Probability Distribution Functions for Data:-

probability distribution functions describe how the probabilities are distributed over the values of a random variable.

Ages = { - - - - - } ⇒ Continuous random variable.

It has two type of function:

(i) Probability mass function.
(ii) Probability density function.

(i) Probability mass function: (PMF): used for discreate random variable.

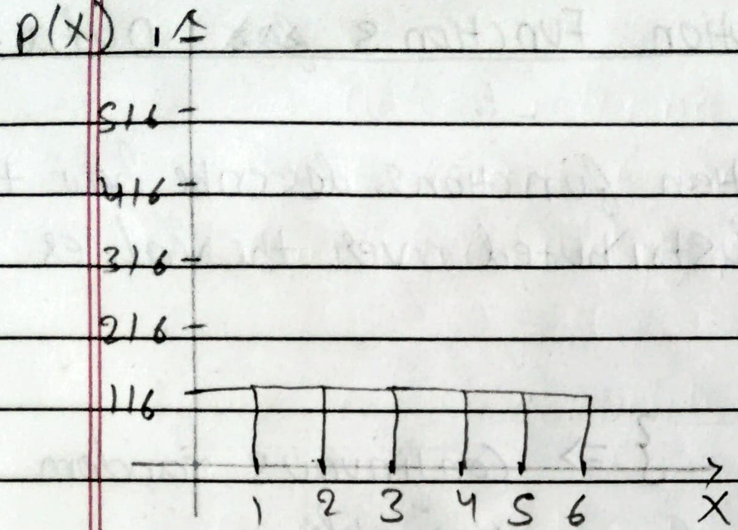
eg: Rolling a dice : { 1, 2, 3, 4, 5, 6 }

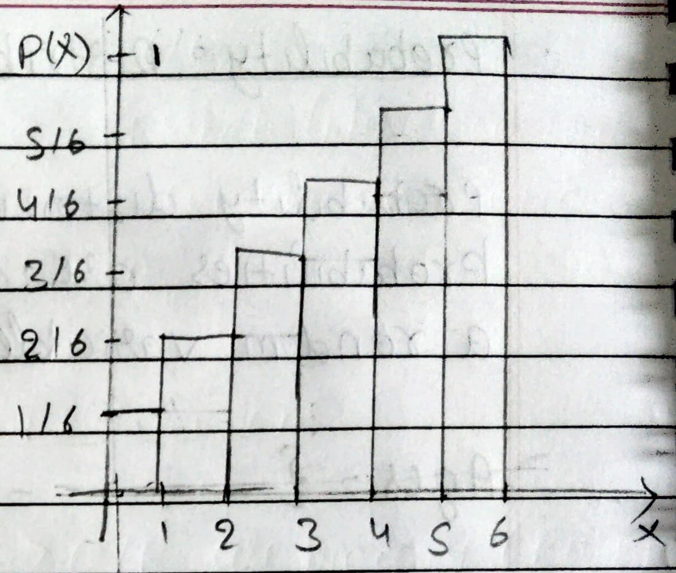
$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$

⇒ Cumulative density function (CDF):

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

$P(4) = 4/6$, $P(5) = 5/6$, $P(6) = 6/6$





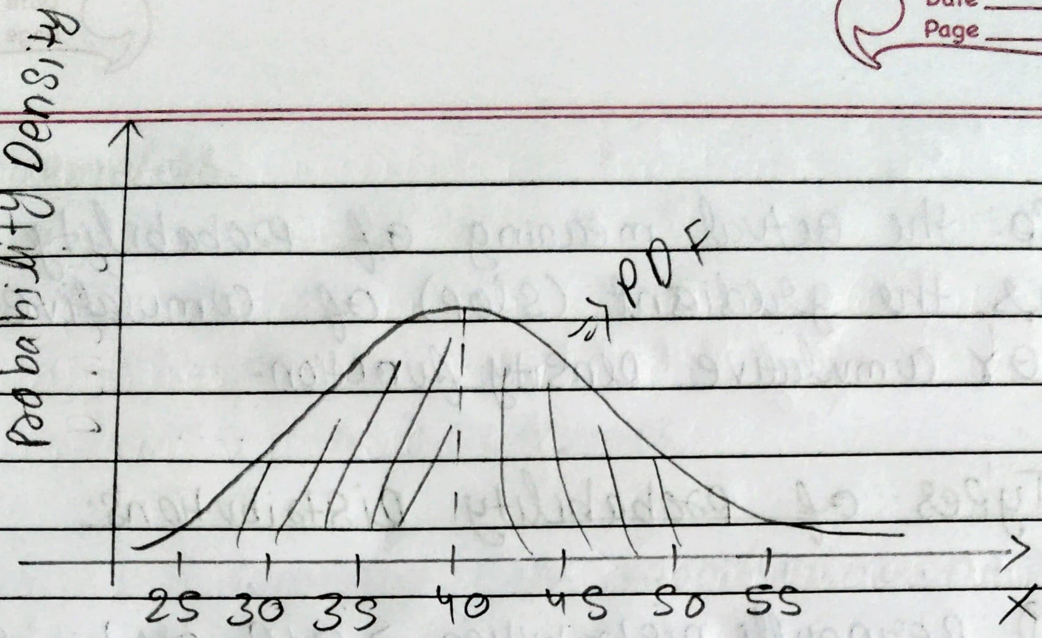
$$
\begin{aligned}
P(X \le 2) &= P(1) + P(2) \\
&= 1/6 + 1/6 \\
&= 2/6 \\
&= 1/3
\end{aligned}
$$

