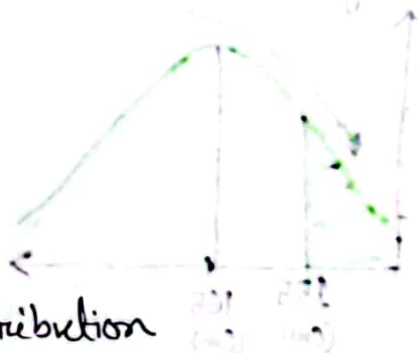
Advanced Statistics 01

Topics:

- Probability Density / Distribution Function
- PDF, PMF and ~~CMF~~ CDF
- Type of probality Distribution
- Bernouli Distribution
- Binomial Distribution
- Poison Distribution
- Normal or Gaussian Distribution
- End



(PdF)

Probability Distribution Fuction/ Density Function:

- Actually it denots the distribution of data

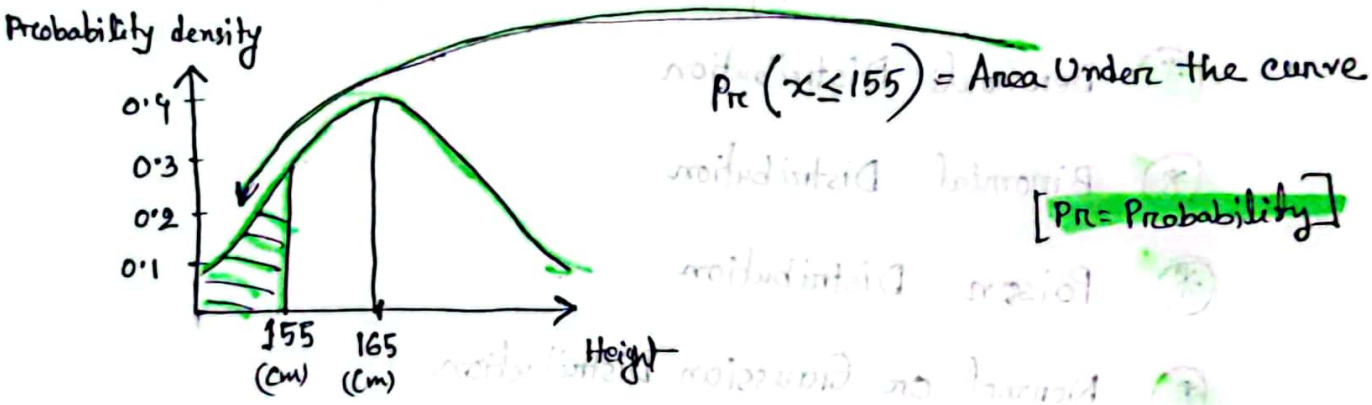
Types

→ ① **Probability Density Function: (PDF)**

→ Continuous Random variable

(Their distribution denotes by PDF)

Example → Height of students in classroom



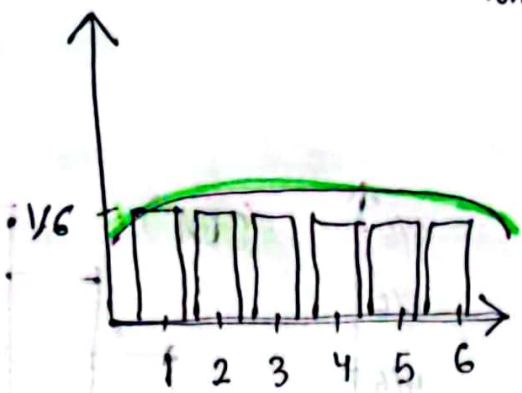
→ ② **Probability Mass Function: (PMF)**

→ When the variable is discrete random variable, their distribution denotes by PMF

Example → Rolling a dice

Value Range = {1,2,3,4,5,6}

Histogram:



What is $\Pr(x \leq 4)$?

