⇒ **Probability Distribution Function**

· **Def:**

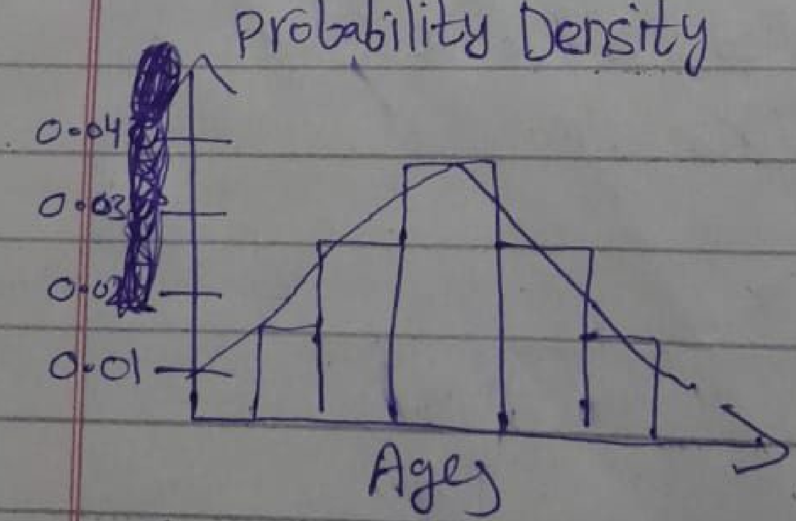
Probability Distribution Function defines/describes how the probabilities are distributed over the values of a random variable.

e.g

Let a continuous random variable

Ages = {. . . .}

We will construct our histogram for that random variable



We smooth the curve for PDF

2 main types of Probability Distribution Function

① Probability Mass Function (PMF): Used for Discrete Random variables

② Probability Density Function (PDF): Used for Continuous Random Variables.

1. Probability Mass Function (PMF)

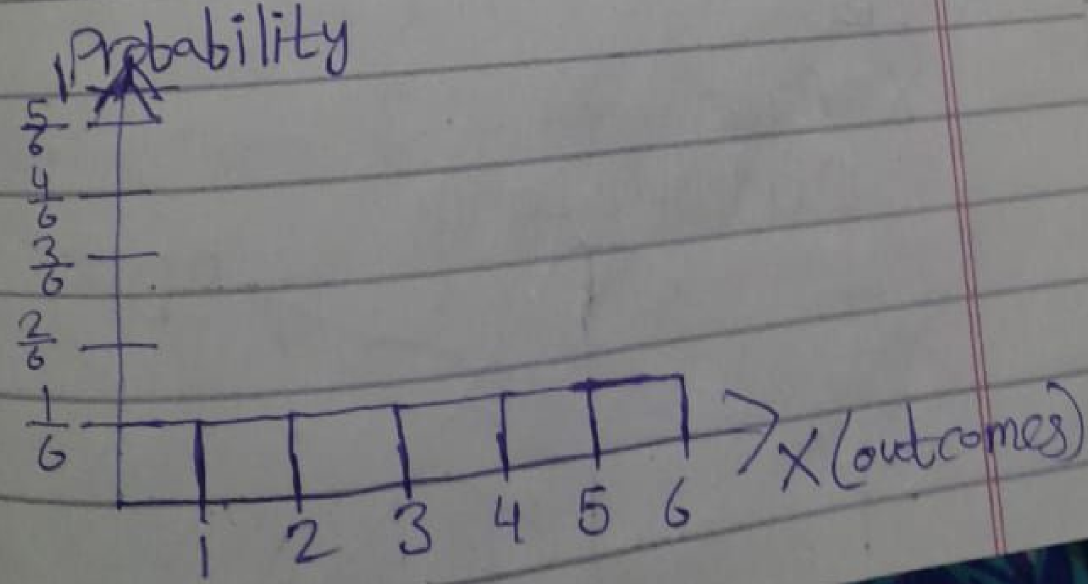
Used for Discrete Random Variables

e.g Rolling a dice {1, 2, 3, 4, 5, 6}

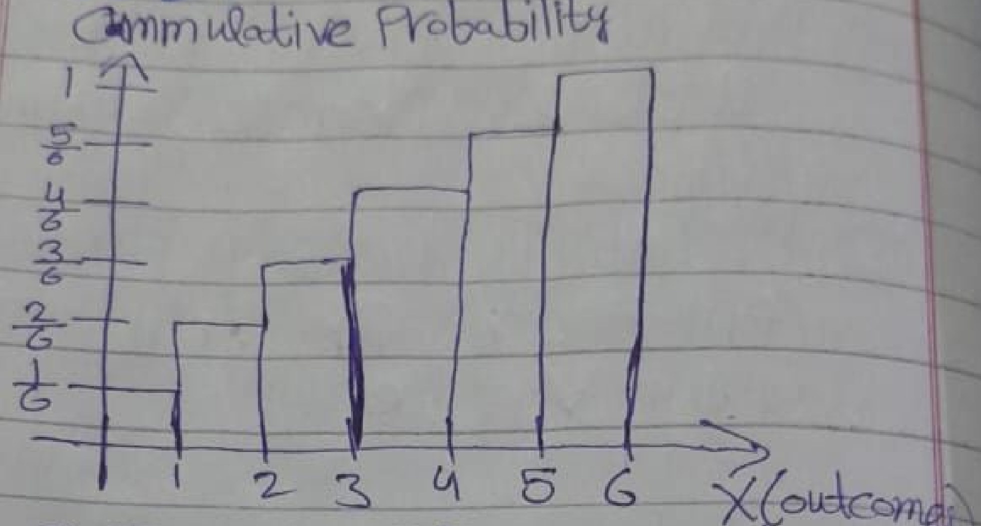
And it is Fair Dice i.e

$P(1) = P(2) = P(3) = P(4) = P(5) = P(6)$

• **Construct PMF**



Cummulative Density Function (CDF)



