Probability & Statistics :-

Random variable :-

ex:- rolling dice → 6 sides = {1, 2, 3, 4, 5, 6} → when rolled → any one of these. } Random Experiment

equal outcome

random variable → X = {1, 2, 3, 4, 5, 6} → Sample Space.

tossing a coin → Y = {H, T}

P(X = 1) = 1/6 P(X = 2) = 1/6 ....

P(X is even) = 3/6 = 1/2

(probability of X being even)

(P(X=2) + P(X=4) + P(X=6) = 1/6 + 1/6 + 1/6)

P(X is odd) = 1/2

P(x = x_i) → P(x_i) same thing diff notations.

Finite set of values → Discrete random value

→ Height of randomly picked student.

y could be 162, 180, 120, 140, ... → infinite values. → Continuous Random Variable

Outliers :-

Y : height of student.

{122.2, 146.4, 132.5, ... 12.2, 156.3, 92.7 ...}

92.7 → could be an outlier

12.2 → outlies → could be human error (or) actual height

→ Outlies can corrupt data.

→ A discrete value is obtained by counting

→ A continuous value is obtained by measuring.

Sample Space :- Set of all possible outcomes of an experiment.

→ A random variable value depends on the outcome of a random phenomenon.

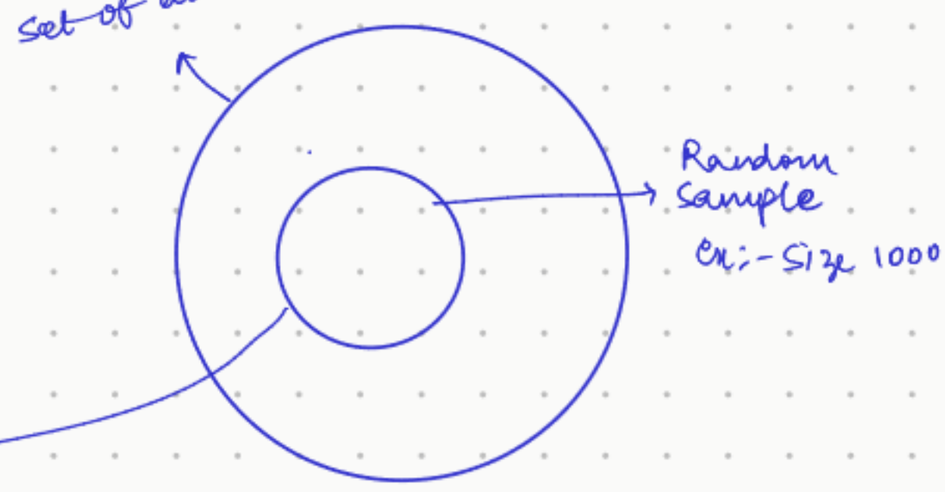
Population & Sample :-

→ Estimating the average height of human

↳ $\mu = \frac{1}{Pop} \sum_{i=1}^{Pop} h_i$ (IMPOSSIBLE)

So we estimate.

often represented by $\bar{x}$ ← $\bar{x} = \frac{1}{1000} \sum_{i=1}^{1600} h_{is}$



→ As sample size increases $\bar{x} = \mu$

$\bar{x}$ → Sample mean

μ → original mean

Sampling is of two types:-

(i) Simple Sampling

(ii) Stratified Sampling ⟶ Unbiased sampling & more accurate results.

ex:- 1000 people ⋯⋯⋯⋯⋯> original sample

- 400 have cars
- 300 have bikes
- 200 have cycles
- 100 have nothing

Sample size ⟶ 250

Simple Random Sampling

250 could have cars (or)

250 could have bikes (or)

100 bikes + 150 cars.

Stratified Random Sampling

- Cars ⟶ 100
- bikes ⟶ 75
- cycles ⟶ 50
- nothing ⟶ 25

} These are random but equal imp to all classes

Gaussian Distribution:- (AKA Normal Distribution)

→ If X is a continuous random variable & X has a PDF curve (⩘), then we say X has a Gaussian distribution.

