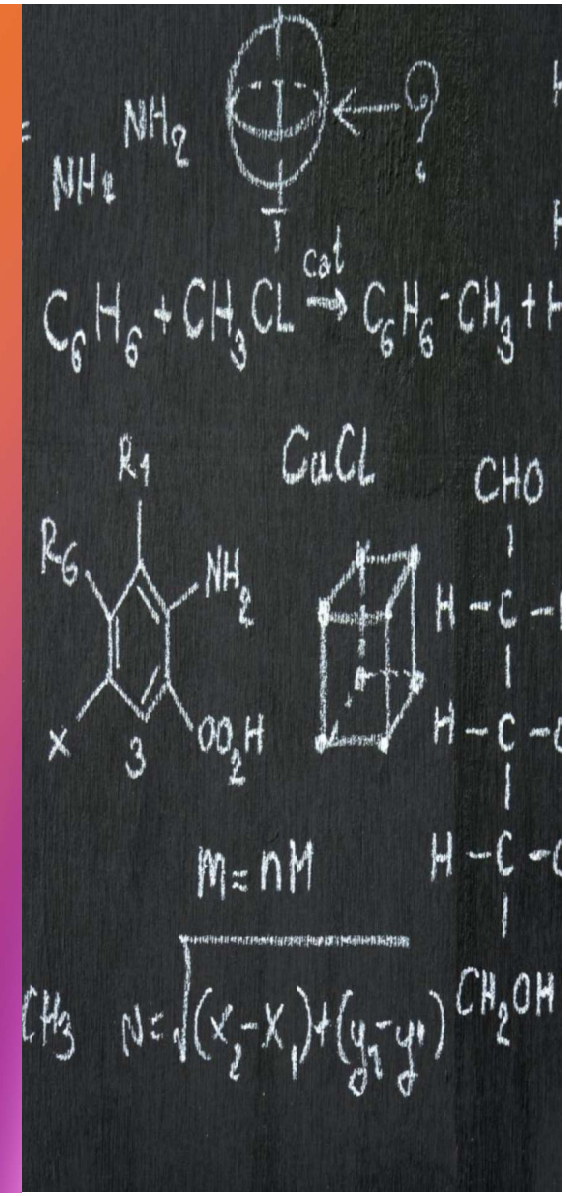
Unit-2

Probability and statistics for Data Science

Random Variable

In probability, a random variable is a real valued function whose domain is the sample space of the random experiment. It means that each outcome of a random experiment is associated with a single real number, and the single real number may vary with the different outcomes of a random experiment. Hence, it is called a random variable and it is generally represented by the letter “X”.



Example

let us consider an experiment for tossing a coin Thrice.

Hence, the sample space for this experiment is

S = {HHH, HHT, HTH, THH,HTT,THT,TTH,TTT}

S1, S2, S3, S4, S5, S6, S7, S8

If X is a random variable and it denotes the number of heads obtained, then the values are represented as follows:

X(S1) = 3; Sample space1 where head comes three times,

X(S2) = X(S3)=X(S4)=2

X(S5) = = X(S6)=X(S7)=1

X(S8) = 0.

Probability Distribution of a Random Variable

The probability distribution of a random variable X for the system of numbers is defined as follows:

$$\sum_{i=1}^{n} p_i = 1$$

Where, $p_i > 0$, and i= 1, 2, 3, ..., n.

The real numbers x_1, x_2, x_3,...x_n are the possible values of the random variable X, and p_1, p_2, p_3, ...p_n are the probabilities of the random variable X that takes the value x_i.

Therefore, $P(X = x_i) = p_i$.

(**Note**: The sum of all the probabilities in the probability distribution should be equal to 1)

