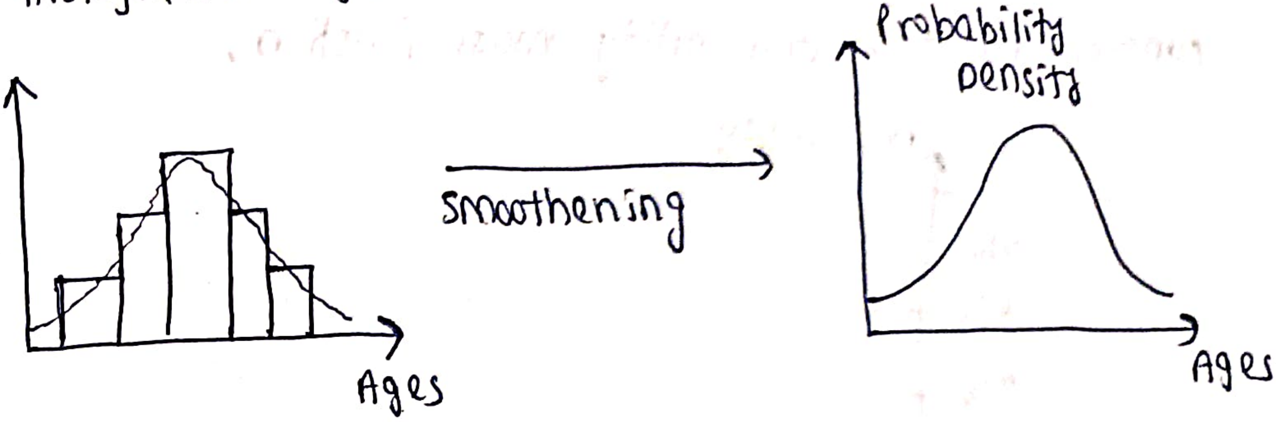
PROBABILITY DISTRIBUTION FUNCTION

⇒ Probability distribution functions describe how the probabilities are distributed over the values of random variable.

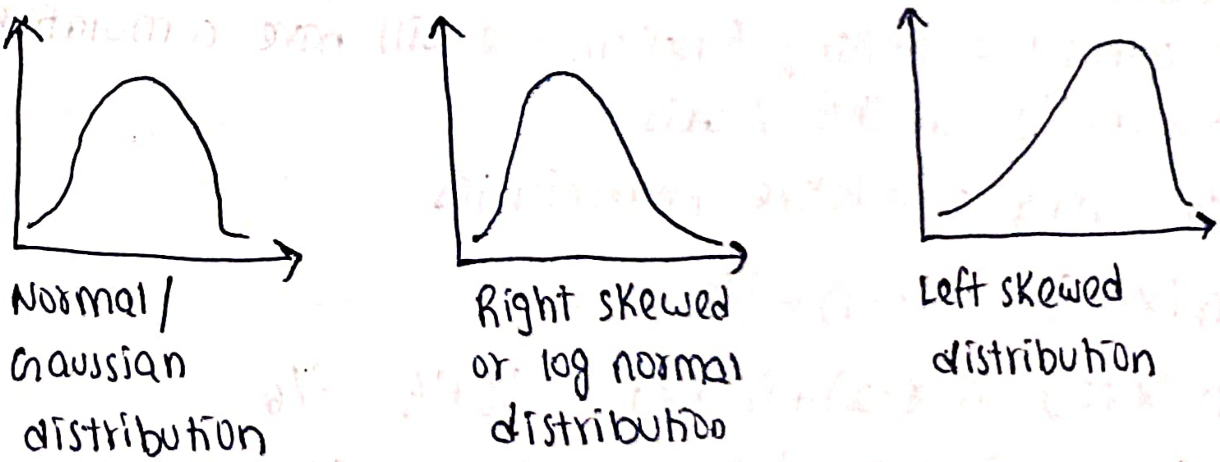
Example:- Ages → continuous random variable

Histogram of Ages



After smoothing the histogram, we will have probability density values at the Y-axis.

Different probability distributions provide different kinds of basic information about data.



⇒ Probability distribution functions are of two types:-

(i) Probability mass Function (PMF), which is used for discrete random variables.

(ii) Probability Density Function (PDF), which is used for continuous random variables.

Probability Mass Function (PMF):-

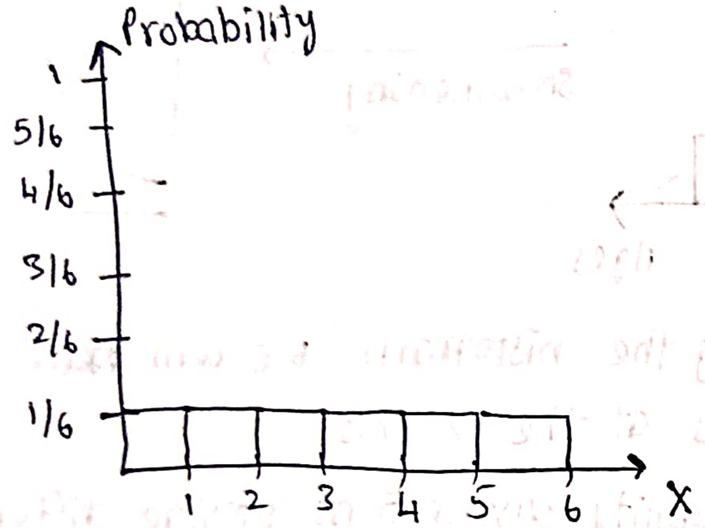
It is specifically used for discrete random variables.

Example:- Rolling a fair dice

Possible outcomes = {1,2,3,4,5,6}

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

constructing the probability mass function,



Cumulative density function (CDF):-

In cumulative density function, we will have cumulative probability on the Y-axis.

Calculating cumulative probabilities,

$Pr(X \leq 1) = Pr(X=1) = 1/6$

$Pr(X \leq 2) = Pr(X=2) + Pr(X \leq 1) = 1/6 + 1/6 = 2/6$

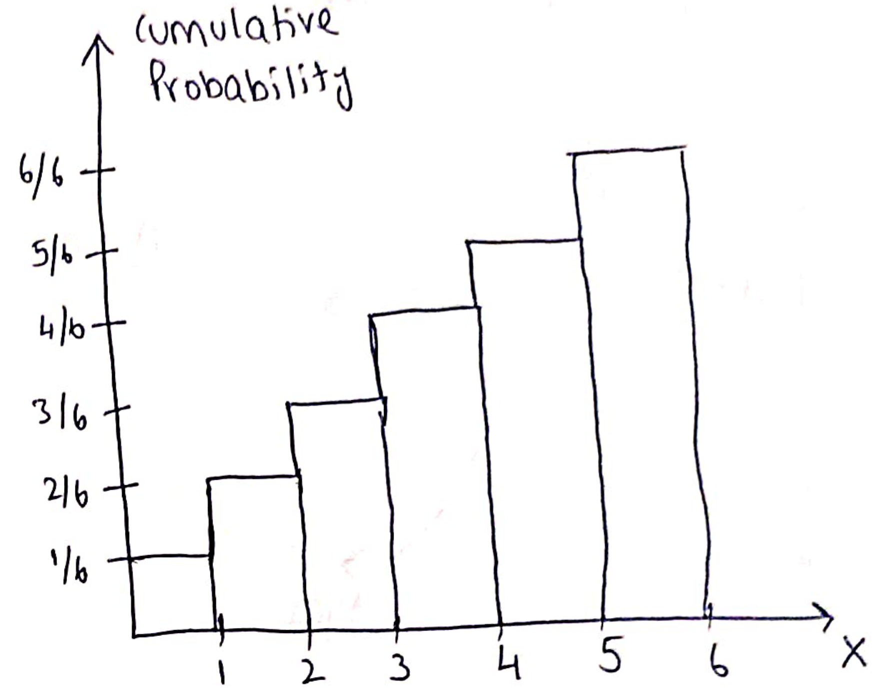
$Pr(X \leq 3) = Pr(X=3) + Pr(X \leq 2) = 1/6 + 2/6 = 3/6$