PDF = Probability Density Function

[Mean, Median & Mode are all same for GD. There are equal number of measurements above & below the mean]

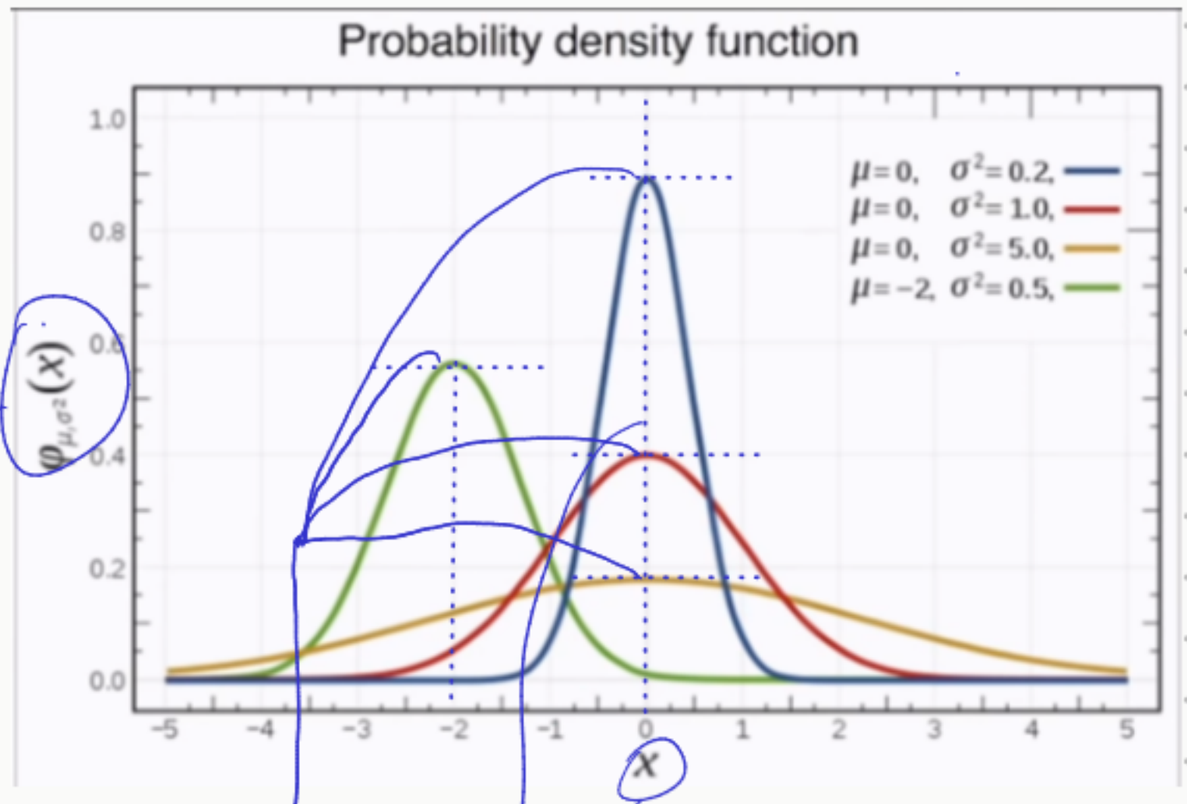
Why learn?

→ Stuff in nature tends to follow GD.

→ heights, weights of people follow it. (Natural Phenomenon)

→ They are simple models that summarize R.V.

ex:-



PDF

μ = mean
σ^2 = variance
} ⟶ parameters

[If this is not known, then we can't]

[If we are given μ, σ^2 & told that X follows GD, we can plot PDF. We don't need the whole data]

Variance is a measure of spread.

Red, Yellow, Blue have $\mu = 0$ but varying variances.

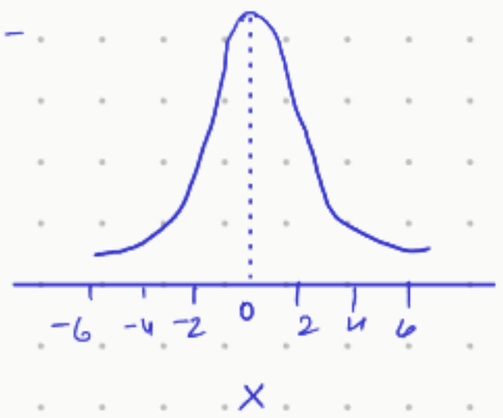
The peak of curve is usually at 'μ'.