X :	x_1	x_2	...	x_n
P(X) :	p_1	p_2	...	p_n

Probability Distribution of a Random Variable

S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}

S1, S2, S3, S4, S5, S6, S7, S8

X(S1) = 3; Sample space1 where head comes three times,

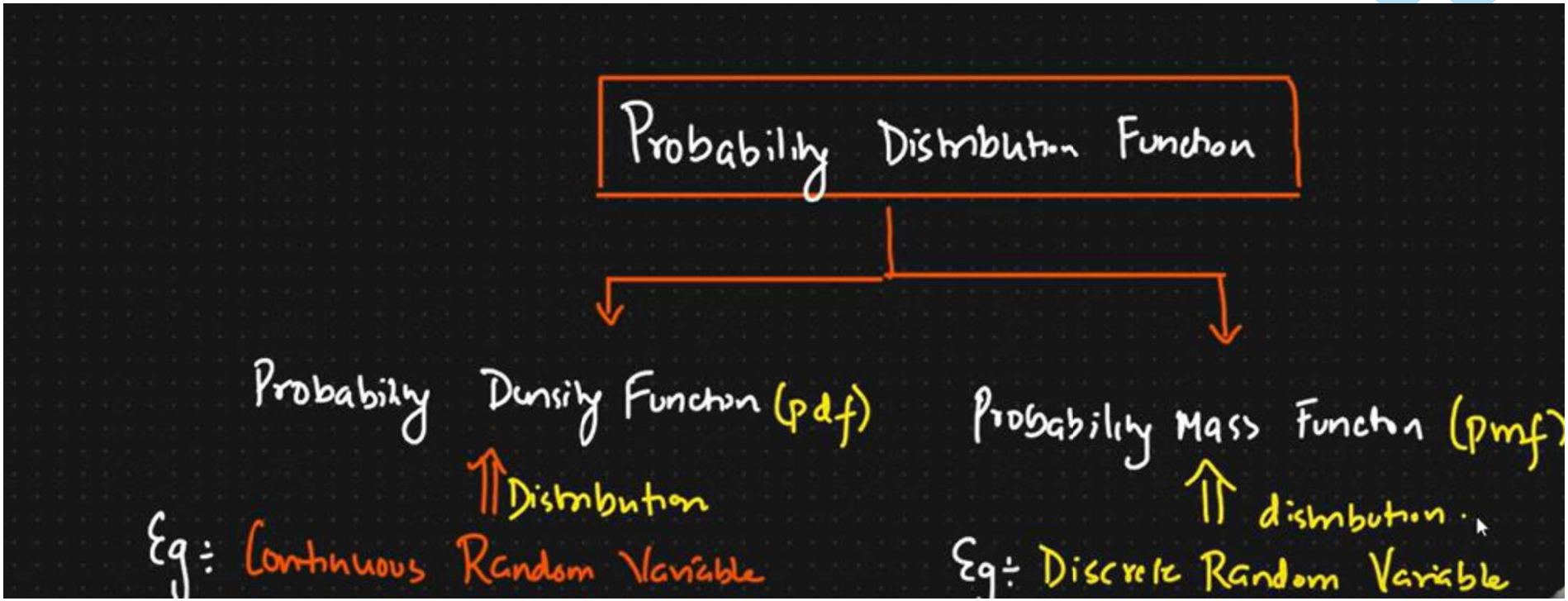
X(S2) = X(S3)=X(S4)=2

X(S5) = = X(S6)=X(S7)=1

X(S8) = 0.

X : No. of head	0	1	2	3
P(X) :				

X : No. of head	0	1	2	3
P(X) :	1/8	3/8	3/8	1/8


Probability Distribution Function
Probability Density Function (pdf)
Distribution
Eg: Continuous Random Variable
Probability Mass Function (pmf)
distribution
Eg: Discrete Random Variable