$$
\left\{ \begin{array}{l} P(x) \geq 0 \text{ for every } X \\ \sum_{i=1}^{n} P(x_i) = 1 \end{array} \right\}
$$

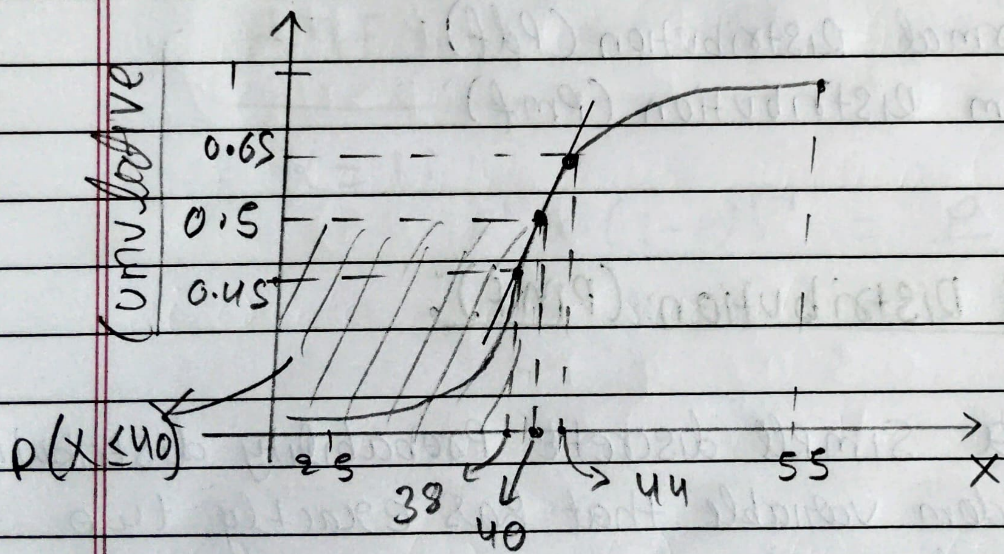
$$
\begin{aligned}
P(X \le 6) &= P(X=1) + P(X=2) + \cdots P(X=6) \\
&= \frac{1}{6} + \cdots \frac{1}{6} \\
&= 1
\end{aligned}
$$

② **Probability density function(PDF):**

~~distid~~ used for continious random variables or distribution of continious random variable.



Cumulative probability



So to finding the $P(X \leq 40)$ youhave to canlculate the area under curve.

$$\boxed{slop = \frac{y_2 - y_1}{x_2 - x_1}}$$

So the actual meaning of probability density is the grediant (slop) of cumulative function or cumulative density function.

Types of probability Distributions:

1. Bernoulli Distribution → O/P are binary (pmf)
2. Binomial Distribution (pmf)
3. Normal / Gaussion Distribution (pdf) → Assumptions
4. Poisson Distribution (pmf)
5. Log Normal Distribution (pdf)
6. Uniform Distribution (pmf)

① Bernoulli Distribution: (PMF):

It is a simple discrete probability distribution of a random variable that has exactly two possible outcomes : ~~with~~ success (with probability p) and the failure (with probability 1-p)

and it is used to model binary outcomes such as coin flip or a yes/No questions.