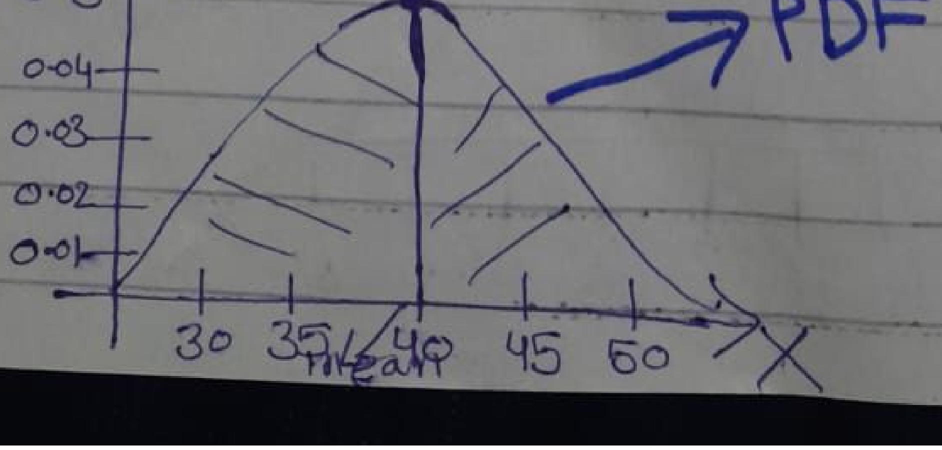
PMF and CDF are used for finding range probabilitie lik $P(x \leq 3) = 0.5 = \frac{3}{6}$

- **Probability Density Function (PDF):**

Used for Distribution of Continuous Random variables

Let

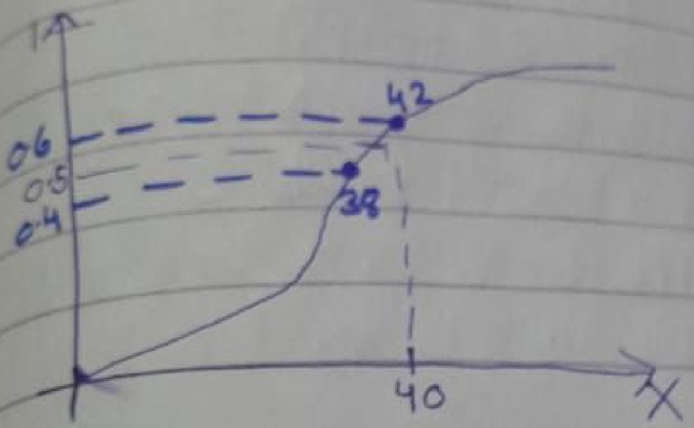
$X = \text{Ages} = \{ \dots \}$



Convert it into Cummulative Density Function

CDF:

Cummulative Probability



So, $P(x \leq 40) = 0.5$

which is **Area under the curve.**

⇒ We calculate the Probibility Density by calculating slope at that points

For calculating slope, pick the two nearest points one below and one above and see their value and commulative probility

$$\text{Slope of } 45 = \left(\frac{0.6-0.4}{42-38}\right)$$

Sog

Probability Density = Gradient of Cummulative Density Function

- **Properties of PDF**

1. It is always Non-Negative
   $f(x) \geq 0$ for all x
2. The total area under the PDF curve is equal to 1.

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

This is entire area under the curve

With respect to different distribution you f(x) function is going to change

Types of Probability Distribution

1) Bernoulli distribution ⇒ Outcomes are binary (We use PMF) as Discrete Random Variables.
2) Bionomial Distribution → PMF
3) Normal / Gaussian Distribution → Most Common One → PDF
4) Poison Distribution → PMF
5) Log Normal Distribution → PDF
6) Uniform Distribution → PMF

1) Bernoulli Distribution

The Bernoulli distribution is the simplest discrete probability distribution. It represents the probability distribution of a random variable that has exactly two possible outcomes: success with probability (P) and failure (1-P). It is used to model binary outcomes.

E.g

i) Tossing a Coin $\{H, T\}$

$Pr(H) = 0.5 = P$

$Pr(T) = 1 - P = 1 - 0.5 = 0.5$

ii) Whether the person will pass or fail

$Pr(Pass) = \boxed{0.6}$

$Pr(Fail) = 1 - Pr(Pass) = 1 - 0.6 = \boxed{0.4}$

⇒ **Parameters:**

$0 \leq P \leq 1$

$q = 1 - P$

$K = \{0, 1\} \Rightarrow 2$ outcomes

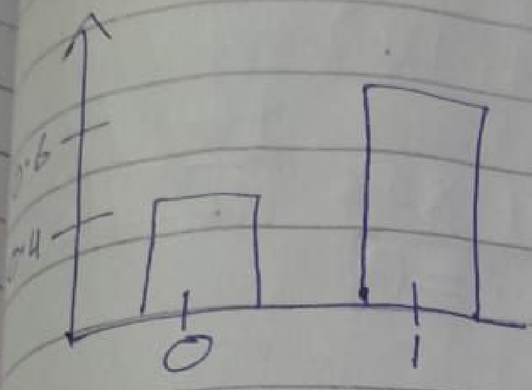
- **Probability Mass Function (PMF):**

A company has launched a new smart phone "A" and I have to provide Review in two scenarios

Use $\Rightarrow (K=1) = 60\% = 0.6 \Rightarrow P$

Not Use $\Rightarrow (K=0) = 40\% = 0.4 \Rightarrow q = 1-P$

PMF



• Formula for PMF [redacted] of Bernoulli Distribution

$$PMF = P^K \times (1-P)^{1-K}$$

if we take $K=1$

$$PMF = P^1 \times (1-P)^{1-1}$$

$$PMF^{(K=1)} = P = 0.6$$

$$Pr(K=0) = P^0 \times (1-P)^{1-0}$$

$$Pr(K=0) = 1-P = q$$

So,

$$\boxed{Pr(K=1) = P}$$

$$\boxed{Pr(K=0) = 1-P}$$

• Simplified:

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

⇒ Mean of Bernoulli Distribution:

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

$\therefore K = \{0, 1\}$

$$E(x) = 0 \times 0.4 + 1 \times 0.6$$

$$\boxed{E(x) = 0.6 = P}$$

⇒ Median of Bernuolli Distribution:

$$\text{Median} = \begin{cases} 0 & \text{if } P < \frac{1}{2} \\ [0, 1] & \text{if } P = \frac{1}{2} \\ 1 & \text{if } P > \frac{1}{2} \end{cases}$$

OR

Median $= 0$ if $q > P$

Median $= 0.5$ if $q = P$

Median $= 1$ if $q < P$

⇒ Mode of Bernoulli Distribution.

if $(P > q)$ then P will be the mode else q will be the mode

Variance of Bernoulli Distribution:

We know:

$K = 0$ and 1, $Pr(K=0) = 0.4 \Rightarrow q$

$Pr(K=1) = 0.6 \Rightarrow P$

$\sigma^2 = 0.4 \times (0-0.6)^2 + 0.6(1-0.6)^2$

$\sigma^2 = 0.4 \times 0.36 + 0.6 \times 0.16$

$\sigma^2 = 0.14 + 0.096$

$\boxed{\sigma^2 = 0.24}$

$\sigma^2(\text{variance}) = Pq = (0.6)(0.4)$

$\boxed{\sigma^2 = 0.24}$

$\Rightarrow$ **Standard Deviation of Bernoulli Distriubution:**

$\sigma = \sqrt{Pq}$

$\sigma = \sqrt{(0.6)(0.4)}$

$\sigma = \sqrt{0.24}$

$\boxed{\sigma = 0.48}$

2. Bionomial Distribution:

Bionomial Distribution with parameters n and p is the discrete probability distribution of the number of success in a sequence of n independent experiments each asking a Yes-No question and each with its own Boolean valued outcome: success (with probability p) or failure (with probability $q = 1 - p$)

⇒ A single success or failure is also called a Bernoulli trial or Bernoulli experiment.

For a single trial $n = 1$, Bionomial distribution will be Bernoulli distribution

⇒ Bionomial distribution is the basis of popular bionomial test of statistical significance

Important points.

1. Discrete random variable (PMF is used)
2. Outcome of every experiment is binary
3. These experiments are performed for n trials

E.g.:

Tossing a Coin 10 times
↓
(H, T)

Notation: $B(n, p)$

Parameters:

$n \in \{0, 1, 2, \dots\} \rightarrow$ No. of trials or experiment

$p \in [0, 1] \rightarrow$ Success probability for each trial

$q = 1 - p$

Support $k \in \{0, 1, 2, 3, \dots, n\}$
→ No. of successes

- PMF:

$Pr(K, n, p) = {}^{n}C_{K} P^{K} (1-P)^{n-K}$

for $K = 0, 1, 2, \cdots n$

$\boxed{{}^{n}C_{K} = \frac{n!}{K!(n-K)!}}$ ⇒ Bionomial Coefficient

- **Mean:** $n \cdot p$
- **Variance:** $n \cdot p \cdot q$
- **Standard Deviation:** $\sqrt{n \cdot p \cdot q}$
- **Example:**

No. of trial$(n) = 5$

Probability of success$(P) = 0.5$

No. of success (K) = Varies from 0 to 5

Question: What is the probability of getting exactly 3 heads in 5 trials

$n = 5, K = 3$

$Pr(x=3) = {}^{5}C_{3} (0.5)^{3} (1-0.5)^{5-3}$

$\therefore P(K, n, p) = {}^{n}C_{K} (P)^{K} (1-P)^{n-K}$

$\boxed{Pr(x=3) = 0.31}$

⇒ Example : Quality Control
Scenario. Inspecting the 10 items in a factory where each item has a 10% chance of being defective.

Data:

No. of trials $(n) = 10$

Probability of success $(P) = 0.1$

No. of successes ~~(x)~~ varies from 1 to 10

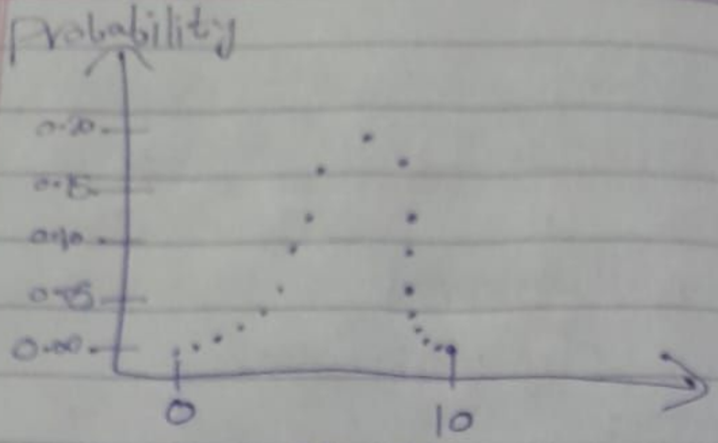
Question : What is the probability of finding exactly two defective items in a sample of 10

$$n = 10, K = 2, P = 0.1$$

$$P(x=2) = {}^{10}C_2 (0.1)^2 (1-0.1)^{10-2}$$

$$\boxed{P(x=2) = 0.193}$$

★ For Probability mass function graph e.g $p = 0.5$ $n = 20$
So, K will be 0 to 20
So we will find Probability for every K and mark point



Something like this

3 Poisson Distribution:

The poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occuring in a fixed interval of time if these events occur with a known constant mean rate and independently of the time since last event.