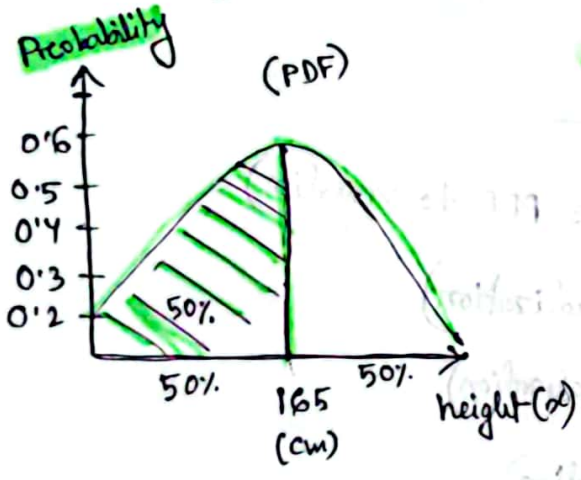
$= \Pr(x=1) + \Pr(x=2) + \Pr(x=3) + \Pr(x=4)$

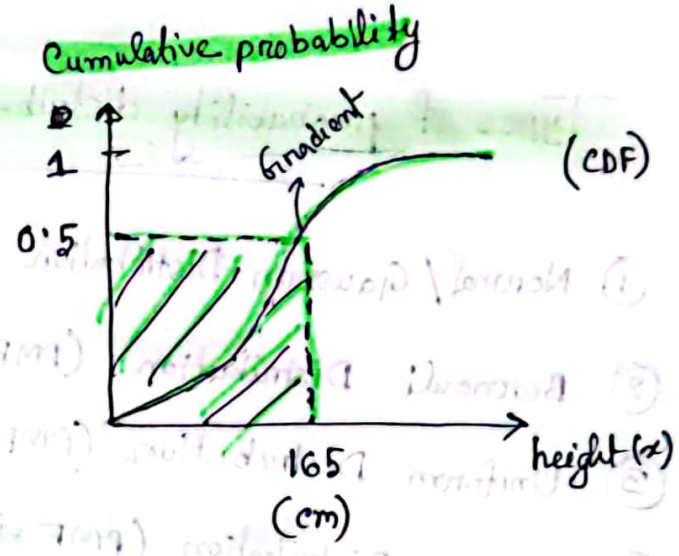
$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$

$= \frac{4}{6} = \frac{2}{3}$

Cumulative Distribution Function:

Again the height example from PDF →





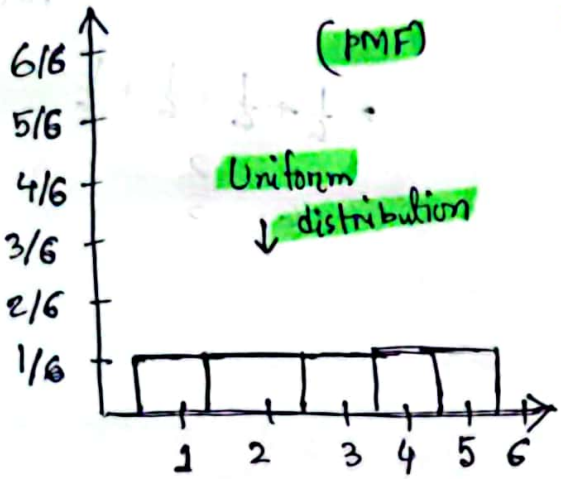
So, using the cumulative Distribution Function, we denote actually how much area height (x) has taken in the first half distribution.

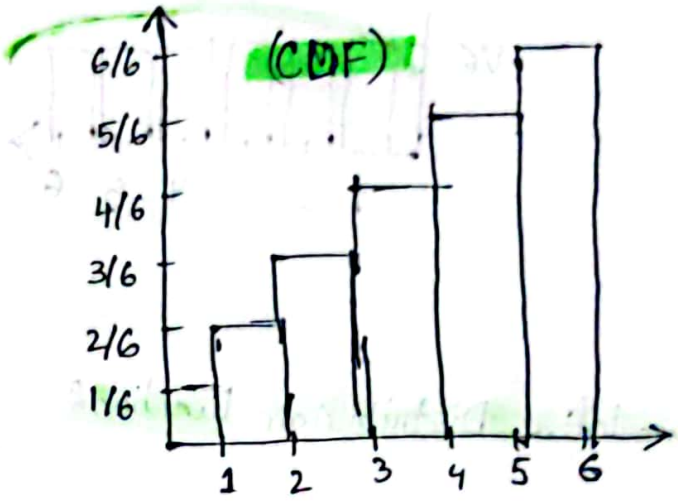
Here in the left chart in the y axis, the values are denoting Gradients (steepness) of the cumulative curve (right) chart)

In the right chart the y axis values are denoting area.