$Pr(X \leq 4) = Pr(X=4) + Pr(X \leq 3) = 1/6 + 3/6 = 4/6$

$Pr(X \leq 5) = Pr(X=5) + Pr(X \leq 4) = 1/6 + 4/6 = 5/6$

$Pr(X \leq 6) = Pr(X=6) + Pr(X \leq 5) = 1/6 + 5/6 = 6/6$

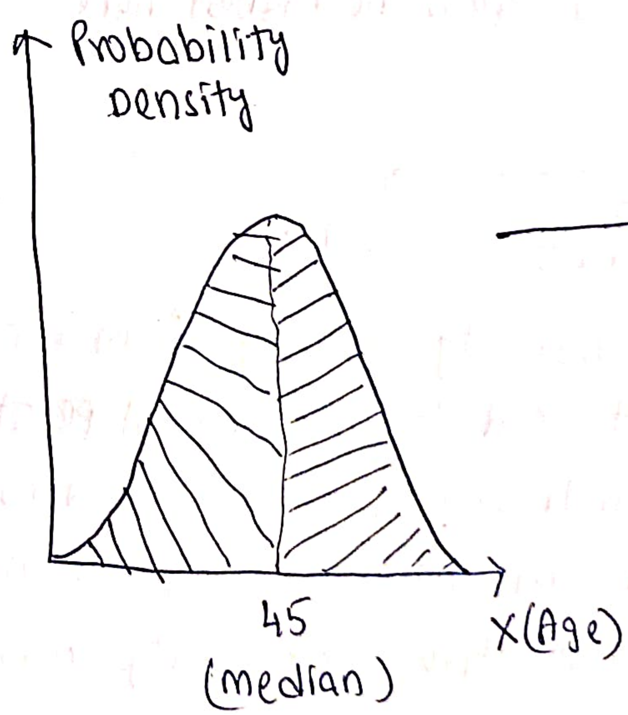
Constructing the cumulative density function (CDF).

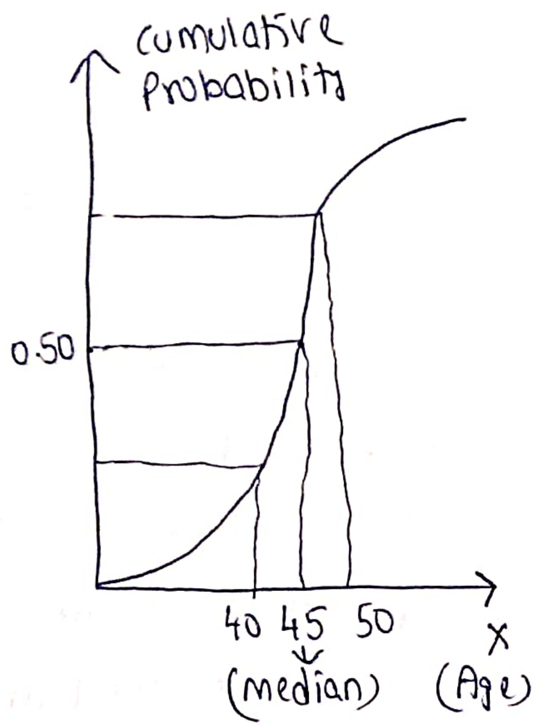


* Probability Density Function (PDF):-

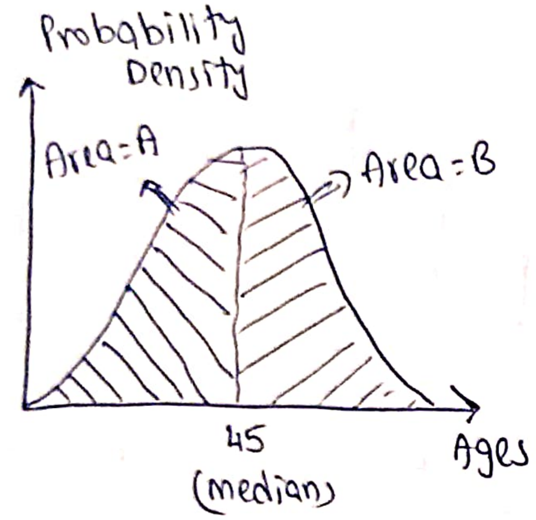
PDF is specifically used for continuous random variables.

Example:- Age dataset which has symmetric distribution and a median of 45



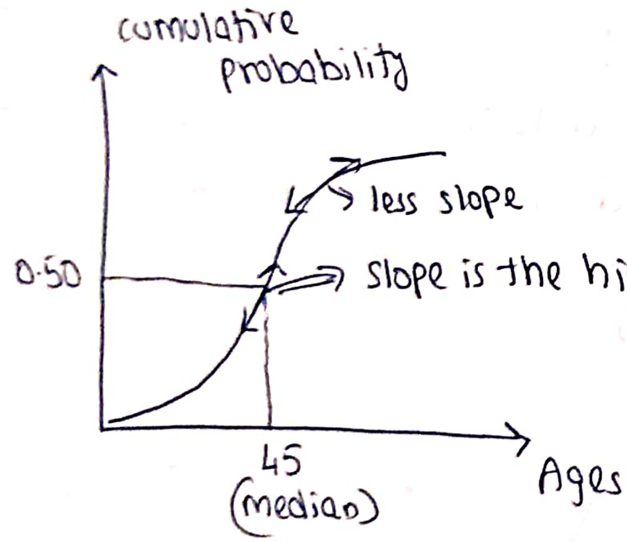


In a symmetric distribution, half of the probability density lies to the left of median and other half lies to the right of median.



A = B

For data that follows symmetric / gaussian probability density curve, the slope will be maximum at the median



To findout the value of probability density from cumulative probability curve, we just need to find the slope at that point. So, for symmetric distribution, probability density is highest at the median and therefore slope is the highest at median in cumulative probability curve.

Probability Density is the gradient of cumulative density function

Properties of probability density function:-

→ Its always non-negative i.e

$$f(x) \geq 0 \text{ for all } x$$

→ The area under the PDF curve is 1

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

with respect to different distributions, f(x) will change but the area under the curve will always be 1.

(A) Types of probability distribution:-

(i) Bernoulli distribution (outcomes are binary) (Discrete) (PMF)

(ii) Binomial distribution (Discrete) (PMF)

(iii) Normal / Gaussian distribution (PDF)

(iv) Poisson distribution (PMF)

(v) Log-normal distribution (PDF)

(vi) uniform distribution (PMF)

→ why is understanding probability distribution important?

It is because, as a data scientist when we receive a dataset, different features may have different type of random variable.

Example:- House price dataset

Size of house	no. of rooms	Floor	sea side	Price
⇓	⇓	⇓	⇓	⇓
continuous random variable (PDF)	Discrete random variable (PMF)	Discrete random variable (PMF)	Binary (0/1)	continuous random variable (PDF)

By analysing different kinds of distributions, we get alot of information from the data and we can make some assumptions about data that can help us in EDA (Exploratory Data Analysis) and FE (Feature Engineering).

BERNOULLI DISTRIBUTION

→ Properties of Bernoulli distribution:-

(i) used for discrete random variables
Bernoulli distribution is the simplest discrete probability distribution.

(ii) used for random variables with exactly two possible outcomes.
Probability of success $= p$
Probability of failure $= 1-p = q$

Since it is used for discrete random variables, we use probability mass function to obtain probability distribution.

→ Examples of when it can be used:-

(i) Tossing a coin
$Pr(X = \text{Head}) = 0.50 = p$
$Pr(X = \text{Tail}) = 0.50 = q = 1-p$

(ii) whether a student passes or fails in an exam
$Pr(X = \text{pass}) = p$
$Pr(X = \text{fail}) = q = 1-p$

In both these examples, random variables have exactly two possible outcomes and are discrete. So, we can use bernoulli distribution in both the cases.

