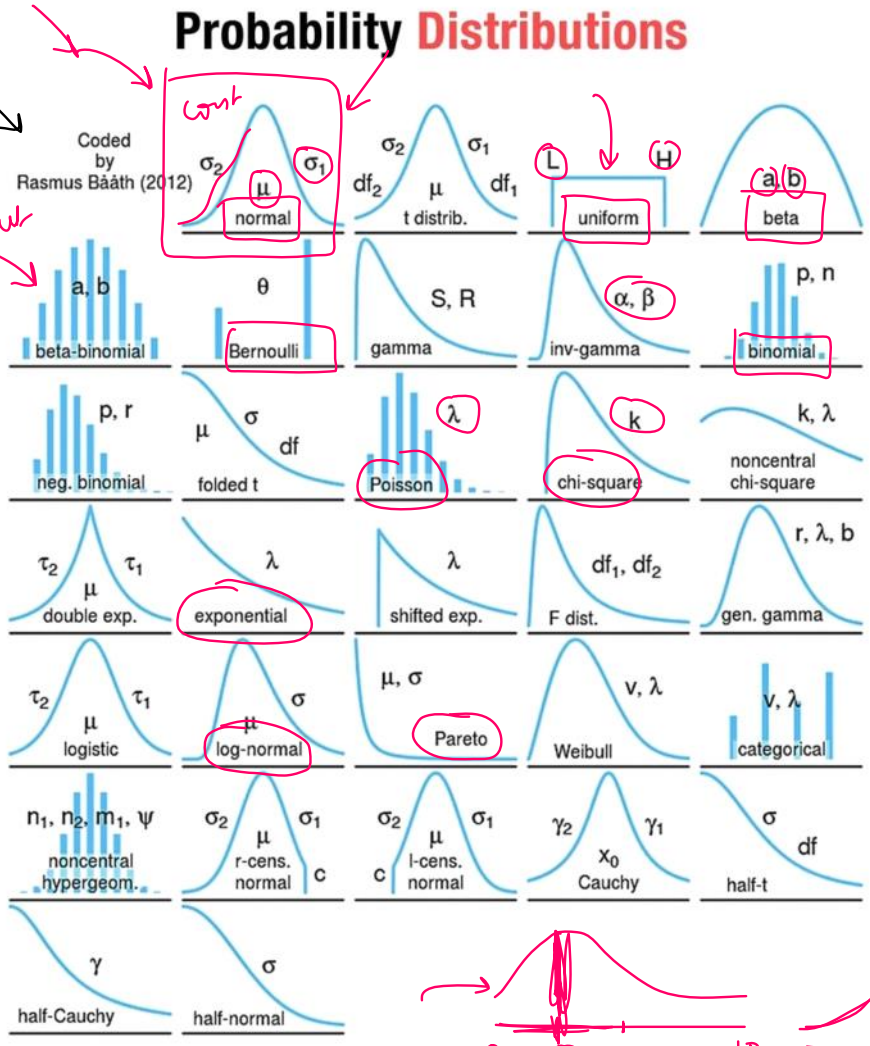
∫ Note - A lot of time Probability Distribution and Probability Distribution Functions are ↴

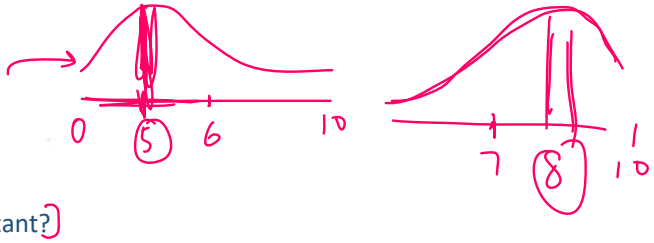
Note - A lot of time Probability Distribution and Probability Distribution Functions are used interchangeably.