Parameters:

P = success
$q = (1-P)$ = failure
$K = \{1 \text{ or } 0\} \Rightarrow$ 2 outcomes.

PMF : A company has launched a new phone 'A

use = 60% = P
not use = 40% $\Rightarrow (q = 1-P)$

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

if $K = 1$:
$$P^1 \times (1-P)^{1-1} = \underline{P}$$

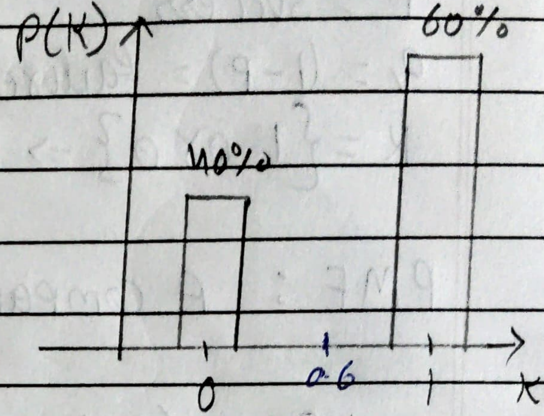
if $K = 0$:
$$P^0 \times (1-P)^1 = (1-P) = \underline{q}$$

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

mean of the Bernoulli distribution!

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

(mean) Expectation of X



median: - $\begin{cases} 0, & \text{if } P \le 1/2 \\ [0,1], & \text{if } P = 1/2 \\ 1 & \text{if } P > 1/2 \end{cases}$

$$\begin{cases} \text{median} = 0 & \text{if } q > P \\ \text{median} = 0.5 & \text{if } q = P \\ \text{median} = 1 & \text{if } q < P \end{cases}$$

mode: $P > q \Rightarrow$ P will be the mode

- else q will be the mode.

Variance: $E[x^2] - E[x]^2$
$\Rightarrow p\cancel{*}q$
$\Rightarrow p(\cancel{1}-p)$

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

Standard deviation $(\sigma) = \sqrt{pq}$

(11) Binomial Distribution (PMF)

- Discrete random variable
- Every outcome of the experiment is binary
- These experiments are performed for n trials

Parameters:

$n \in \{0, 1, 2, 3, \ldots\} \Rightarrow$ no of trials.
$p \in [0, 1] \Rightarrow$ success probability of each trial
$q = 1-p$
$K = \{0, 1, 2, \ldots n\} \Rightarrow$ Number of successes.

$B(n, p) \Rightarrow$ Binomial.

PMF $\Rightarrow$

$$P_r(K, n, p) = ({}^nC_K)(p^K(1-p)^{n-K})$$

classmate

Date

Page

$$\boxed{{}^{n}C_{k} = \frac{n!}{k!\,(n-k)!}}$$ ⇒ Binomial coefficient

mean $= n \cdot p =$ median

variance $= npq$

Standard deviation $= \sqrt{npq}$

Q: What is the probability of getting exactly 3 heads in 5 flips?

Soln $n = 5,\ k = 3$ $\quad p = 0.5 \Rightarrow$ (Head)

$$P(X=3) = {}^{5}C_{3}\,(0.5)^3\,(1-0.5)^{5-3}$$

$$= \underline{0.3125}$$

③ Poisson Distribution:

- Discrete random variable (PMF)
- Discribes the numbers of events occuring in a fixed time intervals.

eg:- No of people visiting hospital every hour.

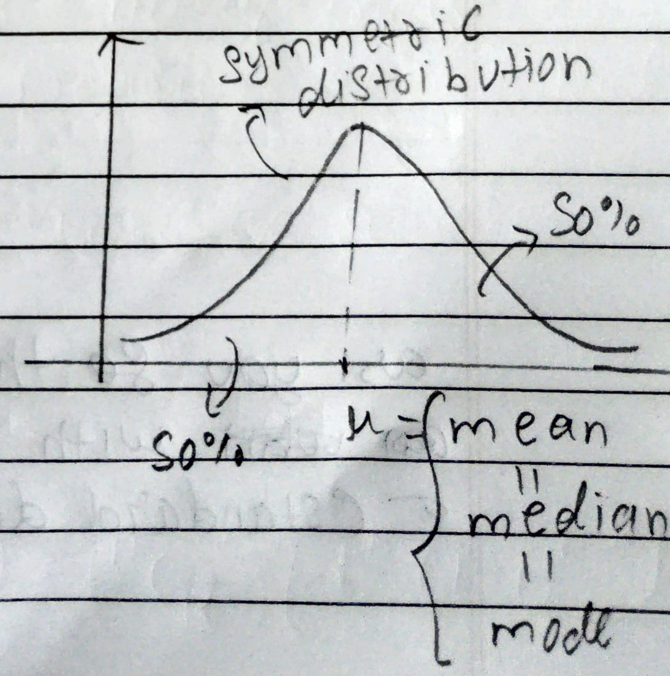
$$\boxed{PMF = \frac{\lambda^{x} \cdot e^{-\lambda}}{x!}}$$