→ Parameters:-

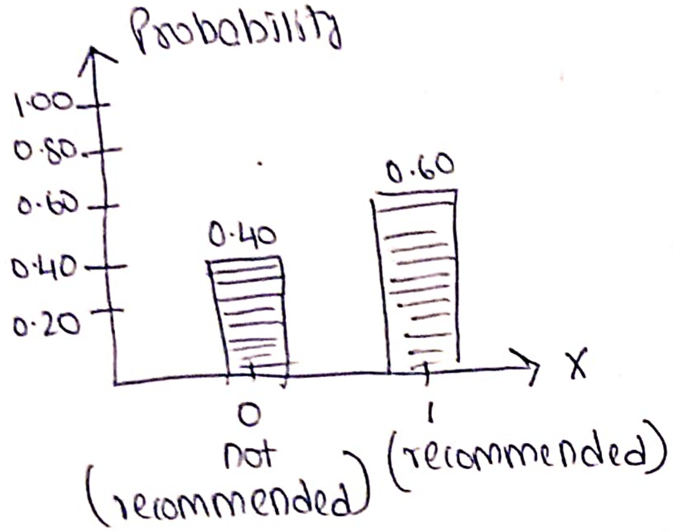
(i) Success probability $= p \in [0,1]$
(ii) Failure probability $= q = 1-p \in [0,1]$
(iii) Possible outcomes $= \{0,1\} = K$

Probability curve/Distribution :-

Let us consider that a company has launched a new smartphone.

$Pr(X = \text{recommended}) = 0.60 = P$

$Pr(X = \text{not recommended}) = 1 - 0.60 = 0.40 = q$



The Probability Mass Function (PMF) used to obtain this distribution is

$$\boxed{PMF = p^K * (1-p)^{1-K}} \text{ , where } K = \{0,1\}$$

In the example,

$P = 0.60$

$Pr(K=1) = p^K (1-p)^{1-K} = p^1 (1-p)^{1-1} = p^1 = 0.60$

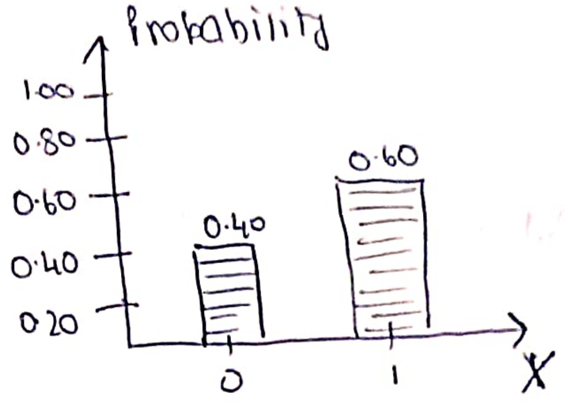
$Pr(K=0) = p^K (1-p)^{1-K} = p^0 (1-p)^{1-0} = 1 - P = 0.40$

Simplified version of PMF:-

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

Mean of the bernoulli distribution:-

Let us take the smartphone example



$Pr(K=0) = 0.40 = 1-p = q$
$Pr(K=1) = 0.60 = p$

Expected value of X,

$$\boxed{E(X) = \sum_{i=0}^{K} K \cdot p(K)} \quad , \text{wher } K = \{0,1\}$$

$$E(X) = \sum_{i=0}^{K} K \cdot p(K) = [0 \cdot p(0) + 1 \cdot p(1)] = [(0 \times 0.40) + (1 \times 0.60)]$$

$$= 0 + 0.60 = 0.60 = p$$

The mean of bernoulli distribution is p which also means that it is probability at K=1 i.e $Pr(K=1) = p$.

Median of bernoulli distribution:-

$$\text{median} = \begin{cases} 0, & \text{if } p < 1/2 \\ 0 \text{ or } 1, & \text{if } p = 1/2 \\ 1, & \text{if } p > 1/2 \end{cases}$$

we can also simplify it as

$$\text{median} = \begin{cases} 0, & \text{if } q > p \\ 0.50, & \text{if } q = p \\ 1, & \text{if } p > q \end{cases}$$

Mode of the Bernoulli Distribution:-

$$\text{mode} = \begin{cases} p, \text{if } p > q \\ q, \text{if } q > p \end{cases}$$

Variance of the Bernoulli Distribution:-

$Pr(K=0) = q = 1-p = 1-0.60 = 0.40$

$Pr(K=1) = p = 0.60$

$\text{mean} = p = 0.60$

$$\sigma^2 = q(0-\text{mean})^2 + p(1-\text{mean})^2$$
$$= 0.40(0-0.60)^2 + 0.60(1-0.60)^2$$
$$= 0.40(0.36) + 0.60(0.16)$$
$$= 0.144 + 0.096 = 0.240 = (0.60)(0.40) = p \times q$$

So, we can write variance as

$$\boxed{\sigma^2 = p \times q}$$

Therefore, standard deviation of Bernoulli Distribution is

$$\sigma^2 = p \times q \Rightarrow \boxed{\sigma = \sqrt{p \times q}}$$

Deriving the variance of Bernoulli Distribution:-

$$\sigma^2 = q(0-p)^2 + p(1-p)^2$$
$$= q(p^2) + p(1+p^2-2p)$$
$$= qp^2 + p + p^3 - 2p^2$$
$$= (1-p)p^2 + p^3 + p - 2p^2$$

$$\sigma^2 = p^2 - p^3 + p^3 + p - 2p^2$$
$$= p - p^2 = p(1-p) = pq$$

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

so variance of bernoulli distribution is the product of probabilites of failure and success

$$\boxed{\sigma^2 = Pr(K=0) \times Pr(K=1)}$$

Binomial distribution

→ In probability theory and statistics, 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 a yes-no question.

→ Each experiment in the sequence has its own boolean valued outcome : success with probability p and a failure with probability $q = 1-p$.

→ A single experiment/trail is called bernoulli trail/bernoulli experiment. A sequence of Bernoulli's/binary outcomes is called a Bernoulli process.

→ For a single trail $(n=1)$, the binomial distribution is a bernoulli distribution.

The binomial distribution is the basis for popular binomial test of statistical significance.

Parameters of Binomial distribution:-

Binomial distribution is denoted by $B(n,p)$

$n \in \{0,1,2,\ldots\} \Rightarrow$ Number of trails/experiments

$p \in [0,1] \Rightarrow$ success probability for each trail

$q = 1-p \in [0,1] \Rightarrow$ Failure probability for each trail

$k \in \{0,1,2,\ldots n\} \Rightarrow$ Number of successes

Since we are dealing with discrete random variables, we can use Probability Mass Function (PMF) to obtain the probability curve.

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

where $K \in \{0,1,2,\ldots n\}$

Mean of the binomial distribution:-

$$\boxed{\text{mean} = np}$$

$n \Rightarrow$ number of trails

$p \Rightarrow$ success probability

variance of the binomial distribution:-

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

$n \Rightarrow$ Number of trails

$p \Rightarrow$ success probability

$q \Rightarrow$ Failure probability $= 1-p$

Examples of Binomial distribution:-

(i) suppose we toss a time 5 times. what is the probability of getting exactly 3 heads?

$n \Rightarrow$ no. of trails / experiments $= 5$

$K \Rightarrow$ no of successes $= 3$

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

for a fair coin $\Rightarrow p = 1/2$

$$Pr(X=3) = {}^{n}C_3\, p^3 (1-p)^{n-3} = {}^{5}C_3\, p^3 (1-p)^{5-2}$$

$$= {}^{5}C_3\, p^3 (1-p)^2 = {}^{5}C_3 (1/2)^3 (1-1/2)^2$$

$$= {}^{5}C_3 (1/2)^3 (1/2)^2 = {}^{5}C_3 (1/2)^5 = \frac{1}{32}\, {}^{5}C_3$$

$$\boxed{{}^{n}C_r = \frac{n!}{r!(n-r)!}} \Rightarrow \text{Binomial coefficient}$$

$$Pr(X=3) = \frac{1}{32}\, {}^{5}C_3 = \frac{1}{32}\left(\frac{5!}{3!2!}\right) = \frac{20}{32\times 2} = \frac{5}{16}$$

$$= 0.3125$$

(ii) Inspecting 10 items in a factory where each item has a 10% chance of being defective.

Number of trails $= n = 10$

Probability of success $= p = 0.10$ (defective item)

Number of successes $= K \Rightarrow$ varies from 0 to 10

What is the probability of finding exactly 2 defective items?