1. Types of Probability Distributions (PDF)



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 (PDF)

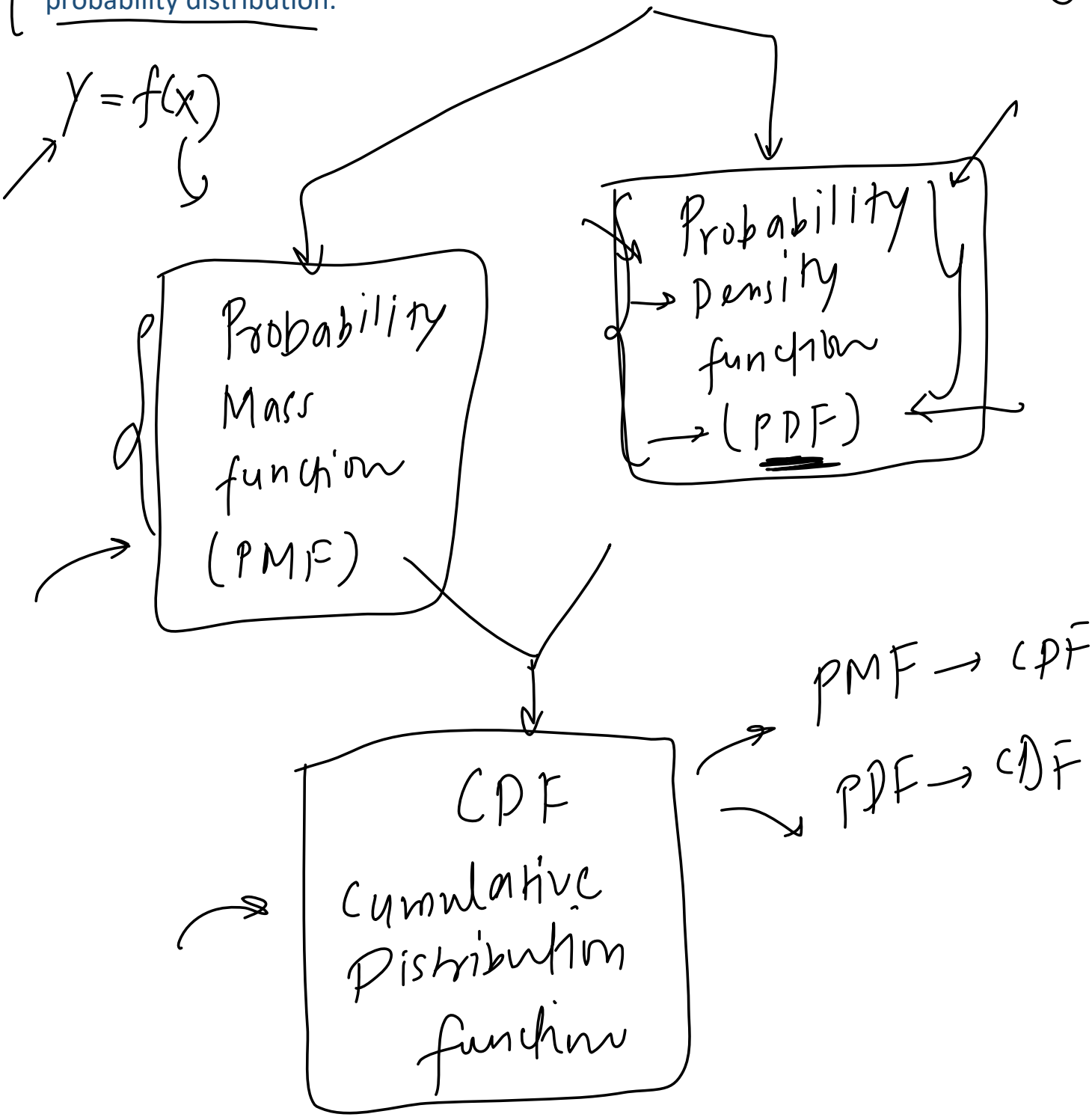
Parameters in probability distributions are numerical values that determine the shape, location, and scale of the distribution.