• Important points:

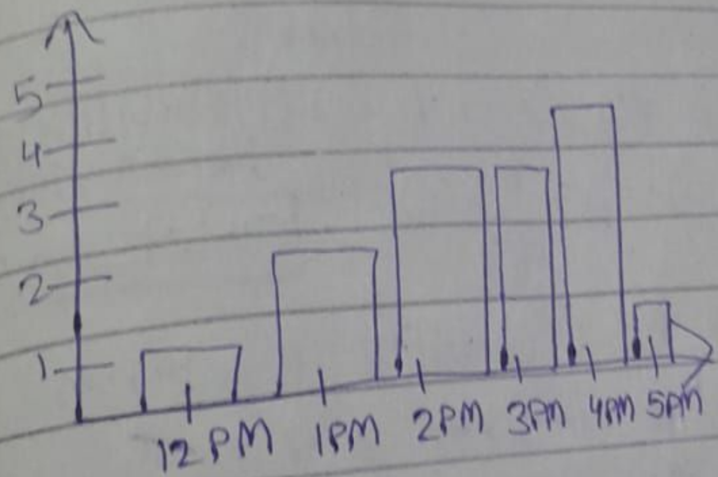
1. Discrete Random Variable (PMF)
2. Describe number of events occaring in a fixed interval

of time

• Examples:

Number of people visiting bank every hour

Let's make a histogram



Parameters:

$\lambda \Rightarrow$ Expected no. of events occuring at every time interval

PMF

$$P(X) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

Let $\lambda = 3 \Rightarrow$ Expected number of people visiting at every time interval

$X = 5$ (5th hour) **(Probability of person visiting at 5 PM)**

$$P(5) = \frac{e^{-3}(-3)^{5}}{5!} = \boxed{0.101 = 10\%}$$

Similarly probability of a person visiting before 3PM

$P(x<3) = P(12) + P(1) + P(2)$

Mean of poison distribution:

Mean $= E(x) = \mu = \lambda \times t$

where t is time interval

$\lambda =$ Expected no. of events at every interval

Variance $= \lambda$

Standard Deviation $= \sqrt{\lambda}$

Normal / Gaussian Distribution

: Normal or Gaussian Distribution is a type of continuous probability distribution for a real valued random variable

- **Important points:**

Continuous random variable
$\Rightarrow$ PDF

Bell curved

Mean = Median = Mode

- if X is a random variable following normal distribution then Bell curve will be **symmetric distribution** i.e 50% of values in random variable are present to left of mean and 50% to right of mean.

Notation: $N(\mu, \sigma^2)$

Parameters:

$\mu \in R \Rightarrow$ mean

$\sigma^2 \in R > 0 \Rightarrow$ Variance

$x \in R$

$\downarrow$

data in Random Variable

$\Rightarrow$ As variance increases the spread of data also increases

Example:

Weights of student in a class

Heights of student in a class

- **PDF:**

$$\boxed{PDF = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}}$$

- **Mean:**

$$\mu = \sum_{i=1}^{N} \frac{X_i}{N}$$

- **Variance:**

$$\sigma^2 = \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{N}$$

- **Standard deviation:**

$$\sigma = \sqrt{\text{Variance}}$$

- Empirical Rule (68–95–99.7 Rule)

Standard Deviation Region

1st ⇒ $Pr(\mu - \sigma \leq x \leq \mu + \sigma) \approx 68\%$.

2nd ⇒ $Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 95\%$.

3rd ⇒ $Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 99.7\%$.

If a Random variable X follows Normal/Gaussian Distribution.

5) Standard Normal Distribution:

Let a Random variable

$X = \{1, 2, 3, 4, 5\}$

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

$$\mu = \frac{1+2+3+4+5}{5}$$

$$\boxed{\mu = 3}$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

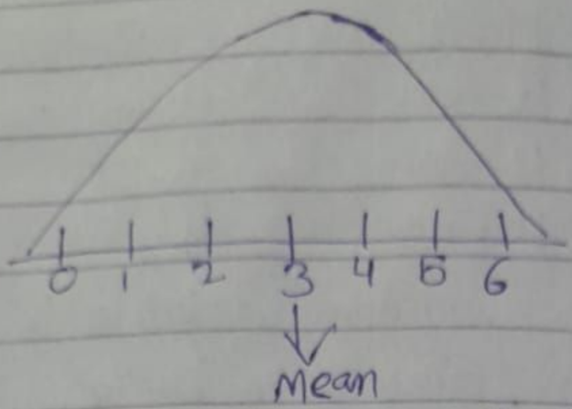
$$\sigma^2 = 2$$

$$\sigma = \sqrt{\sigma^2}$$

$$\boxed{\sigma \text{ (Standard Deviation)} = 1.414}$$

Let it

$\sigma \approx 1$



Ad as $\sigma = 1$ then 1 step

Let's say I have a gaussian distribution and we convert it into mean $\mu = 0$ and $\sigma = 1$ then this will become standard Normal Distribution