λ is the expected rate of occurrences in the given time.

$\{mean = E(X) = \mu = \lambda * t\}$ t = time interval

④ Normal / Gaussian Distribution (pdf)

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



Parameter:

$\mu \in R$ = mean

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

e to the power

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

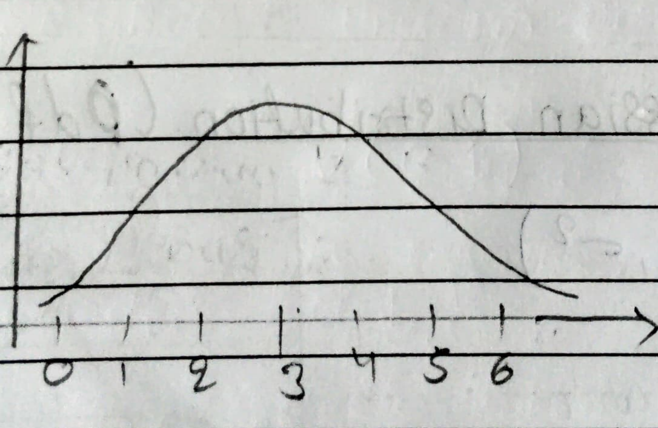
$$mean = \sum_{i=1}^{n} \frac{x_i}{n} = media = mode$$

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

⑤ Standard Normal Distribution and Z-Scor

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

$\sigma = 1.414 \approx 1$



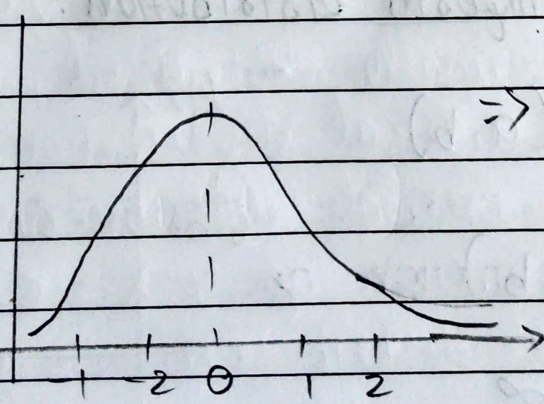
as you so the graph of distribution ~~...~~ with the value of some μ and σ (standard deviation), If we transform the distribution curve in other distributio where the mean (μ) = 0 and standard deviation (σ) = 1, to achive this we have used the Z-score

$$\boxed{\text{Z-Score} = \frac{x_i - \mu}{\sigma}}$$

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

$$y = \left(\frac{x_i - \mu}{\sigma}\right)$$

$y = \{-2, -1, 0, 1, 2\}$ ⇒ x after applying Z-score.



⇒ Standard normal Distribution.

⇒ Why we use the Z-score?

When you are going to build any model (machine learning), you use the dataset and the ~~d~~ ~~b~~ unit of features are different and the values are not in same range they are scattered so using Z-Score we are standardize all the features to bring in same range for better model working or learning. we apply Z-Score on all the features.

⑥ **Uniform Distribution:**

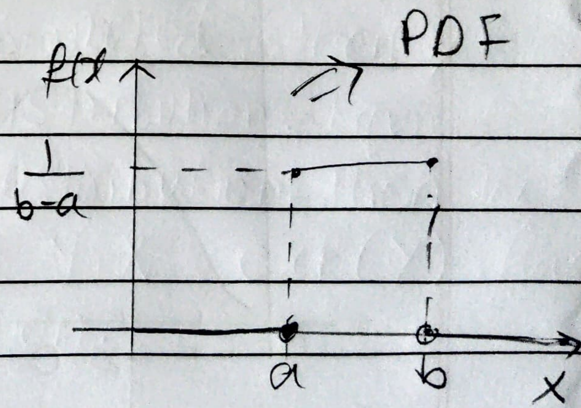
(i) Continious Distribution &
(ii) Discrete Distribution.