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 Distribution Functions] → PDF

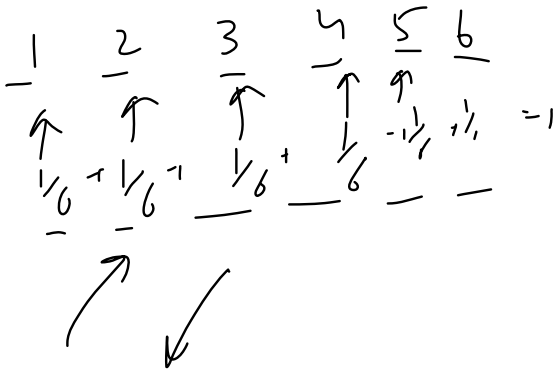
15 March 2023 20:08

A probability distribution function (~~PDF~~) is a mathematical function that describes the probability of obtaining different values of a random variable in a particular probability distribution.



Probability Mass Function (PMF)

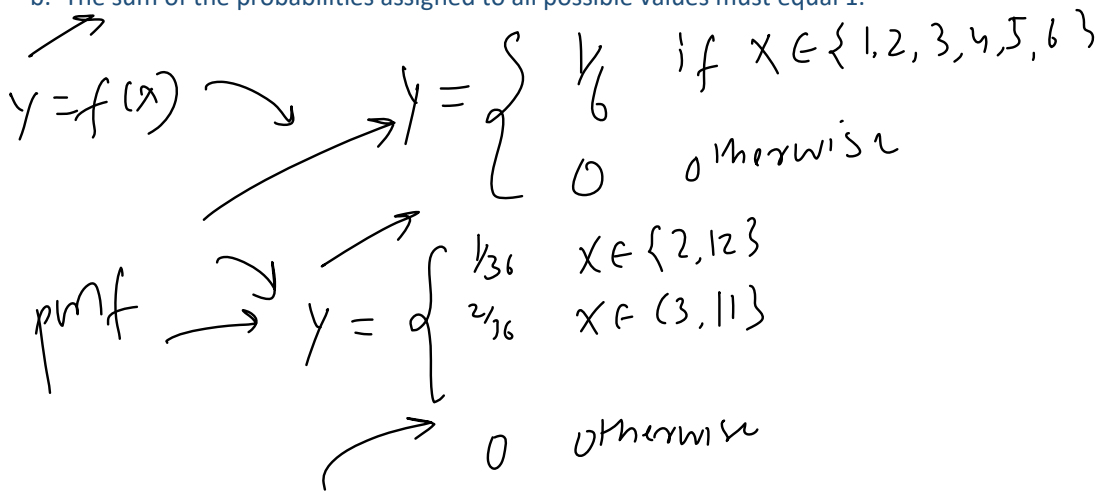
15 March 2023 15:25



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

https://en.wikipedia.org/wiki/Bernoulli_distribution

https://en.wikipedia.org/wiki/Binomial_distribution

Cumulative Distribution Function(CDF) of PMF

15 March 2023 20:09

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 <= x)]

f(x) =

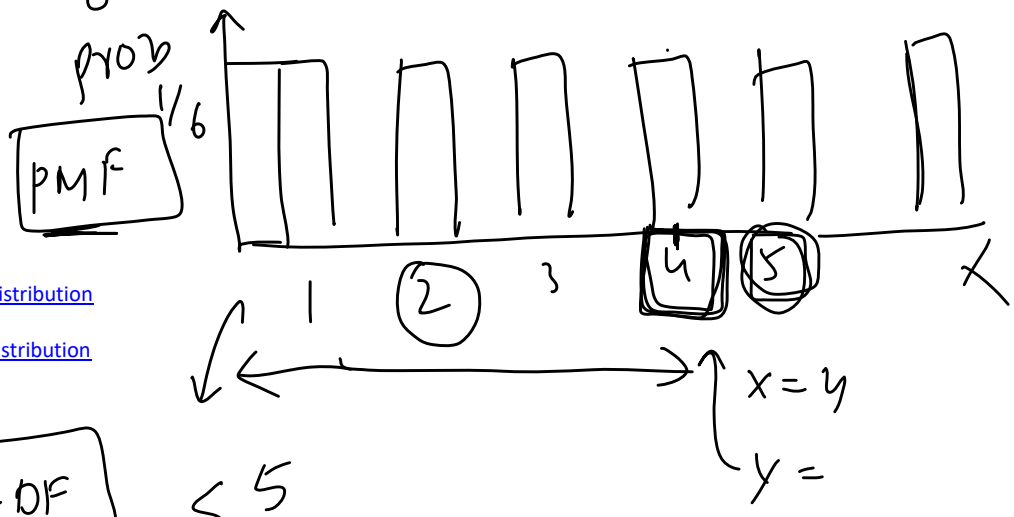
PMF f(x)

f(X=6) = 1/6

f(X=5) = 1/6

f(X≤4) = f(X=4) + f(X=3) + f(X=2) + f(X=1)