Now, $K =$ no. of successes $= 2$

$$Pr(X=K) = {}^{n}C_K\, p^K (1-p)^{n-K} = {}^{10}C_K\, p^K (1-p)^{10-K}$$

$$Pr(X=2) = {}^{10}C_2\, p^2 (1-p)^8 = \left(\frac{10!}{2!\,8!}\right)\left(\frac{1}{10}\right)^2\left(1-\frac{1}{10}\right)^8$$

$$= \left(\frac{10\times 9}{2}\right)\left(\frac{1}{100}\right)\left(\frac{9}{10}\right)^8 = \frac{9^9}{2\times 10^{10}} = 0.1937$$

Poisson Distribution

⇒ In probability theory and statistics, 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 independent of the time since the last event.

⇒ Properties of poisson distribution:-

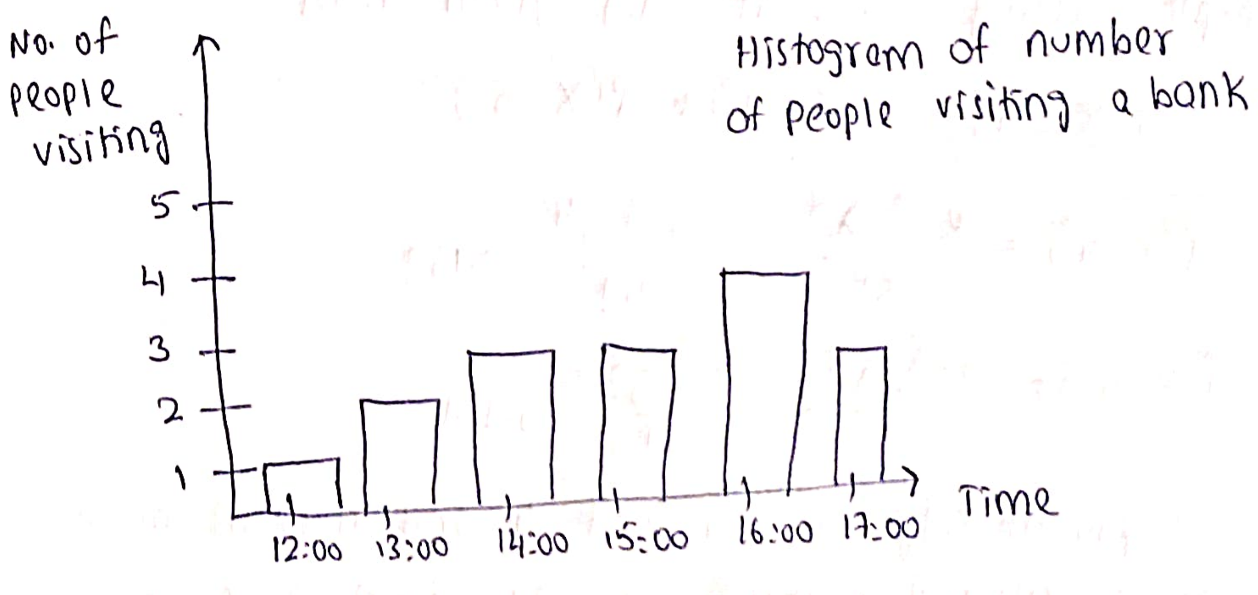
(i) Applicable for discrete random variables. So, we use PMF (Probability Mass Function) to plot the probability distribution graph

(ii) It describes the number of events occuring in fixed time intervals.

Example:-

(a) Number of people visiting a hospital every hour

(b) Number of people visiting banks every hour.



In poisson distribution, we have another parameter λ, where

$\lambda \Rightarrow$ Expected number of events occuring at every time interval.

Probability mass function of poisson distribution is:-

$$\boxed{P(X=K) = \frac{e^{-\lambda} \cdot \lambda^K}{K!}}$$

considering previous example,

Let $\lambda=3$ and $K=5$, then

$$P(X=3) = \frac{e^{-3} \cdot 3^5}{5!} = \frac{e^{-3} \times 243}{120} = 0.101$$

The probability of 5 people exactly visiting the bank in a given hour is 10.1%.

$\Rightarrow$ Finding probability of atleast 4 and maximum 5 people visiting bank at any given hour

$$P(X \geq 4) = P(X=4) + P(X=5)$$

$$P(X=4) = \frac{e^{-\lambda} \cdot \lambda^K}{K!} = \frac{e^{-3} \cdot 3^4}{4!} = 0.168$$

$$P(X=5) = \frac{e^{-\lambda} \cdot \lambda^K}{K!} = \frac{e^{-3} \cdot 3^5}{5!} = 0.101$$

The probability of atleast 4 and maximum 5 people visiting a bank at any given hour is

$Pr(X=4) + Pr(X=5) = 0.101 + 0.168 = 0.269$

The probability of atleast 4 and maximum 5 people visiting bank at any given hour is 26.9%.

Mean of the poisson distribution:-

mean = Expected value of $X = E(X)$

$$\boxed{E(x) = \mu = \lambda \times t}$$

$\lambda \Rightarrow$ Expected number of events occuring at every time interval

$t \Rightarrow$ Time interval.

Variance of the poisson distribution:-

variance is same as the mean in poisson distribution

$$\boxed{\sigma^2 = \lambda \times t}$$

$\lambda \Rightarrow$ Expected number of events occuring at every time interval

$t \Rightarrow$ Time interval

Some properties of poisson distribution:-

→ The entire distribution is defined by a single parameter λ, which represents the average rate or expected number of events in the given interval.

→ when the probability mass function (PMF) is plotted for the poisson distribution, the values of $P(X=K)$ will rise to the peak and then fall

The mode (the value of K at the highest probability) is the integer less than or equal to λ i.e mode is the largest integer less than or equal to lambda if λ is not an integer.

But, if λ is an integer, it has two modes λ and $\lambda-1$.

For example, $\lambda = 3$

$$Pr(X=2) = \frac{e^{-\lambda}.\lambda^2}{2!} = \frac{e^{-3}.3^2}{2!} = 0.224$$

$$Pr(X=3) = \frac{e^{-\lambda}.\lambda^3}{3!} = \frac{e^{-3}.3^3}{3\times2\times1} = 0.224$$

$$Pr(X=2) \approx Pr(X=3), \text{ at } \lambda = 3$$

So, here 3 (λ) and 2 ($\lambda-1$) both are the modes of the poisson distribution.

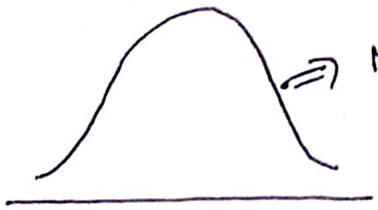
Normal/Gaussian Distribution

⇒ In probability theory and statistics, a normal/gaussian distribution is a type of continuous probability distribution for a real-valued random variable.

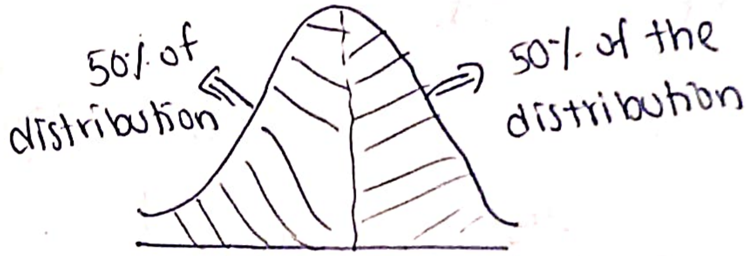
⇒ Properties of normal/Gaussian distribution:-

(i) It is applicable for continuous random variable.

(ii) It's distribution follows a bell-curve



(iii) The distribution will be symmetric i.e 50% of the distribution lies on either sides of the mean



(iv) In normal/Gaussian distribution, mean, median and mode are the same

$$\boxed{\text{mean} = \text{median} = \text{mode}}$$

⇒ Parameters of normal/Gaussian distribution:-

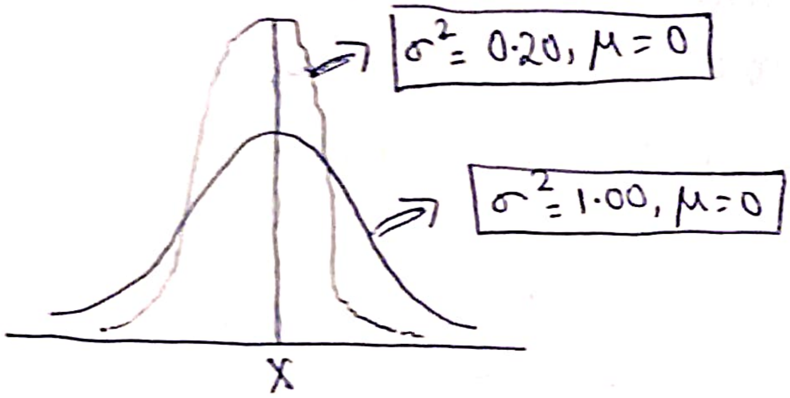
$\mu \Rightarrow$ mean $\in R$ (Real numbers)

$\sigma^2 \Rightarrow$ variance $\in R$

$x \in R \Rightarrow$ Data points

Since we deal with continuous random variables in normal distribution, we use Probability Density Function (PDF) to obtain the curve.