① **Continuous Uniform Distribution:**

Notation: $U(a, b)$

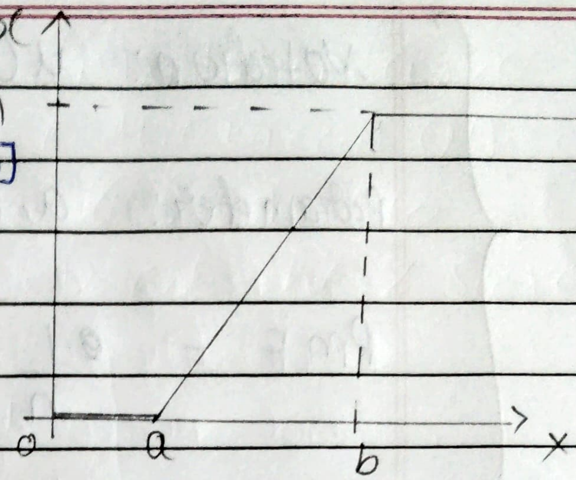
Parameters: (a, b)

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



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

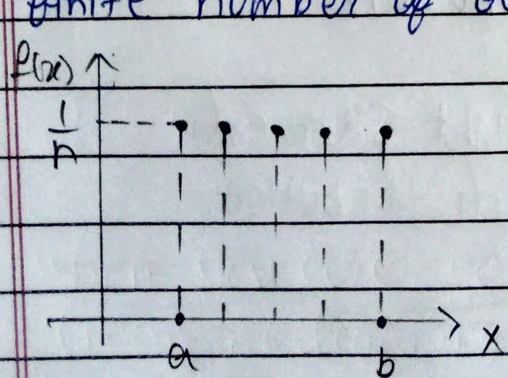
$$CDF = \begin{cases} 0 & \text{for } x < a \\ x-a/b-a & \text{for } x \in [a,b] \\ 1 & \text{for } x > b \end{cases}$$



$$\left\{ \begin{array}{l} \text{mean} = \frac{1}{2}(a+b) = \text{median} \\ \text{variance} = \frac{1}{12}(b-a)^2 \end{array} \right\}$$

② <u>Discrete Uniform Distribution:</u>

It is a symmetric probability distribution wherein a finite number of values are equally likely to be observed: everyone of n values has equal probability 1/n. Another way of saying "discrete uniform distribution" would be 'a known finite number of outcomes equally likely to happen'.