1/6 + 1/6 + 1/6 + 1/6



Examples:

https://en.wikipedia.org/wiki/Bernoulli_distribution

https://en.wikipedia.org/wiki/Binomial_distribution

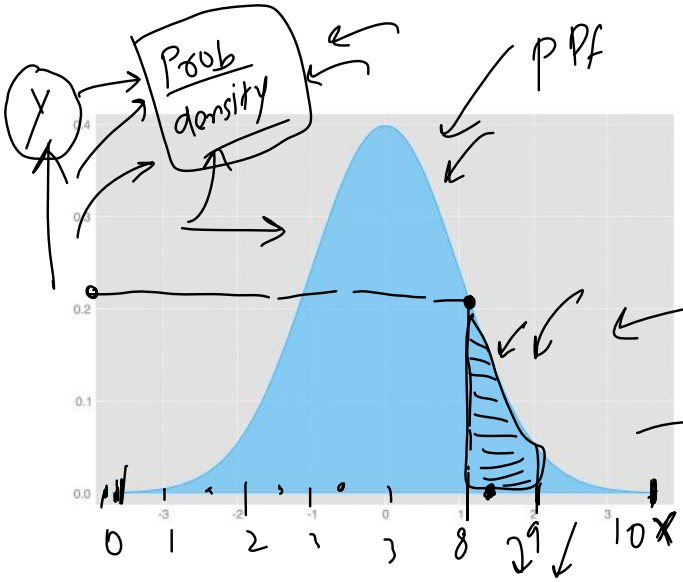
CDF ≤ 5

	PMF	CDF
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6 → 1

Probability Density Function (PDF)

15 March 2023 15:25

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



sample PDF

cgpa = 7.912

7.912345 fg

$\int_8^9 f(x)\,dx$

$P(0 \le x \le 10) = 1$

area

10 (marks)

0 1 2 3 8 9

8 → 9

probability between

8 → 8.1

8 → 8.01

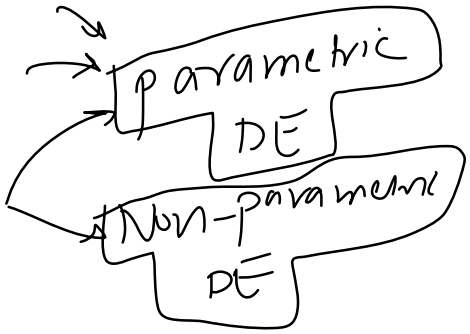
8 → 8.001

1. Why Probability Density and why not Probability?
2. What does the area of this graph represents?
3. How to calculate Probability then?
4. Examples of PDF
   a. https://en.wikipedia.org/wiki/Normal_distribution
   b. https://en.wikipedia.org/wiki/Log-normal_distribution
   c. https://en.wikipedia.org/wiki/Poisson_distribution
5. How is graph calculated?

Refer to chatgpt chat in the folder to understand all these questions

Density Estimation

16 March 2023 06:54



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.