PMF relation with CDF :

Dice rolling probability





Types of probability distribution :

1. Normal / Gaussian Distribution (use PDF to visualize)
2. Bernouli Distribution (PMF visualization)
3. Uniform Distribution (PMF visualization)
4. Poisson Distribution (PMF visualization)
5. Log normal Distribution (PDF visualization)
6. Binomial Distribution (PMF visualization)

Bernoulli Distribution:

Discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1-p$

In a easy way it can be said that,

It is a model for the set of possible outcomes of any single experiment that asks a yes-no question.

Key things to remember in Bernoulli Distribution :-

1. Discrete Random Variable {PMF}
2. Outcomes are binary [(0,1), (Head, tail), (yes, no)]
3. Example - Tossing a coin.

   Pr(Head) = 0.5 (suppose p)

   Pr (Tail) = 1 - 0.5 (1 - p)

4. Example → Pass or fail in exam

   Pr(Pass) = p = 0.7

   Pr (Fail) = q = 1 - p = 1 - 0.7 = 0.3

Parameters:

$0 \leq p \leq 1$

$q = 1-p$

$K = \{0, 1\}$ [yes or no]

① **PMF:** (Probability Mass Function)

$$PMF = p^K * (1-p)^{1-K}$$

Here, $K \in \{0, 1\}$

if $K=1$, $Pr(K=1) = p^1 * (1-p)^{1-1}$

$= p * 1$

$= p$

if, $K=0$ then $Pr(K=0) = p^0 * (1-p)^{1-0}$

$= 1-p$

PMF simplified:

$$PMF \begin{cases} q = 1-p & \text{if, } K=0 \\ p & \text{if, } K=1 \end{cases}$$

Mean of Bernoulli Distribution:

$$E(K) = \sum_{j=0}^{K} K \cdot P(K)$$

if, suppose
$Pr(K=1) = 0.6$
$Pr(K=0) = 0.4$

$= [0 \times p(0) + 1 \times p(1)]$

$= [0 \times 0.4 + 1 \cdot (0.6)]$

$= 0.6$

$= p$

So, p is the mean of Bernoulli Distribution

Median of Bernoulli Distribution:

Median
- 0 if $p < 0.5$
- [0,1] if $p = 0.5$ ← Based on problem Statement
- 1 if $p > 0.5$

Variance:

$$Var = p*(1-p)$$
$$= p*q \quad [q = 1-p]$$

Std:

$$Std = \sqrt{pq}$$

Binomial Distribution: B(n,p)

The binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking yes-no question, and each with it's own boolean valued outcome: Success (p) or failure ($q = 1-p$)

Parameters:

n = number of trials

Suppose I am tossing a coin 7 time

So, trial $n = 7$

p = ~~probability~~ [0, 1] → Success probability for each trial

$q = p - 1$

Key things to note:

⓪ For discrete random variable (PMF Function)

① Every outcome is binary

② This experiment is performed for n trials.
where each trail is a bernouli distribution

③ Every single trial from the n trial is called Bernouli distribution.

Support: $K = \{1, 2, 3, \ldots n\}$ → Number of success for n trials

PMF: $Pr(K, n, p) = {}^{n}C_{K}\, p^{K}(1-p)^{n-K}$ $\left\{ K \in 0 \xrightarrow{to} n \right.$

Mean: np, **Variance:** npq **Std:** $\sqrt{npq}$

Poisson Distribution:

→ For Discrete Random variable (PMF Function)

→ Describes the number of events in a fixed time interval.

Example: Number of people (n) visiting hospital every hour
↳ fixed time

$\lambda = 3$

[λ = Number of people expected in medical in each hour]

Some statistical question that can be achieved:

1) What is the probability of number of people at 5-th hour

2) What is the probability of a particular person to come at 5th hour?