$$\begin{cases} \text{pmf} = \frac{1}{n} \\ n = b-a+1 \end{cases}$$

Notation: $U(a,b)$

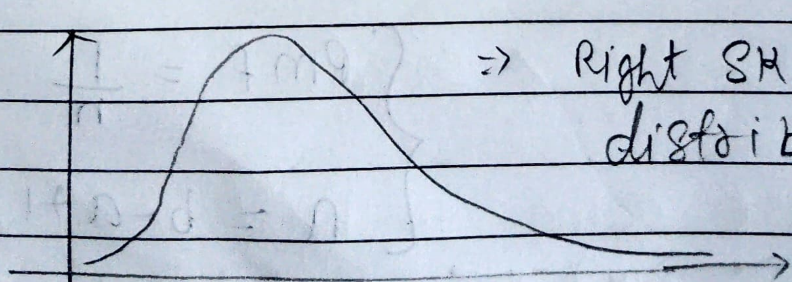
Parameter: a, b, where $a \le b$

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

$$mean = \frac{a+b}{2} = median$$

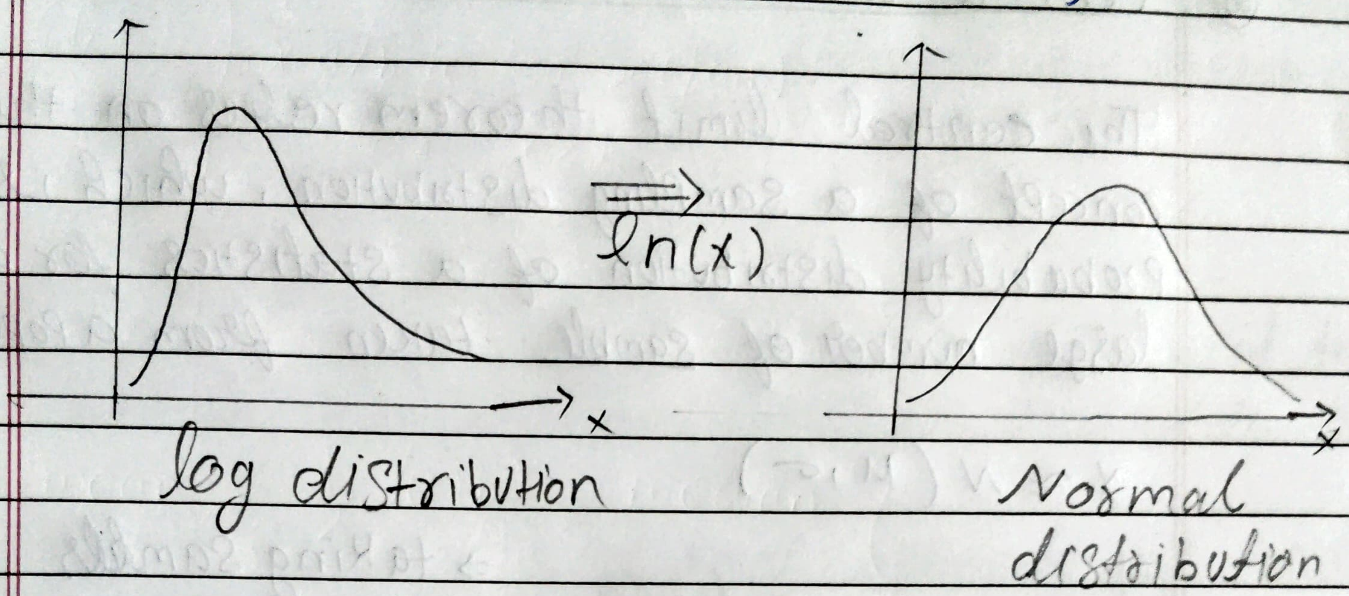
(7) Log Normal distribution:

In Probability theory, a log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then $Y = \ln(X)$ has a normal distribution. Equivalently, if Y has a normal distribution, then the exponential function of Y, $X = \exp(Y)$, has a log-normal distribution.



⇒ Right skewed distribution.