⇒ For this conversion we use Z-Score

- **Z-Score** $= \dfrac{x_i - \mu}{\sigma}$

We apply it to all values of random variable

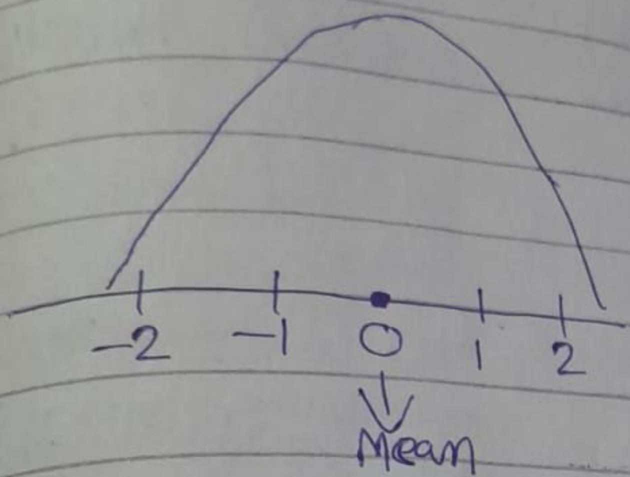
$X = \{1, 2, 3, 4, 5\}$

① $\dfrac{1-3}{1} = -2$ ③ $\dfrac{3-3}{1} = 0$

② $\dfrac{2-3}{1} = -1$ ④ $\dfrac{4-3}{1} = 1$

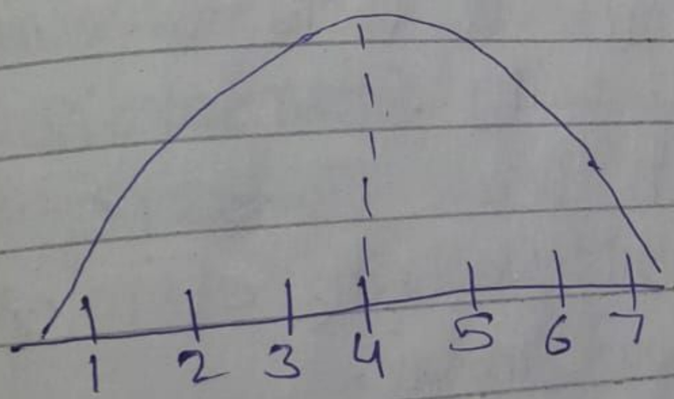
⑤ $\dfrac{5-3}{1} = 2$

$Z = \{-2, -1, 0, 1, 2\}$



and Standard Deviation = 1

• Example:



$\mu = 4 \qquad \sigma = 1$

Q: How many Standard Deviation is 4.25 away from mean?

• Sol: We will use Z-score

$X_i = 4.25$

$$Z = \frac{4.25 - 4}{1} = \frac{X_i - \mu}{\sigma}$$

it is positive so to right

• Example:

Dataset

Age (years)	Weight (kg)	Height (cms)	Salary
24	70	175	40k
25	60	160	50k
26	55	160	60k
22	40	130	30k
30	30	175	20k
31	25	180	70k

Before training model we convert all field to same unit using **Standardization** which uses **Z-score**

⇒ By this we bring efficiency in our model.

5) Uniform Distribution

(i) Continuous Uniform Distribution (PDF)

Continuous uniform distributions or rectangular distributions are family of symmetric probability distributions. Such distribution describes an experiment where there is an arbitrary outcome that lies b/w certain bounds. The bounds are defined by parameters a and b which are maximum and minimum values.

- **Notation:** $U(a,b)$

Params:

$-\infty < a < b < \infty$

PDF (Probability density function)

$$PDF = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if otherwise} \end{cases}$$

CDF (Commulative distribution Function)

$$CDF = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } x \in [a,b] \\ 1 & \text{if } x > b \end{cases}$$

- **Mean:**

Mean $= \frac{1}{2}(a+b)$

- **Median:**

Median $= \frac{1}{2}(a+b)$

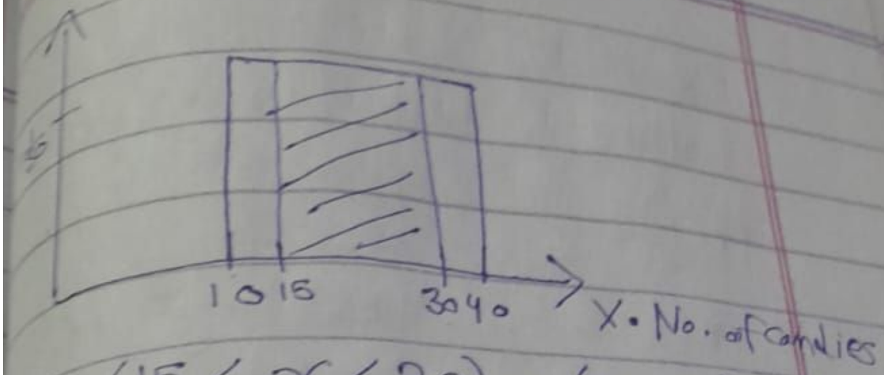
- **Variance:**

Variance $= \frac{1}{12}(b-a)^2$

Example:

The number of candies sold daily at a shop is uniformly distributed with a maximum of 40 candies and minimum of 10

(i) Probability of daily sales to fall between 15 and 30



$$Pr(15 \le x \le 30) = (x_2 - x_1) \times \frac{1}{b-a}$$

$\therefore x_1 = 15, \; x_2 = 30$

~~$Pr(15 \le x \le 30) = (30-15) \times \frac{1}{30} = \frac{15}{30}$~~

$$\boxed{Pr(15 \le x \le 30) = 0.5}$$

(ii) Discrete Uniform Distribution:

Discrete Uniform Distribution is a symmetric distribution wherein a finite number of values are equally likely to be observed; every one of n values has equal probability $\frac{1}{n}$.

- ⇒ Discrete Random Variable
- ⇒ PMF

Example:

Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

$Pr(1) = \frac{1}{6}$, $Pr(2) = \frac{1}{6}$, $Pr(3) = \frac{1}{6}$

$Pr(4) = \frac{1}{6}$, $Pr(5) = \frac{1}{6}$ $Pr(6) = \frac{1}{6}$

$a = 1$ $b = 6$ (Values ranging between 1 to 6)

$\boxed{\text{Probability} = \frac{1}{n}}$

Where $\boxed{n = b - a + 1} = 6 - 1 + 1$

$\boxed{n = 6}$

Notation: $U(a, b)$

Parameters: a, b where $b \geq a$

PMF: $\frac{1}{n}$

Mean: $\frac{1}{2}(a+b)$

Median: $\frac{1}{2}(a+b)$

① Log Normal Distribution:

Log Normal Distribution is a Continuous probability Distribution of a random variable whose algorithm is normally distributed. Thus if X is random variable and it is normally distributed in terms of algorithm then $Y = \ln(x)$ has normal distribution. Equivalently if Y has normal distribution then exponential function of Y, $\cancel{X} = \exp(Y)$ has log normal distribution.