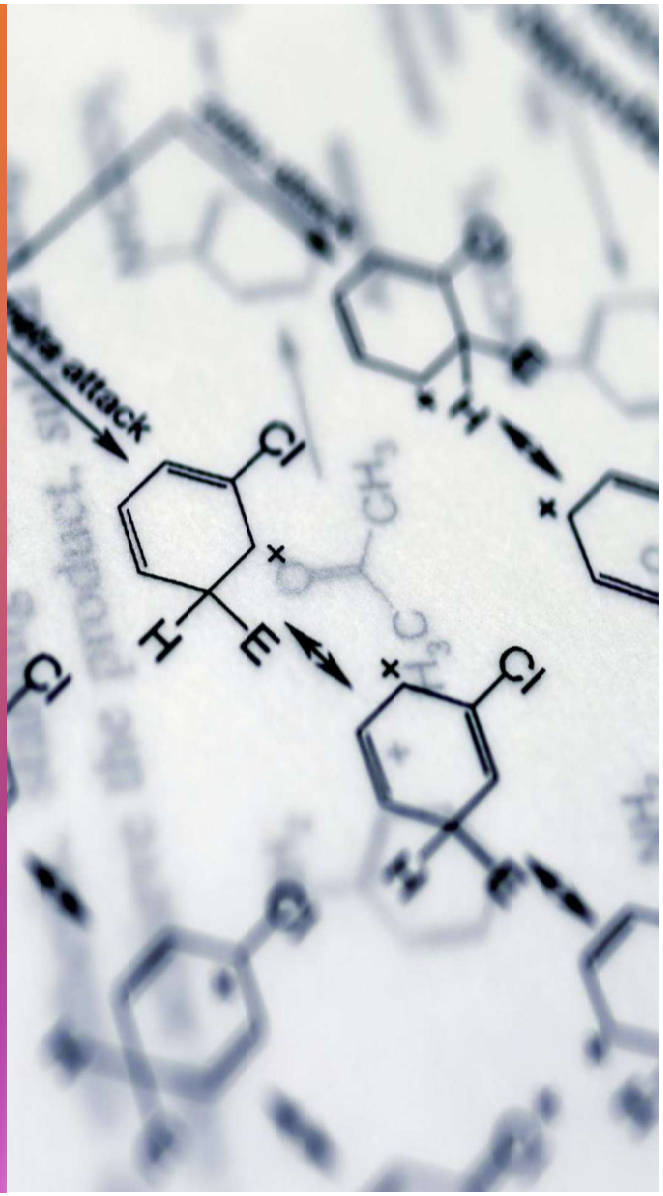
Types of Random Variable

There are two random variables, such as:

- Discrete Random Variable
- Continuous Random Variable

Let's understand these types of variables in detail along with suitable examples below.

Discrete random variable can take only a finite number of distinct values such as 0, 1, 2, 3, 4, ... and so on. The probability distribution of a random variable has a list of probabilities compared with each of its possible values known as probability mass function.



Continuous Random Variable

A numerically valued variable is said to be continuous if, in any unit of measurement, whenever it can take on the values a and b. If the random variable X can assume an infinite and uncountable set of values, it is said to be a continuous random variable. When X takes any value in a given interval (a, b), it is said to be a continuous random variable in that interval.

Probability Mass Function (PMF)

X : No. of head	0	1	2	3
P(X) :	1/8	3/8	3/8	1/8

The **Probability Mass Function (PMF)** is also called a **probability function** or **frequency function** which characterizes the distribution of a discrete random variable. Let X be a discrete random variable of a function, then the probability mass function of a random variable X is given by:

$\mathbf{P_x(x) = P(X=x)}$, For all x belongs to the range of X

It is noted that the probability function should fall on the condition

$\mathbf{P_x(x) \geq 0}$ **and**

$\mathbf{\Sigma_{x \epsilon Range(x)} P_x(x) = 1}$

Distribution function

In Probability and Statistics, the Cumulative Distribution Function (CDF) of a real-valued random variable, say "X", which is evaluated at x, is the probability that X takes a value less than or equal to the x.

$f(x) = \sum Pi = P(x \leq xj)$

X : No. of head	0	1	2	3
P(X) :	1/8	3/8	3/8	1/8

F (x)= 1/8 $x \leq 0$
F(x) = ? $x \leq 1$

= 4/8

F (x)= = ? $x \leq 2$
7/8

F(x) = 1 $x \leq 3$

Mean of a Random Variable

If X is a random variable, and its possible values are x_1, x_2, x_3,...x_n associated with the probabilities p_1, p_2, p_3, ..p_n, respectively, then the mean of the random variable X is given by the formula:

$E(X) = \mu = \sum_{i=1}^{n} x_i p_i$

The mean of the random variable (μ) is also called the expectation of the random variable E(X).

The mean of the random variable X can also be represented by

$E(x) = x_1p_1+x_2p_2+x_3p_3+.....+x_np_n$

Thus, the mean or the expectation of the random variable X is defined as the sum of the products of all possible values of X by their respective probability values.

Example: **Find the probability distribution for the number of doublets in the three throws of a pair of dice.**

Let X be a random variable and it denotes the number of doublets. Hence, the possible number of doublets are (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), and (6, 6)

Given that, X can take the values 0, 1, 2 or 3.

If a pair of dice is thrown, the number of sample space S is 36.

Therefore, the probability of getting a doublet

$= 6/36 = 1/6$.

The probability of not getting a doublet =

$1 - (1/6) = 5/6$.

Solution Continue...

Therefore,

The probability of no doublet, $P(X=0) = (5/6)(5/6)(5/6) = 125/216$.

The probability of one doublet and two non-doublet, $P(X=1)$

$= (1/6)(5/6)(5/6) + (5/6)(1/6)(5/6) + (5/6)(5/6)(1/6) = 75/216$.