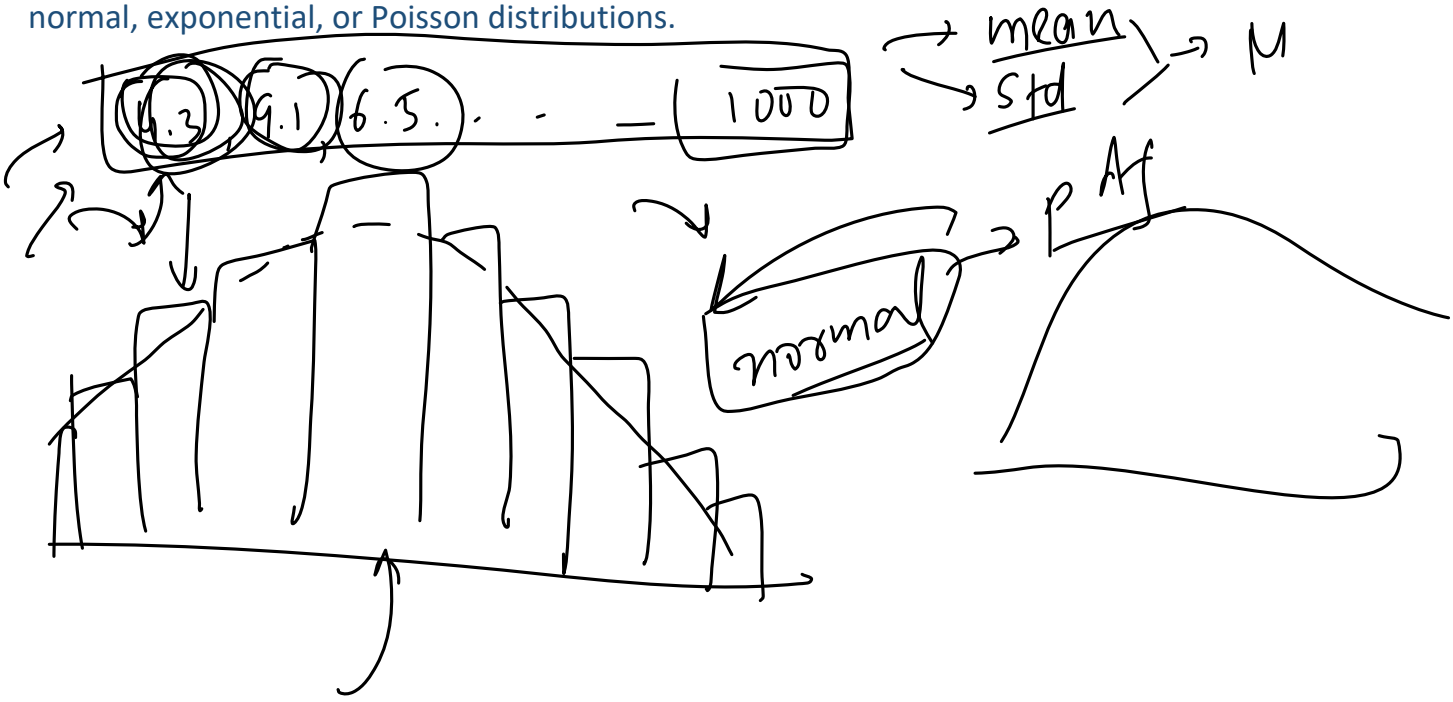
There are various methods for density estimation, including **parametric** and **non-parametric approaches**. Parametric methods assume that the data follows a specific probability distribution (such as a normal distribution), while non-parametric methods do not make any assumptions about the distribution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

16 March 2023 06:54

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.



Consider we have a sample data now to check whether sample matches with the famous distribution we will plot histogram lets say in this example it matches with the normal distribution.
then we will first calculate the sample mean and sample std then infer for the population mean and std. then to get density estimation we will fit the values in the formula of pdf of normal distribuion

Non-Parametric Density Estimation (KDE)

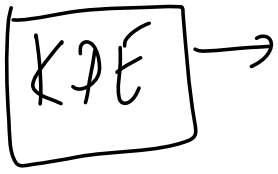
16 March 2023 06:55

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.