- **Important points:**

$X \approx$ Log Normal Distribution (μ, σ)

- Log Normal Distribution is a Right Skewed Distribution.

Let

$X \approx$ log Normal Distribution (μ, σ)

$\Downarrow$

$y \approx \ln(X) \Rightarrow$ Normal Distribution

$\rightarrow$ Natural Log ($\log_e$)

Now let $y \approx$ Normal Distribution

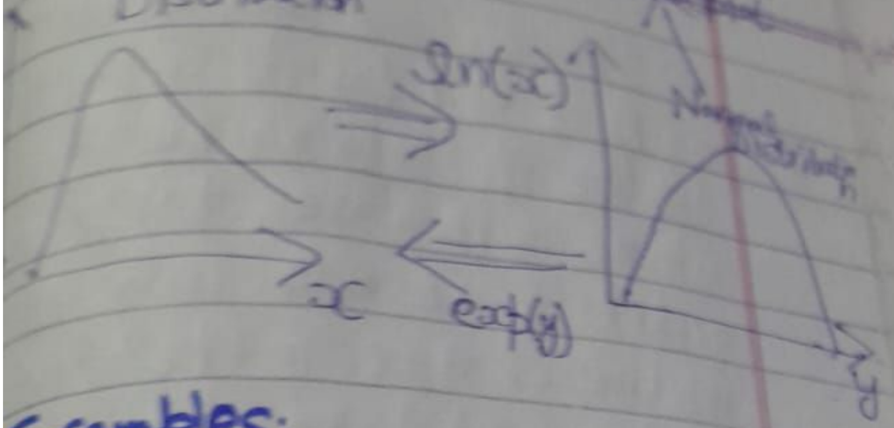
$\Downarrow$

$X = \exp(Y) \Rightarrow$ Log Normal Distribution

Mean if X is log Normal distribution and you apply ln on it then result will be Normal Distribution and if Y is Normal Distribution and we apply exp on it then result will be log Normal distribution

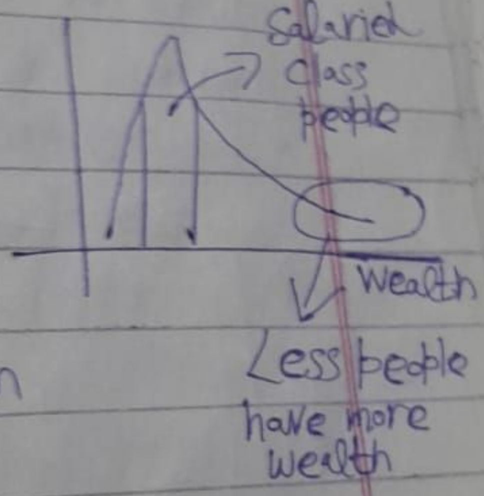
Log Normal Distribution

Checked by



Examples:

1) Wealth Distribution of the World



Discussion Forum
→ length of Comments

Length of Chess games

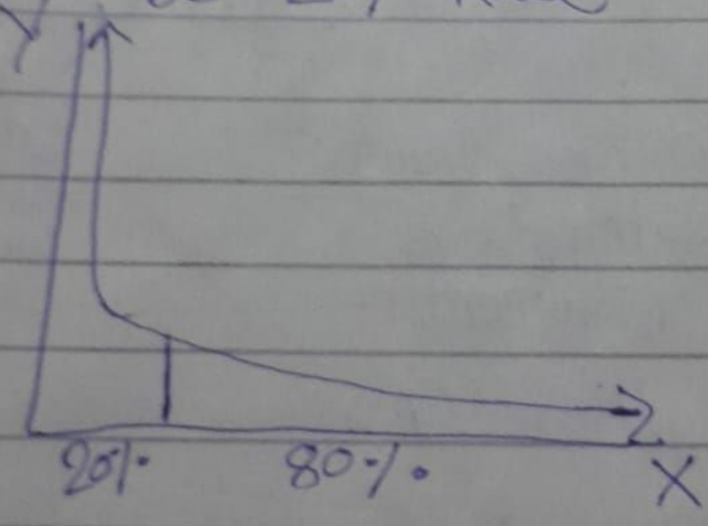
Spend time on online articles

8) Power Law Distribution:

A power law is a functional relationship between two quantities where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities, one quantity varies as power of other.

- **Important Point:**

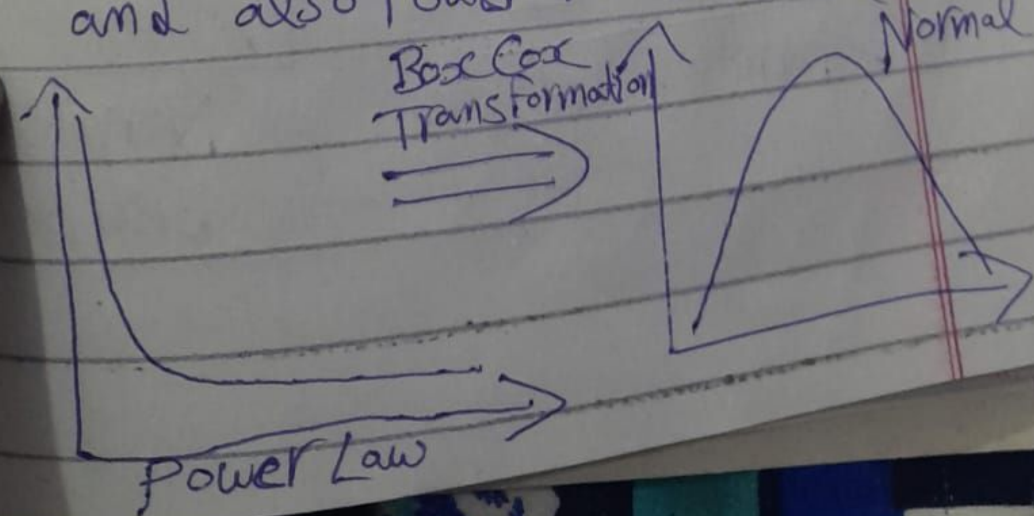
80–20% Rule



Example:

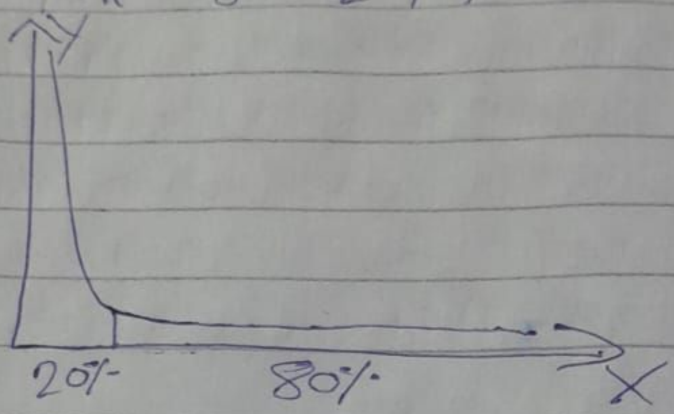
1. PSL (20% of Team is responsible for winning 80% match)
2. 80% of wealth are distributed with 20% of the total population
3. 80% of the total oil is with 20% of the nature
4. Frequencies of words in most languages
5. 20% of major defects fixes 80% of upcoming defects in a software product

★ There are a lot of machine learning algorithms that want data in Normal Distribution that's why we convert log Normal distribution to Normal and also Power law to Normal



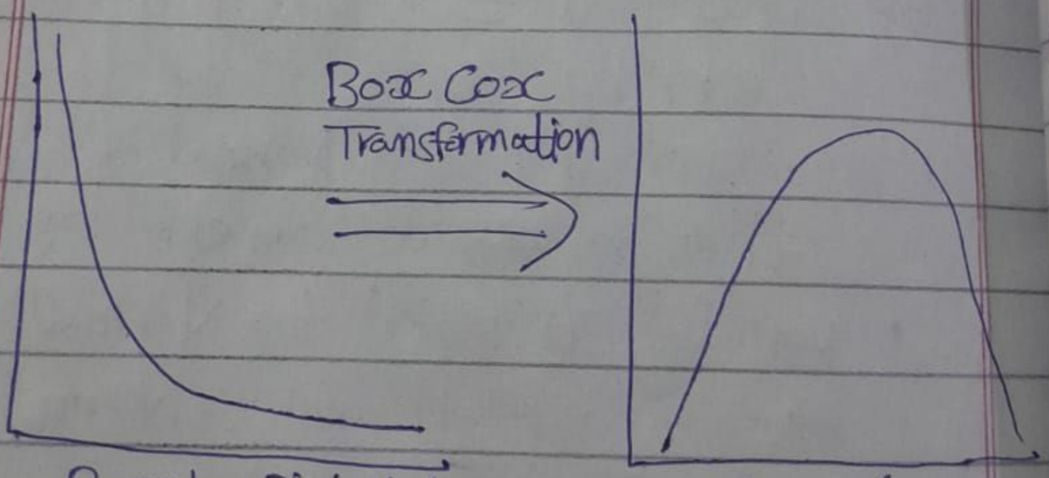
9) Pareto Distribution
(Non-Gausian Distribution)

Follow 80-20% Rule



Pareto Distribution has α

As α increases height of PDF also increases



Parreto Distribution

Normal Distribution

Examples.

① 80% of the entire project is done by 20% of the team.

② 80% of defects can be solved if we solve 20% of major defects.

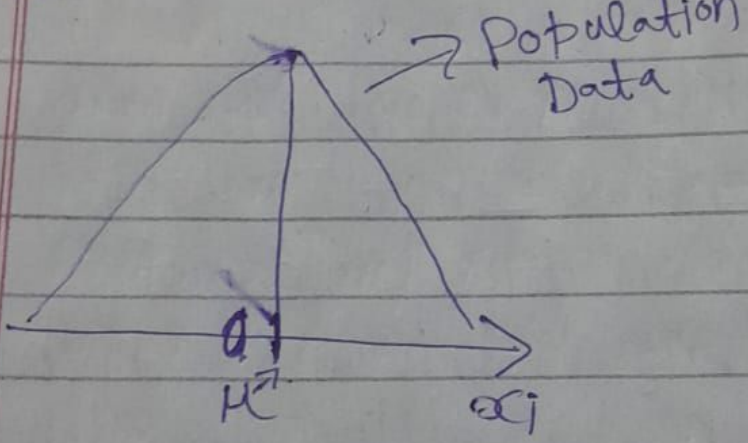
→ Pareto Distribution always follow power law distribution

⇒ Central Limit Theorem

Central Limit Theorem relies on the concept of sampling distribution which is the probability distribution of statistics for a large no. of samples taken from a population.

Central Limit theorem says that the sampling distribution of mean will always be normally distributed as long as sample size is large enough. Regardless of whether the population has normal, poisson or any other distribution, the sampling distribution of mean will be normal.