PMF: $P_n(x=5) = \frac{e^{-\lambda}\lambda^x}{x!}$ { $x = 5$th hour, $\lambda = 3$

$= \frac{e^{-3}3^5}{5!}$

$= 10.1\%$

Mean: $\mu = \lambda t$ [λ = Number of expected events occurred at every time interval, t = Time interval]

Variance: $Var = \lambda t$ (same)

Normal or Gaussian Distribution: $N(\mu, \sigma^2)$

→ Continuous probability distribution for a real-valued random variable

→ Mean = Median = Mode

Parameters:

$\mu \in R$ = mean

$\sigma^2 \in R > 0$ = variance

$x \in R$

[R = Real number]

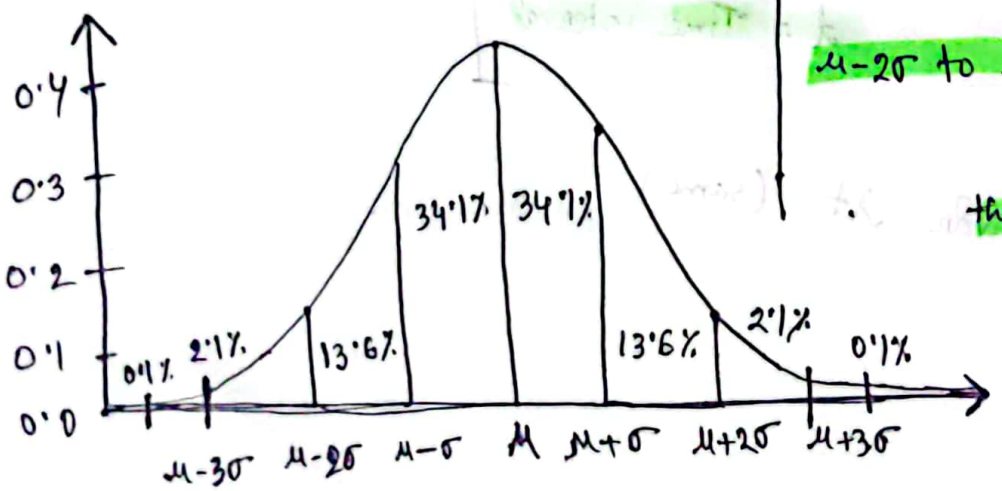
PDF: $\frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$

Mean: μ = Average value

Variance: σ^2

Std: $\sqrt{\sigma^2}$

Empirical Rule of Normal Distribution:



μ-σ to μ+σ → 1st std Region

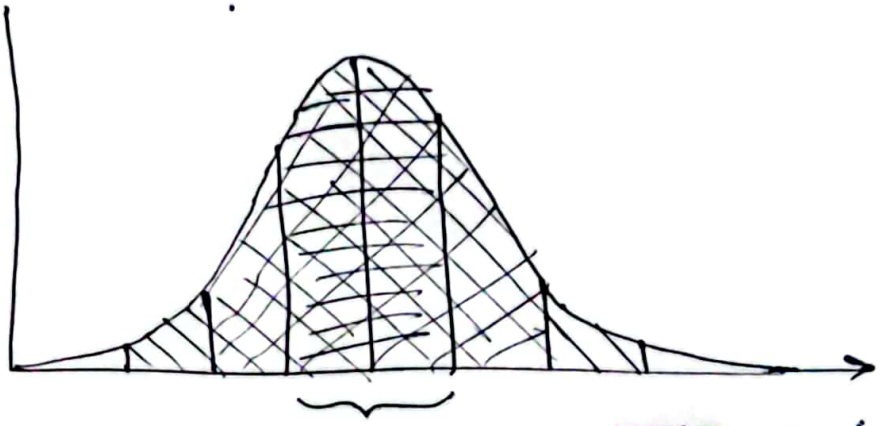
μ-2σ to μ+2σ → 2nd Std Region

then 3rd Std Region

In first std Region → There is 68% data distribution available

In second std Region → There is 95% of data distribution available

In third std Region → There is 99% of data distribution available.



1st Region → (—) This sign area

2nd Region → (/) This sign area

3rd Region → (\) This sign area

In terms of probability

for 1st std Region → $Pr(\mu-\sigma <= x <= \mu+\sigma) \approx 68\%$

2nd std Region → $Pr(\mu-2\sigma <= x <= \mu+2\sigma) \approx 95\%$

3rd std Region → $Pr(\mu-3\sigma <= x <= \mu+3\sigma) \approx 99\%$

Data that follows Normal Distribution:

1) Height, weight
2) Iris