The probability of two doublets and one non-doublet, $P(x=2)$

$= (1/6)(1/6)(5/6) + (1/6)(5/6)(1/6) + (5/6)(1/6)(1/6) = 15/216$

The probability of three doublets, $P(X=3)$

$= (1/6)(1/6)(1/6) = 1/216$.

Therefore, the required probability distribution is

X	0	1	2	3
P(X)	125/216	75/216	15/216	1/216

Verification:

We know that the sum of all the probabilities in the probability distribution is 1.

$= (125/216)+(75/216)+(15/216)+(1/216)$
$= 216/216 = 1$

Example: Assume that the pair of dice is thrown and the random variable X is the sum of numbers that appears on two dice. Find the mean or the expectation of the random variable X.

Solution: If two dice are thrown, then the total number of sample spaces obtained is
36.

Given that, the random variable X is the sum of numbers that appear on two dice, such as 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 or 12.

Therefore,

P(X= 2) = 1/36 P(X= 3) = 2/36 P(X= 4) = 3/36 P(X= 5) = 4/36 P(X= 6) = 5/36

P(X= 7) = 6/36 P(X= 8) = 5/36 P(X= 9) = 4/36 P(X= 10) = 3/36

P(X= 11) = 2/36 P(X= 12) = 1/36

Hence, the probability distribution of the random variable X is:

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Therefore,

The mean or the expectation of the random variable X is:

$E(X) = \mu = \sum_{i=1}^{n} x_i p_i$

$= 2(1/36) + 3(2/36) + 4(3/36) + 5(4/36) + 6(5/36) + 7(6/36) + 8(5/36) + 9(4/36) + 10(3/36) + 11(2/36) + 12(1/36)$

$= (2+6+12+20+30+42+40+36+30+22+12)/36$

$=7$

Therefore, the mean of the random variable X is 7.

Question : Let X be a random variable, and $P(X=x)$ is the PMF given by,

X	0	1	2	3	4	5	6	7
$P(X=x)$	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2+k$

1. Determine the value of k
2. Find the probability (i) $P(X \leq 6)$, (ii) $P(3<x \leq 6)$

(1) We know that;

$\sum P(x_i)=1$

Therefore,

$0 + k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$

$9k + 10k^2 = 1$

$10k^2 + 9k - 1 = 0$

$10k^2 + 10k - k - 1 = 0$

$10k(k + 1) - 1(k + 1) = 0$

$(10k - 1)(k + 1) = 0$

So, $10k - 1 = 0$ and $k + 1 = 0$

Therefore, $k = 1/10$ and $k = -1$

$k=-1$ is not possible because the probability value ranges from 0 to 1.

Hence, the value of k is 1/10.

(2) (i) $P(X \leq 6) = 1 - P(x > 6)$

$= 1 - (7k^2+k)$

$= 1 - (7(1/10)^2 + (1/10))$

$= 1 - (7/100 + 1/10)$

$= 1 - (17/100)$

$= (100 - 17)/100$

$= 83/100$

Therefore, $P(X \leq 6) = 83/100$

What is the Probability Density Function?

The Probability Density Function(PDF) defines the probability function representing the density of a continuous random variable lying between a specific range of values. In other words, the probability density function produces the likelihood of values of the continuous random variable. Sometimes it is also called a probability distribution function or just a probability function.

Probability Density Function Formula (PDF)

$$P(a < X < b) = \int_a^b f(x)\,dx$$

Or

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

In the case of a continuous random variable, the probability taken by X on some given value x is always 0. In this case, if we find P(X = x), it does not work. Instead of this, we must calculate the probability of X lying in an interval (a, b). Now, we have to figure it for P(a< X< b), and we can calculate this using the formula of PDF.

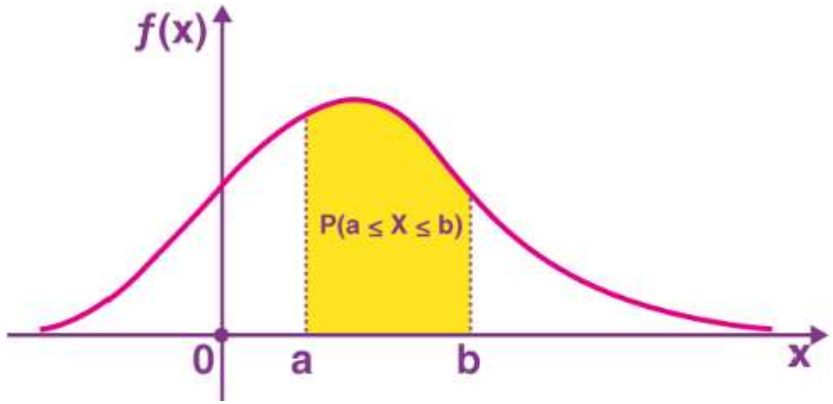
Probability Density Function Formula (Continue...)