① $X \in N(\mu, \sigma) \Rightarrow$ Normal Distribution



$n \Rightarrow$ Sample size any value

$S_1 = \{x_1, x_2, x_4, \dots, x_n\} = \bar{x}_1$

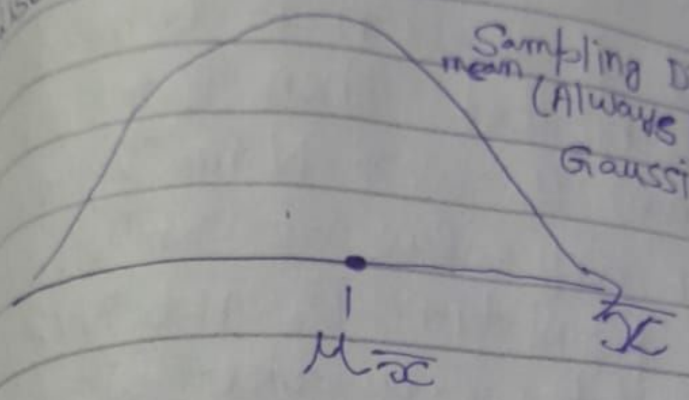
$S_2 = \{x_2, x_3, \dots x_n\} = \bar{x}_2$

S_3 $\qquad\qquad\qquad \bar{x}_3$

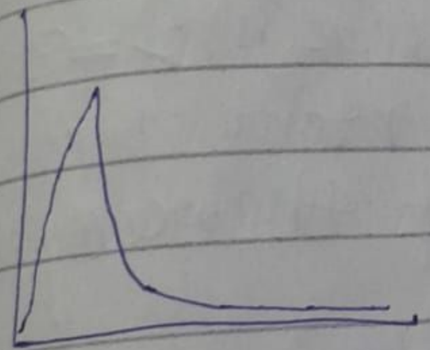
$\vdots$

S_m $\qquad\qquad\qquad \bar{x}_m$

Now we will draw the sampling distribution of mean



$X \neq N(\mu, \sigma)$ Not Normal Distribution



Now $n \geq 30 \Rightarrow$ Sample size
Now pick sample take mean and plot it the sample distribution of mean will be normal distribution

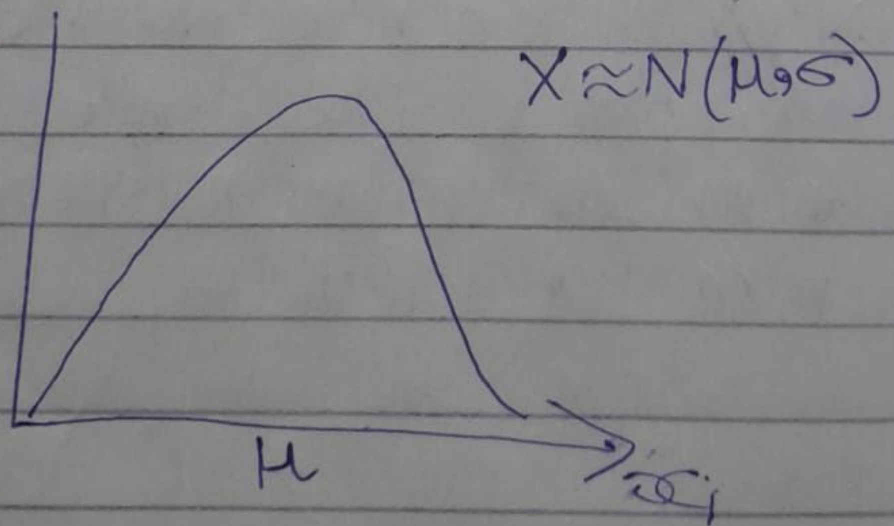
• Summary:

1. If population is Normally distributed then Sample Distribution of Mean is also Normally distributed for any value of sample size (n)

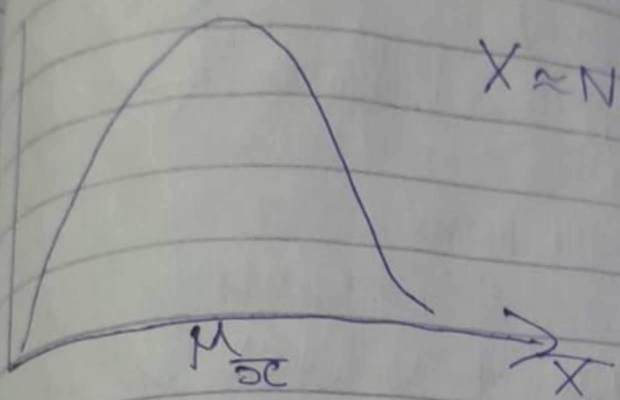
2. If population is not Normally distributed (any other distribution) then sample distribution of mean is also Normal Distribution if sample size(n) >= 30

Another important point Central Limit theorem says that if

Population Data



Sampling Mean Distribution:



Population Mean

Population standard devn

$$X \approx N\left(\mu, \frac{\sigma}{n}\right)$$

Sample size

⇒ Estimates:

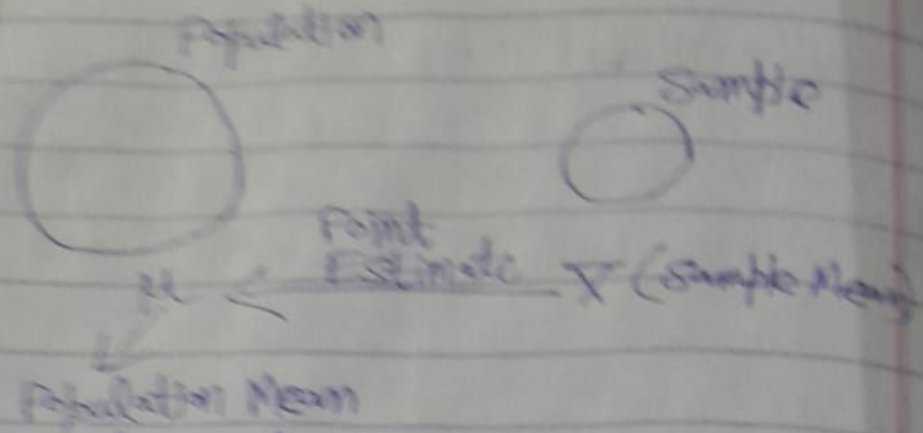
★ **Estimate** is a specified observed numerical value used to estimate an unknown population parameter

. Types of Estimate

① Point Estimate:

Single numerical value used to estimate unknown population parameter

e.g Sample mean is a point estimate of population mean



② **Interval Estimate:**

Range of values is used to estimate unknown population parameters