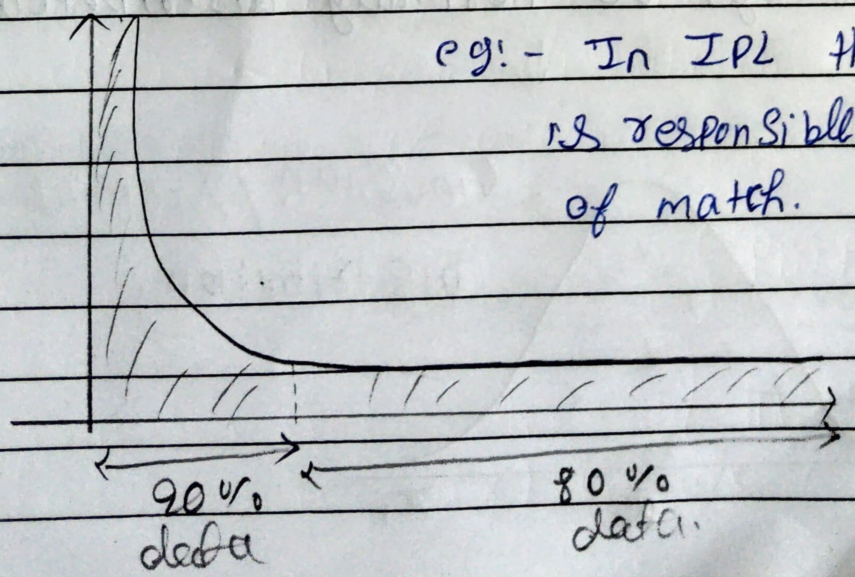
X $\quad\quad\quad\quad$ $Y = \ln(X)$



⑧ **Power law distribution:**

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

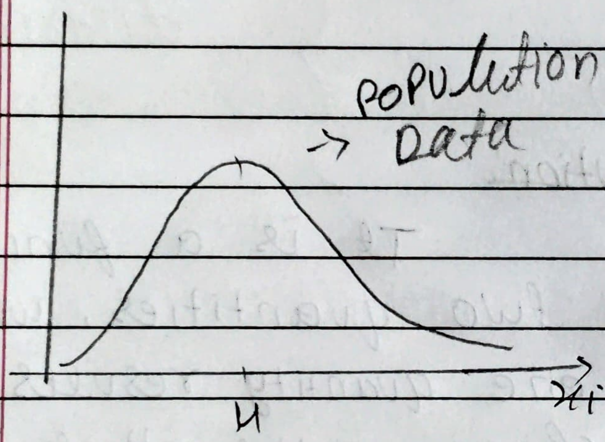
eg:- In IPL the 20 % of team is responsible for winning 80% of match.



Central limit theorem.

The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistics for a large number of sample taken from a population.

$X \approx N(\mu, \sigma)$



⇒ taking samples

n = sample size ⇒ any value

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

$S_2 = \{ \quad \} = \bar{x}_2$

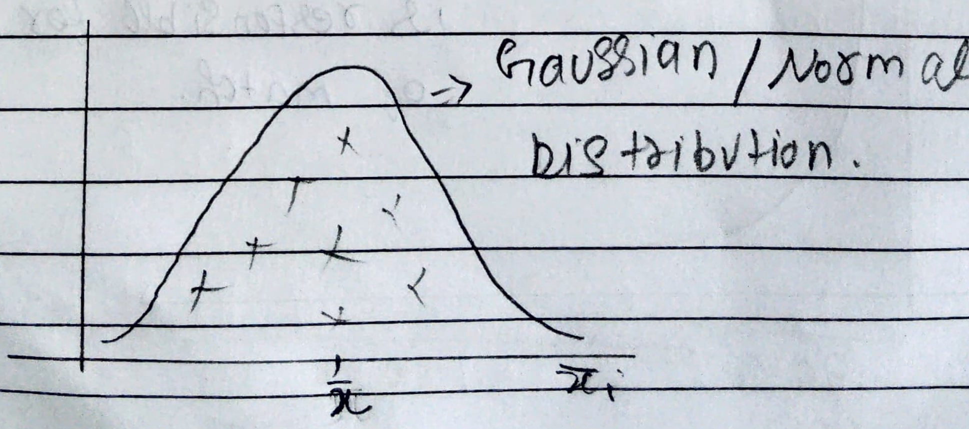
$\vdots$

$S_m \{ \quad \} = \bar{x}_m$

⇓

It says that sampling distribution of mean will ~~be~~ always be normally distributed.

⇓

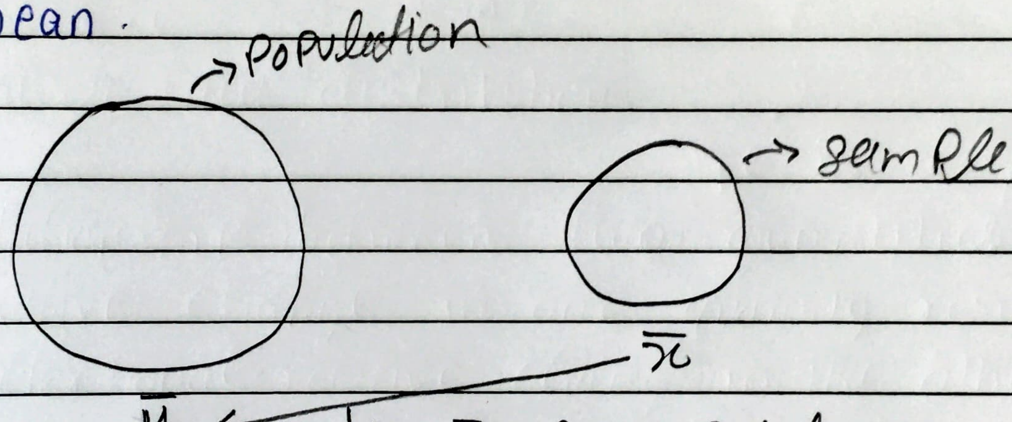


Estimate: It is specified observed numerical value used to estimate an unknown population parameter.

① **Point Estimate:**

Single value (numerical) used to estimate an unknown population parameter.

eg: Sample mean is a point estimate of population mean.



↳ $\bar{x}$ is a point estimate of $\bar{\mu}$

② **Enterval Estimate:**

Range of values is used to estimate unknown population parameter.