⇒ As the value of variance increases, the spread of the distr increases as well



Example:-

(i) weights of students in a class

(ii) Heights of students in a class

⇒ Probability Density Function of normal distribution:-

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

σ ⇒ standard deviation

σ^2 ⇒ variance

x_i ⇒ Data points

μ ⇒ mean of gaussian distribution

⇒ Mean of the normal distribution:-

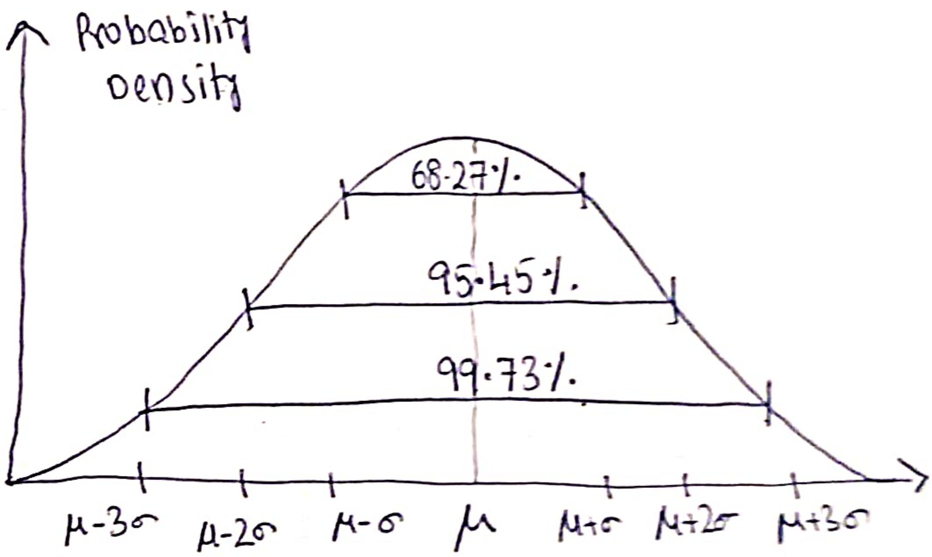
$$\mu = \sum_{i=1}^{n} \frac{x_i}{n}$$

⇒ Variance of Gaussian Distribution:-

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

(*) Empirical rule for normal distribution:-

It is also called as 68-95-99.7 rule, which is one of the most important assumptions in normal distribution.



$Pr(\mu-\sigma \le X \le \mu+\sigma) \approx 68\%$

$Pr(\mu-2\sigma \le X \le \mu+2\sigma) \approx 95\%$

$Pr(\mu-3\sigma \le X \le \mu+3\sigma) \approx 99.7\%$

The probability of random variable to exist between first standard deviation is approximately 68%, in second standard deviation is approximately 95% and in the third standard deviation is 99.7%.

Standard normal Distribution and Z-score

→ The standard normal distribution is a special case of normal distribution where the mean is 0 and the variance is 1

$$\boxed{\text{Standard normal distribution} \Rightarrow \mu = 0, \sigma = 1}$$

⇒ Example of converting normal distribution to standard normal distribution :-

let us take an example, $X = \{1, 2, 3, 4, 5\}$

mean of $X = 3$

variance of $X = 1.414 \approx 1$ (for calculation)

Suppose that X follows normal distribution and now we would like to change it into standard normal distribution.

we can do this using Z-score

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

New data points would be

$X_0' = \frac{X_0 - \mu}{\sigma} = \frac{1-3}{1} = -2$ $X_3' = \frac{X_3 - \mu}{\sigma} = \frac{4-3}{1} = 1$

$X_1' = \frac{X_1 - \mu}{\sigma} = \frac{2-3}{1} = -1$ $X_4' = \frac{X_4 - \mu}{\sigma} = \frac{5-3}{1} = 2$

$X_2' = \frac{X_2 - \mu}{\sigma} = \frac{3-3}{1} = 0$

New data points $(X') = \{-2, -1, 0, 1, 2\}$

so, for the new data points

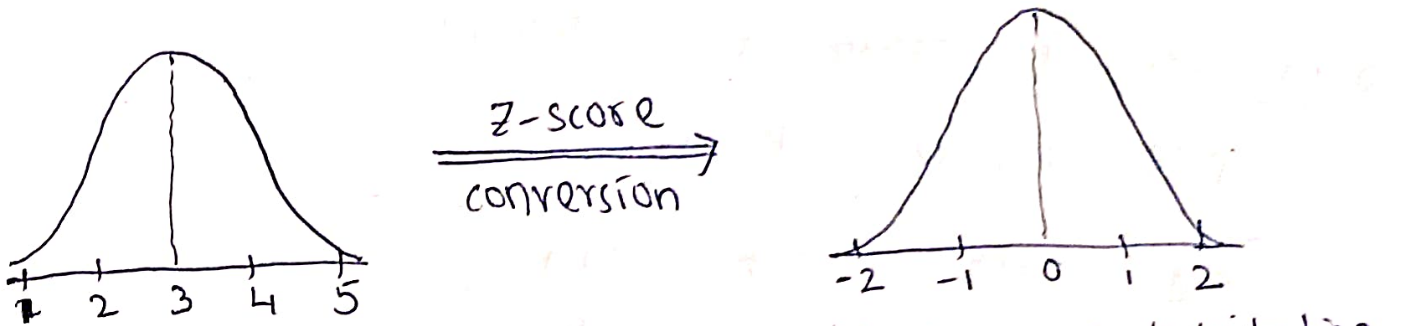
(i) mean is 0

(ii) standard deviation is approximately 1

(iii) still follow normal distribution

Properties of standard normal distribution

In this way, we can utilize z-score to obtain standard normal distribution from normal distribution.



normal distribution

standard normal distribution.

⇒ Another application of z-score is that we can findout how many standard deviations from the mean a datapoint lies.

Example:- $\mu = 4, \sigma = 1$. How many standard deviations away is 4.25 from the mean

$$z\text{-score} = \frac{x_i - \mu}{\sigma}$$

Here ⇒ $x_i = 4.25,\ \mu = 4,\ \sigma = 1$

$$z\text{-score} = \frac{4.25 - 4}{1} = 0.25$$

4.25 is 0.25 standard deviations away from the mean in the right direction (since z-score > 0) of the mean.

Example:- $\mu = 4, \sigma = 1$, how many standard deviations is 2.5 away from the mean

$$\text{Z-score} = \frac{X_i - \mu}{\sigma} = \frac{2.5-4}{1} = -1.5$$

2.5 is 1.5 standard deviations away from the mean on left hand side (since Z-score < 0).

⊛ An important application of Z-score :-

suppose we have a dataset

Age	weight	Height	salary
24	70	175	40K
25	60	160	50K
26	55	160	60K
27	40	130	30K
30	30	175	20K
31	25	180	70K
years	kg	cm	INR

All features are in different units.
Different machine learning algorithms require all features to be in the same unit for better performance.

so, to convert all the features by a technique called STANDARDIZATION to bring them all to same unit.

In standardization,
we use Z-score on every feature to bring all of them into the same unit.

$$\text{Z-score}_{age} = \frac{Age_i - \mu_{age}}{\sigma_{age}}$$

$$\text{Z-score}_{weight} = \frac{weight_i - \mu_{weight}}{\sigma_{weight}}$$

same units

By converting entire data into same unit, we will be able to improve the model performance and accuracy.