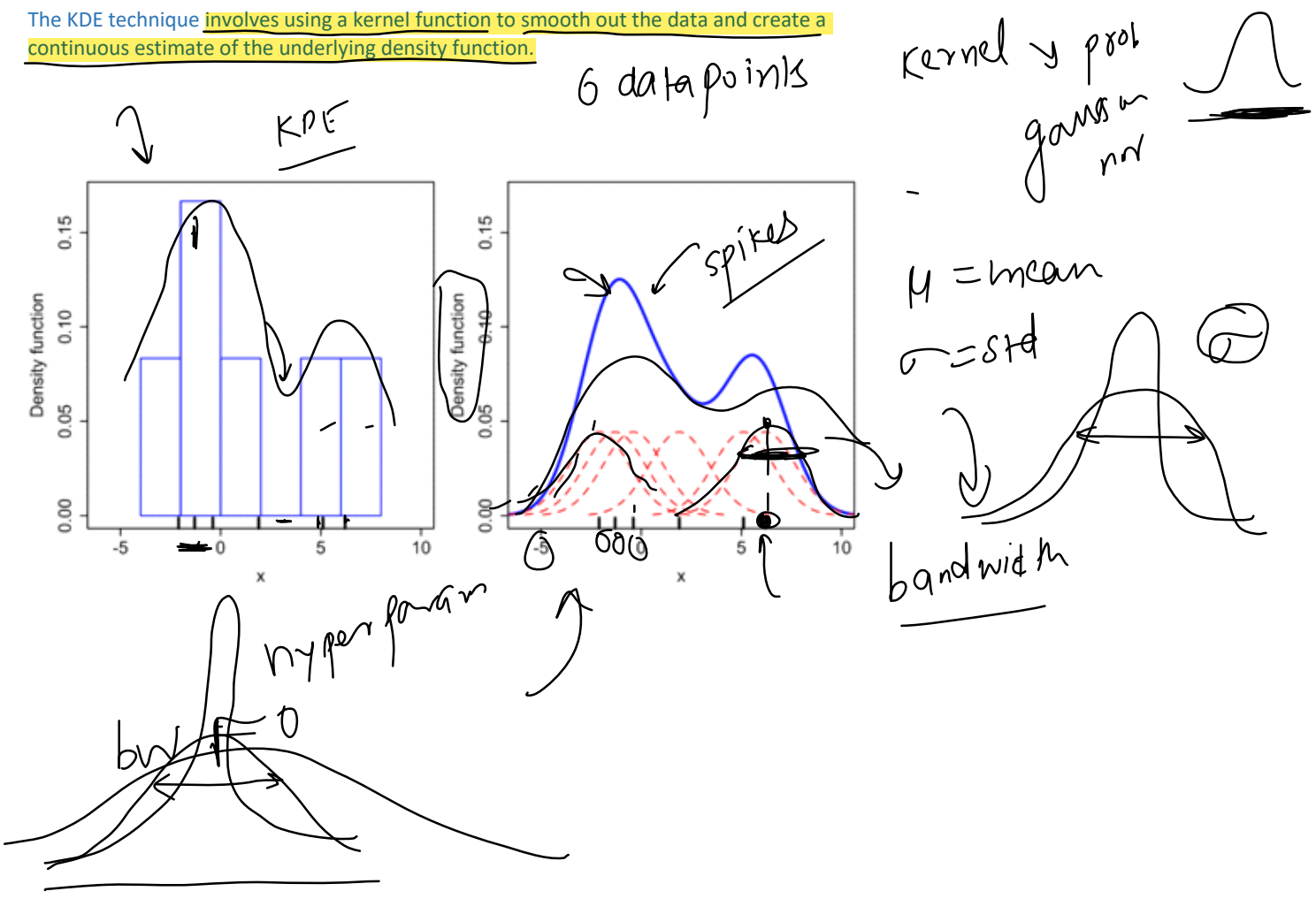


it means in the parametric we were using parameters such as mean and std but in non parametric we will use data points for density estimation

Kernel Density Estimate(KDE)