This is because, when X is continuous, we can ignore the endpoints of intervals while finding probabilities of continuous random variables. That means, for any constants a and b,

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

Probability Density Function Graph

The probability density function is defined as an integral of the density of the variable density over a given range. It is denoted by f (x). This function is positive or non-negative at any point of the graph, and the integral, more specifically the definite integral of PDF over the entire space is always equal to one. The graph of PDFs typically resembles a bell curve, with the probability of the outcomes below the curve. This picture depicts the graph of a probability density function for a continuous random variable x with function f(x).



Probability Density Function Properties

- Let x be the continuous random variable with density function f(x), and the probability density function should satisfy the following conditions: $P(x) = \int_a^b f(x)\,dx$
- For a continuous random variable that takes some value between certain limits, say a and b, the PDF is calculated by finding the area under its curve and the X-axis within the lower limit (a) and upper limit (b).

Mean of Probability Density Function

Mean of the probability density function refers to the average value of the random variable. The mean is also called as expected value or expectation. It is denoted by μ or E[X] where, X is random variable.

Mean of the probability density function f(x) for the continuous random variable X is given by:

$$\mathbf{E[X]} = \mu = \int_{-\infty}^{\infty} \mathbf{xf(x)dx}$$

Median of Probability Density Function

Median is the value which divides the probability density function graph into two equal halves. If x = M is the median then, area under curve from -∞ to M and area under curve from M to ∞ are equal which gives the median value = 1/2.

Median of the probability density function f(x) is given by:

$$\int_{-\infty}^{\mathbf{M}} \mathbf{f(x)dx} = \int_{\mathbf{M}}^{\infty} \mathbf{f(x)dx} = \frac{1}{2}$$

Variance Probability Density Function

Variance of probability density function refers to the squared deviation from the mean of a random variable. It is denoted by Var(X) where, X is random variable.

Variance of the probability density function f(x) for continuous random variable X is given by:

$$Var(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Standard Deviation of Probability Density Function

Standard Deviation is the square root of the variance. It is denoted by σ and is given by:

$$\sigma = \sqrt{Var(X)}$$

Probability Density Function Properties...

- The probability density function is non-negative for all the possible values,
- i.e. **f(x)≥ 0**, for all x.
- The area between the density curve and horizontal X-axis is equal to 1, i.e.

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

- Due to the property of continuous random variables, the density function curve is **continued** for all over the given range. Also, this defines itself over a range of continuous values or the domain of the variable.

Probability Density Function Example

Example:

Let X be a continuous random variable with the PDF given by:

$$f(x) = \begin{cases} x; & 0 < x < 1 \\ 2 - x; & 1 < x < 2 \\ 0; & x > 2 \end{cases}$$

Find $P(0.5 < x < 1.5)$.

Given PDF is:

$$f(x) = \begin{cases} x;\ 0 < x < 1 \\ 2 - x;\ 1 < x < 2 \\ 0;\ x > 2 \end{cases}$$

$P(0.5 < X < 1.5) = \int_{0.5}^{1.5} f(x)dx$

Let us split the integral by taking the intervals as given below:

$- \int_{0.5}^{1} f(x)dx + \int_{1}^{1.5} f(x)dx$

Substituting the corresponding values of f(x) based on the intervals, we get;

$= \int_{0.5}^{1} xdx + \int_{1}^{1.5} (2 - x)dx$

Integrating the functions, we get;

$= \left(\frac{x^2}{2}\right)_{0.5}^{1} + \left(2x - \frac{x^2}{2}\right)_{1}^{1.5}$

$= [(1)^2/2 - (0.5)^2/2] + \{[2(1.5) - (1.5)^2/2] - [2(1) - (1)^2/2]\}$

$= [(½) - (⅛)] + \{[3 - (9/8)] - [2 - (½)]\}$

$= (⅜) + [(15/8) - (3/2)]$

$= (3 + 15 - 12)/8$

$= 6/8$

$= 3/4$

Example:

Given f(x)=0.048(5-x),
i. Verify that f is a probability density function.
Ii. What is the probability that x is a greater than 4.
iii. What is the probability that x is between 1 and 3 inclusive?