Uniform Distribution

⇒ There are two types of uniform distribution:-

(i) continuous uniform Distribution (we used PDF)

(ii) Discrete uniform distribution (we use PMF)

(i) continuous uniform distribution:-

In probability theory and statistics, the continuous uniform distribution/ rectangular distributions are a family of symmetric probability distributions.

This distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds. The bounds are defined by the parameters a and b which are minimum and maximum values.

Properties of continuous uniform distribution:-

(i) It is applicable for continuous random variable, so we use Probability Density Function (PDF).

(ii) It has two parameters

a ⇒ minimum value

b ⇒ maximum value

$$\boxed{-\infty < a < b < \infty}$$

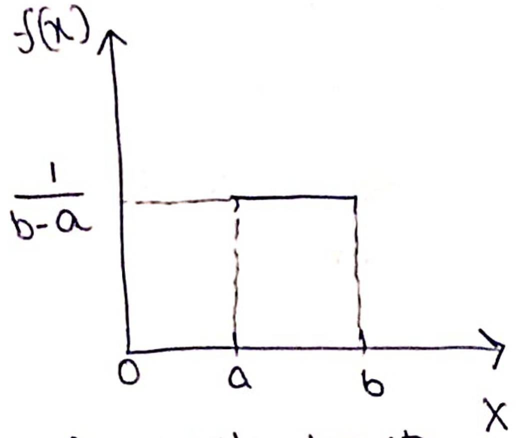
The probability distribution only is present between a and b and in rest of the areas, it is zero.

Probability density function for continuous uniform distribution:-

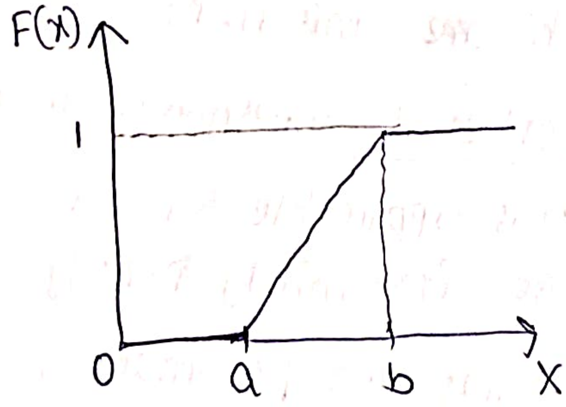
$$PDF = \begin{cases} \dfrac{1}{b-a}, & \text{if } x \in [a,b] \\ 0, & \text{if } x<a \text{ or } x>b \end{cases}$$

Cumulative distribution function is

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



Probability density function



Cumulative distribution function

Mean of the continuous uniform distribution:-

$$\boxed{\mu = \text{mean} = \frac{1}{2}(a+b)}$$

a ⇒ minimum value

b ⇒ maximum value

Median of the continuous uniform distribution:-

$$\boxed{\text{median} = \frac{1}{2}(a+b)}$$

a => minimum value

b => maximum value

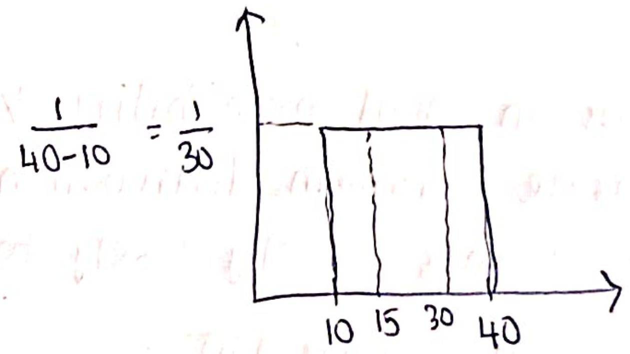
Variance of the continuous uniform distribution:-

$$\boxed{\text{variance}\ (\sigma^2) = \frac{1}{12}(b-a)^2}$$

=> Example of continuous uniform distribution:-

① The number of candies sold daily at a shop is uniformly distributed with a maximum of 40 candies and a minimum of 10.

The probability of daily sales to fall between 15 and 30 candies is:-



$$Pr(15 \le X \le 30) = (X_2 - X_1) \times \frac{1}{b-a} = (30-15)\left(\frac{1}{30}\right)$$

$$= (15/30) = 1/2$$

$$\boxed{Pr(X_1 \le X \le X_2) = (X_2 - X_1)\left(\frac{1}{b-a}\right)}$$

The probability of sales to be more than 20 candies is

$Pr(X \geq 20) \Rightarrow Pr(X \geq 20$ and $X \leq$ maximum value$(b))$

$\Rightarrow Pr(20 \leq X \leq b)$

$$Pr(20 \leq X \leq 40) = (X_2 - X_1)\frac{1}{b-a} = (40-20)\frac{1}{40-10}$$

$$= \frac{20}{30} = 2/3 = 0.667$$

So, the probability of sales to be more than 20 candies is 0.667.

④ Discrete uniform distribution:-

In probability theory and statistics, the discrete uniform distribution is a symmetric probability distribution where in a finite number of values are equally likely to be observed.

Every one of n values has an equal probability $1/n$. Another way of describing discrete uniform distribution would be a finite number of outcomes equally likely to happen.

Properties of discrete uniform distribution:-

(i) It is applicable for discrete random variable

(ii) We use Probability mass Function (PMF).

Example of discrete uniform distribution:-

Rolling a fair dice

$$Pr(1) = Pr(2) = Pr(3) = Pr(4) = Pr(5) = Pr(6) = 1/6$$

Parameters of discrete uniform distribution:-

a ⇒ minimal value i.e minimum possible value

b ⇒ maximum value

n ⇒ b-a+1 ⇒ Total number of possible outcomes

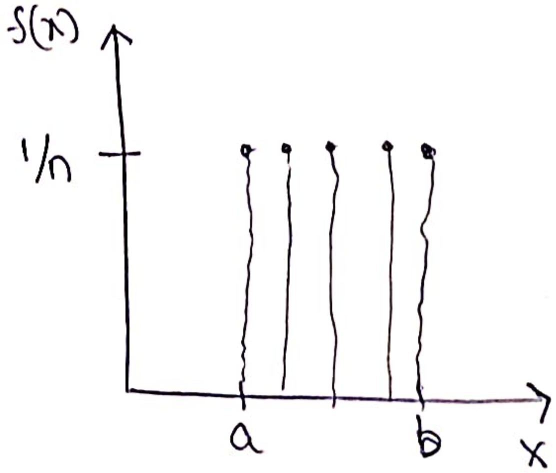
$$\boxed{b \geq a}$$

Probability mass function of discrete uniform distribution:-

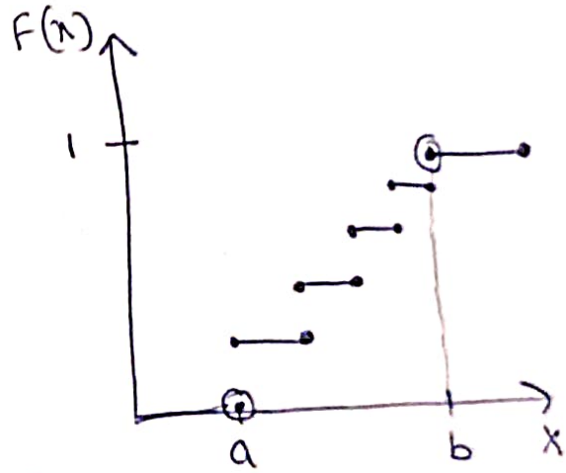
$$\boxed{PMF = \frac{1}{n} = \frac{1}{b-a+1}}$$