16 March 2023 16:08

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

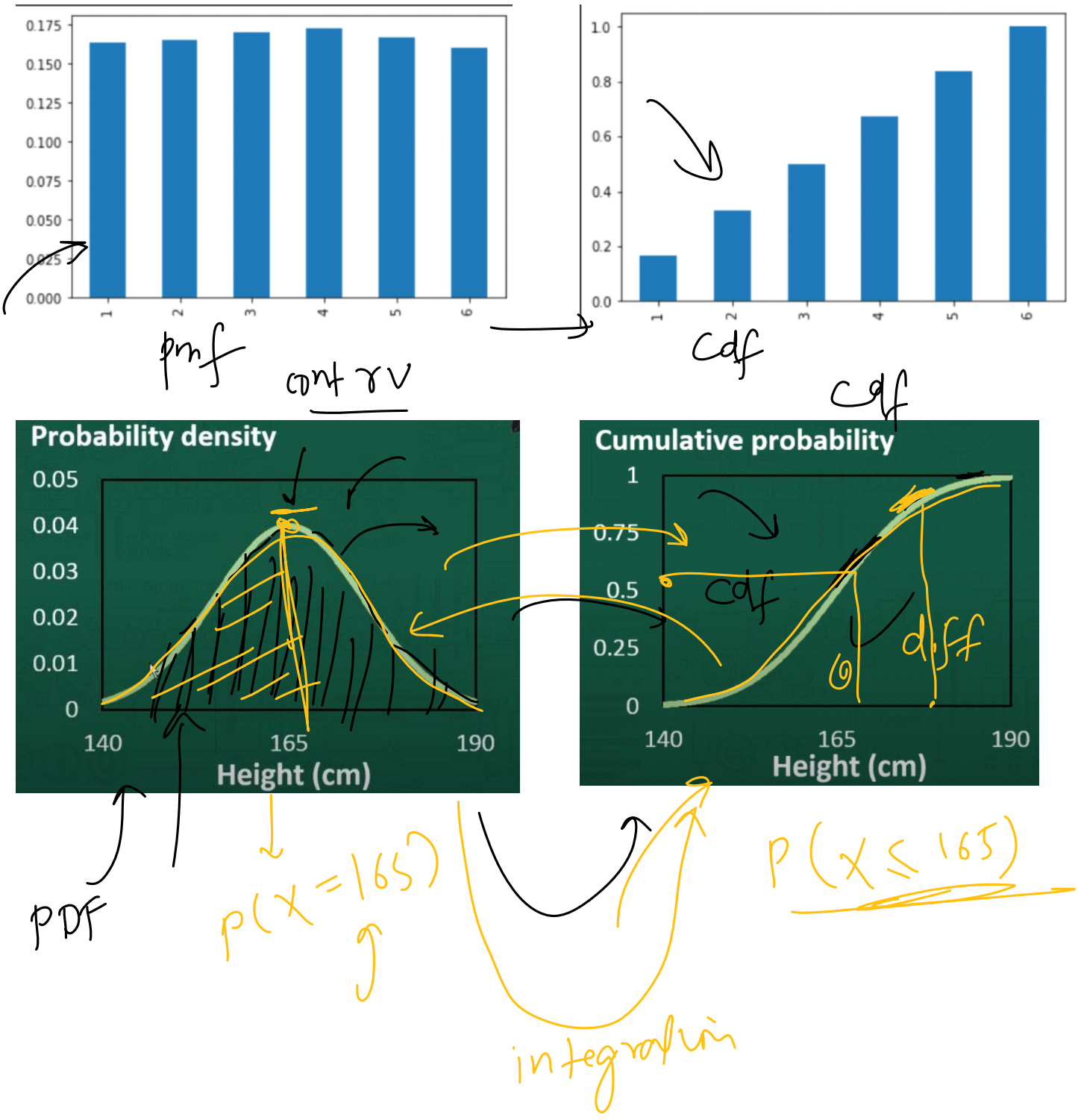


Explaination - Since the above distribution does not match any famous distribution here we will use non parametric.(kde).Here, you will use a kernel (Kernel is basically probability distribution generally Gaussian) .In this what you will do is you will grab each data point considering them as center (mean) and you will make a gaussian distribution for each data point.and now you will grab each point and move in perpendicular direction and us point ke upar jitne bhi norma curve aare hai usko add krenge unke distance ko and you will have your y value

Important note- since you are using gaussian as kernel we know it has two parameters that is mean and std. we know that in case of kde the mean is point on x axis itself , but what about std , thats why std is hyperparameter known as bandwidth if bandwidth is very low then it will show peakedness as all points are closed to mean . then what will happen is the blue curve will start having spikes if bandwidth increases then the curve will be more spreaded out and the y values will be less and overall blue curve will start to smooth

Cumulative Distribution Function(CDF) of PDF

15 March 2023 15:25



How to use PDF and CDF in Data Analysis

15 March 2023 20:10

2D Probability Density Plots

16 March 2023 06:50