b ⇒ maximum possible outcome

a ⇒ minimum possible outcome



Probability mass function



cumulative distribution function

Mean of discrete uniform distribution:-

$$\boxed{mean = \mu = \frac{1}{2}(a+b)}$$

a ⇒ minimal possible outcome

b ⇒ maximum possible outcome

Median of discrete uniform distribution:-

$$\boxed{\text{median} = \frac{1}{2}(a+b)}$$

$a \Rightarrow$ minimum possible outcome

$b \Rightarrow$ maximum possible outcome

Variance of discrete uniform distribution:-

$$\boxed{\text{variance } (\sigma^2) = \frac{1}{12}(n^2-1)}$$

$n \Rightarrow$ Possible number of outcomes $= b-a+1$

Log normal distribution

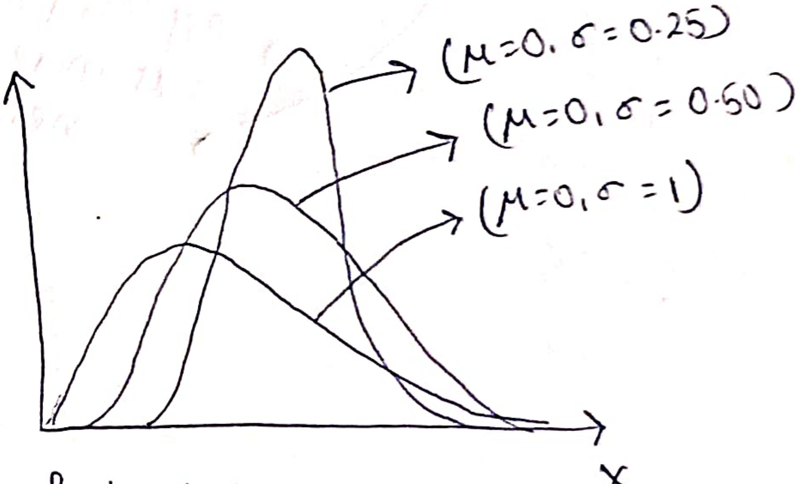
⇒ Log normal distribution / Right skewed distribution in the probability theory is a continuous probability distribution of a random variable whose logarithm is normally distributed.

$$\boxed{\text{If } X \text{ belongs to log-normal distribution then } Y = \ln(X) \text{ belongs to normal distribution}}$$

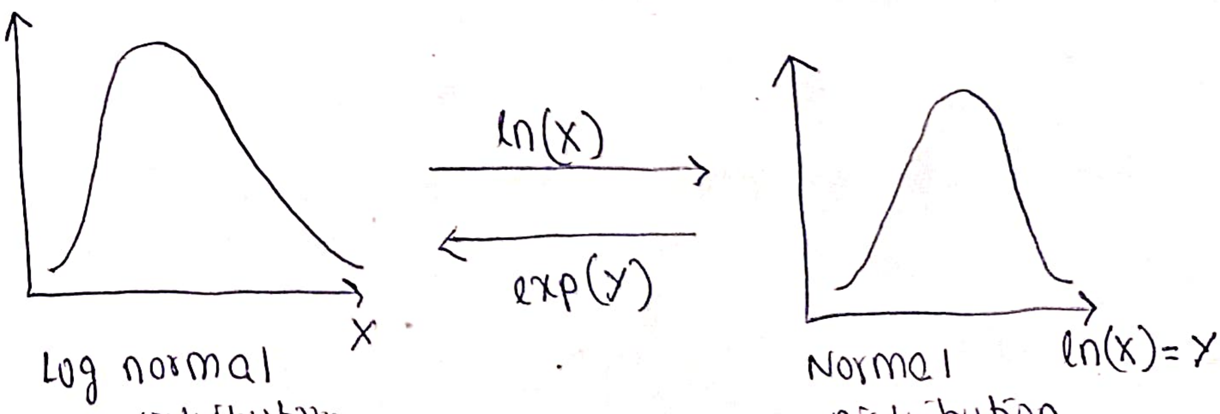
Properties of log-normal distribution:-

(i) It is applicable for continuous random variable.

(ii) We use Probability Density Function (PDF) to obtain curve.



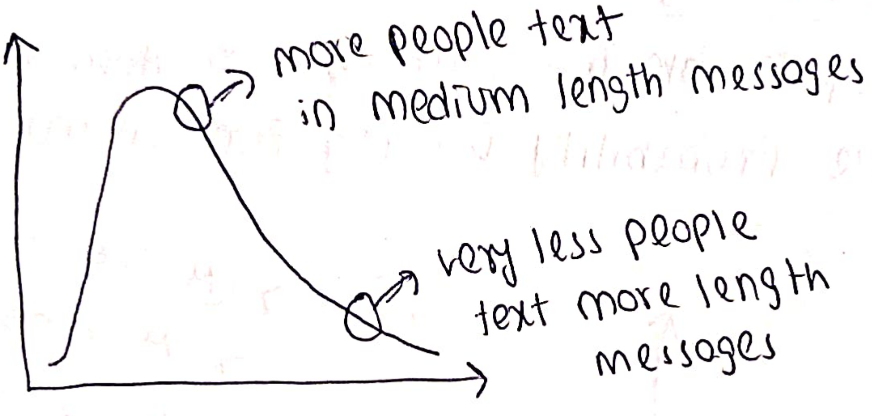
Probability Density Function



Examples of log-normal distribution:-

(i) wealth distribution of the world

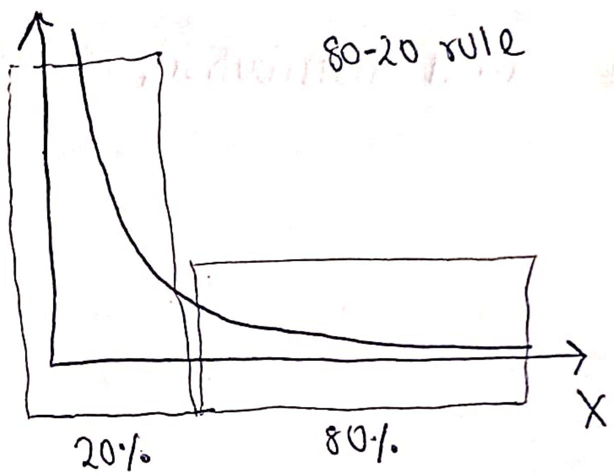
(ii) Length of the comments in discussion forum



Power law distribution

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

One quantity varies as a power of another.



Power law distribution

⇒ Examples of power law distribution:-

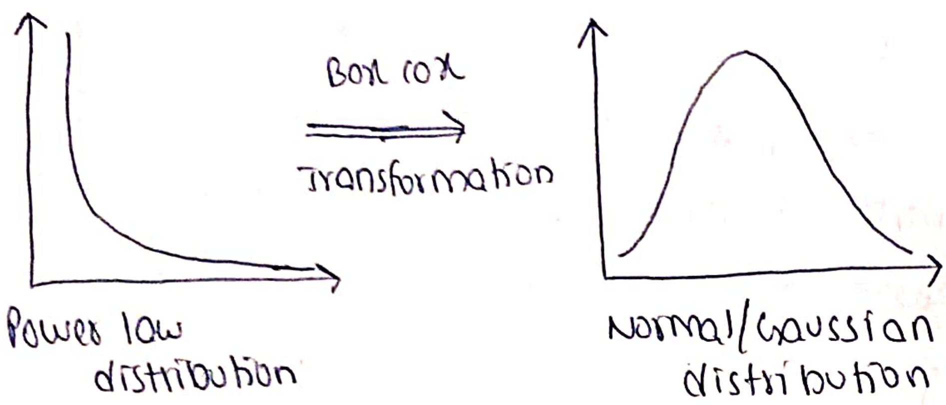
(i) In IPL, 20% of the teams win 80% of the matches.

(ii) 80% of wealth is the world are distributed with 20% of the total world population.

(iii) 80% of the total crude oil is with the 20% of the nations in the world.

(iv) Frequencies of words in sentences in most languages.

→ <u>converting power law distribution into normal distribution:</u>



Any data that has power law distribution is specifically call it as pareto distribution.

In-order to validate normal distribution, we use QQ plot.

Pareto Distribution

⇒ The 80-20 rule - often called as Pareto principle is the direct manifestation of power law distribution, specifically the Pareto distribution.

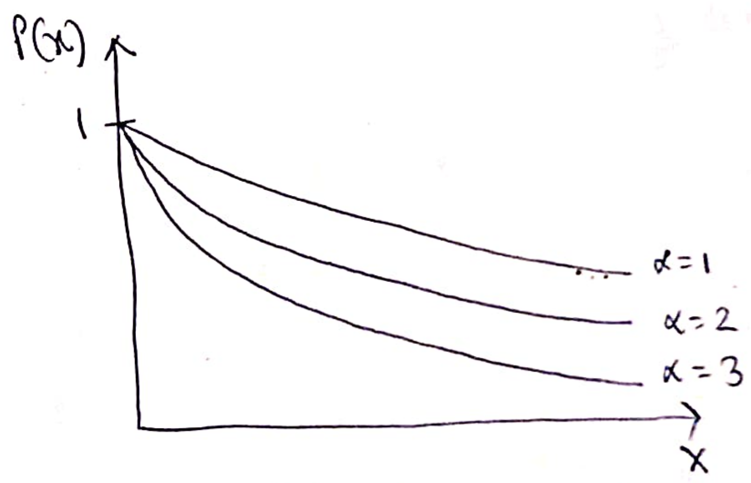
Pareto's 80-20 rule states that about 80% of the effects come from 20% of the causes.

⇒ <u>Hyper-parameter of power law distribution:-</u>

The hyper-parameter of power law distribution is α which is also called exponent or scaling parameter.

Smaller α ⇒ Heavier tail and huge number of outliers.

Larger α ⇒ Lighter tail and less number of outliers.



sample plot of power law distribution with varying α.

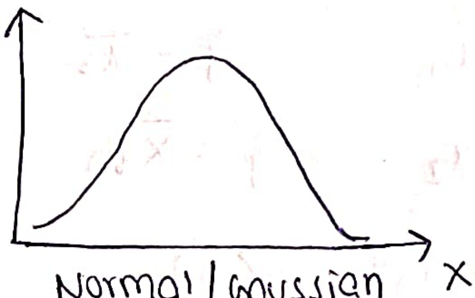
Central limit theorem

⇒ central limit theorem relies on the concept of sampling distribution. sampling distribution is a probability distribution of a statistic for a large number of samples taken from a population.

⇒ central limit states that the sampling distribution of the mean will always be normally distributed, as long as sample size is large enough.

central limit theorem is applicable regardless of whether population has a gaussian or non-gaussian distribution.

For Gaussian distributed population :-



Normal/Gaussian distribution of the population

let us take some samples,

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

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

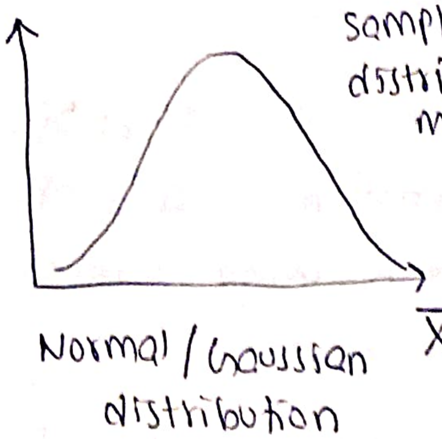
$S_m = \{ ----- \} = \bar{x}_m$

so, the sample means are $\{\bar{x}_1, \bar{x}_2, \bar{x}_3 ----, \bar{x}_m\}$

central limit theorem states that the probability distribution of sampling means will follow a normal/Gaussian distribution i.e

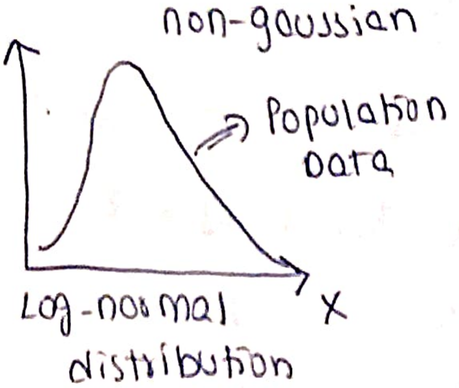
$\bar{x} = \{\bar{x}_1, \bar{x}_2, \bar{x}_3 -- \bar{x}_m\}$

follows normal distribution



If the population data already follows normal distribution, the sample mean will be normally distributed for any sample size, as low as 1 or 2.

<u>For non-Gaussian distribution:-</u>



Let us take samples,

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

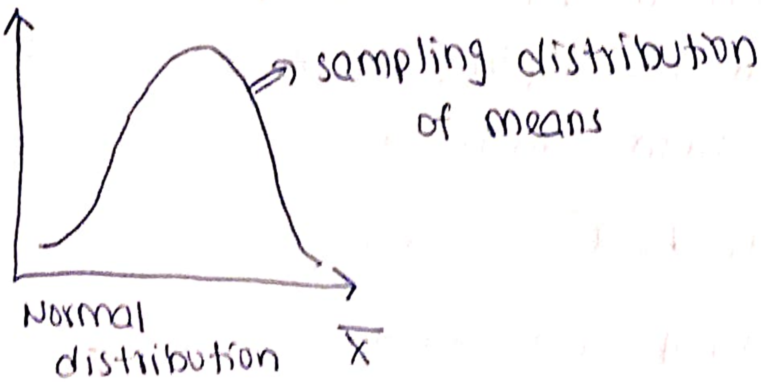
$S_2 = \{x_1, \dots\} = \bar{x}_2$

$S_m = \{x_2 \dots\} = \bar{x}_m$

sample size $(n) \geq 30$

The sample means are $\bar{x} = \{\bar{x}_1, \bar{x}_2, \dots \bar{x}_m\}$

As per the central limit theorem, sampling distribution of means $(\bar{x})$ will also follow normal distribution



For distributions that are non-gaussian, a sample size (n) of atleast 30 is generally recommended to ensure that the sampling distribution of mean becomes approximately normal.

Change in mean and standard deviation:-

Population data

mean $= \mu$

variance $= \sigma^2$

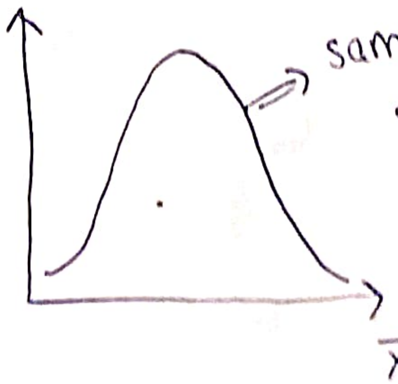
Let us take sample data of sample size n

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

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

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

As per central limit theorem, $\bar{X} = \{\bar{x}_1, \bar{x}_2 \ldots \bar{x}_m\}$ follows normal distribution



sampling distribution of means

$\text{mean}(\bar{x}) = \mu_{\bar{x}} \approx \mu$ (population mean)

$\text{variance}(\bar{x}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

The mean of sampling means will be equal to the population mean approximately.

while, the variance of sample means will be population variance divided by n^2

sample means ⇒ mean $= \mu_{\bar{x}} = \mu$ (population mean)

variance $= \sigma^2_{\bar{x}} = \dfrac{\sigma^2 \text{(Poplation variance)}}{n \text{ (sample size)}}$

(*) so, as per central limit theorem :-

(i) sampling distribution of means of population data will follow normal / gaussian distribution.

(ii) mean of sampling means will be approximately equal to the population mean.

$$\boxed{\mu_{\bar{x}} = \mu}$$

(iii) variance of sampling means will be equal to population variance divided by sample size.

$$\boxed{\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}}$$

Estimates

Estimates is a statistical process of using sample data to approximate unknown population parameters.

⇒ **Point Estimate:-**

A point estimate is a single value calculated from sample data used as a best guess for an unknown population parameter.

Example:-

(i) mean (If a sample of 15 students has an average height 165cm, then 165cm is the point estimate for the population mean of all students in the class)

(ii) If 50 out of 100 randomly sampled people prefer tea over coffee, the sample proportion 0.50 is the point estimate of the overall population.

Drawback of point estimate:-

while point estimate is simple, intuitive and easy to compute, it doesnot account for uncertainity / sampling error.

⇒ **Interval Estimates:-**

An interval estimate expresses uncertainity by giving a range that is likely to contain the population parameter often with a specified confidence interval.

Examples of interval estimates:-

→ Suppose the mean of randomly sampled students is 165cm and the confidence interval of 95% is calculated as 160cm to 170cm, this means that the interval estimate for population mean is [160cm, 170cm] with 95% confidence.

→ For a sample proportion of 0.50, a 95% confidence interval might be [0.45, 0.55], meaning the true population proportion is likely within the range of 0.45-0.55 at the 95% confidence interval.

Key aspect of interval estimates:-

Interval estimates account for sample variability and provide more useful information about reliability of estimates than point estimates.

The width of an interval in interval estimates depends on sample size, variability and chosen confidence interval.