X → Continuous Random Variable

→ The parameters of Gaussian Distribution are μ & σ^2

$X \sim N(\mu, \sigma^2)$ ($\Longrightarrow$ X follows Gaussian Distribution with μ & σ^2)

ex:-



→ $X \sim N(0, 2)$

→ $P(X = x) = p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$

↑ Probability Density at a point (x). PDF at any given point gives the probability density at that point. Probability of getting a single discrete value is 0.

ex:- If $\mu = 0$, $\sigma^2 = 1$, $\sigma = 1$

$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{\left(\frac{-1}{2}\right)x^2\right\} = y$

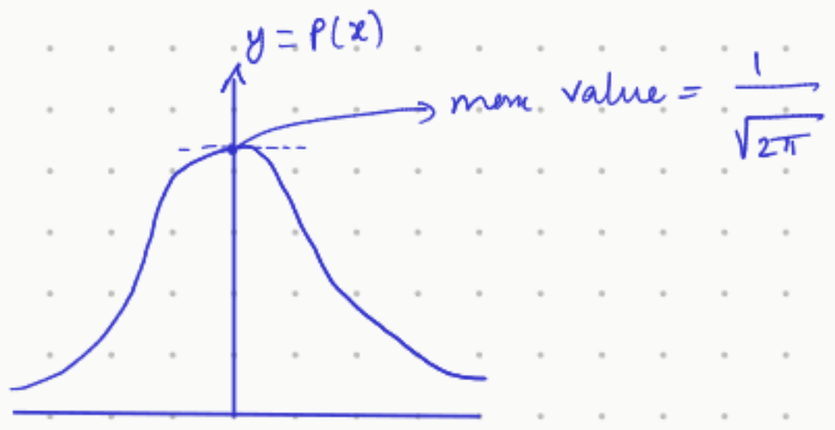
↓ constants

further simplifying

$y = \exp(-x^2)$ → when plotted

as x increases y decreases.

as x decreases y decreases.



conclusions :-

① x moves away from μ, y decreases.
② Graph is symmetric.
③ In this particular graph, y is reducing exponential squared. (e^{-x^2})

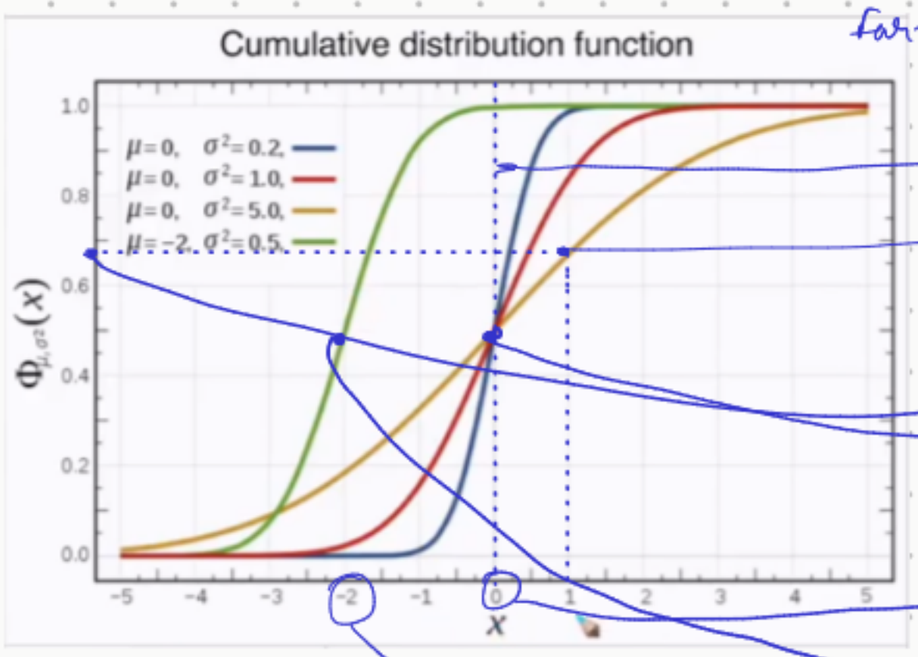
→ If p(x) is the probability density at a point 'x', the probability can be obtained by computing the integral of p(x) over a given interval.

i.e., probability of getting $X \in [a, b]$ is $\int_a^b p(x)\,dx$

Cumulative Distribution Functions (CDF) of Gaussian Distributions/Normal Distributions :-

As σ^2 increases, CDF goes farfromcenter line

CDF of a random variable looks like



→ $P(X \leq 1) = 0.65$

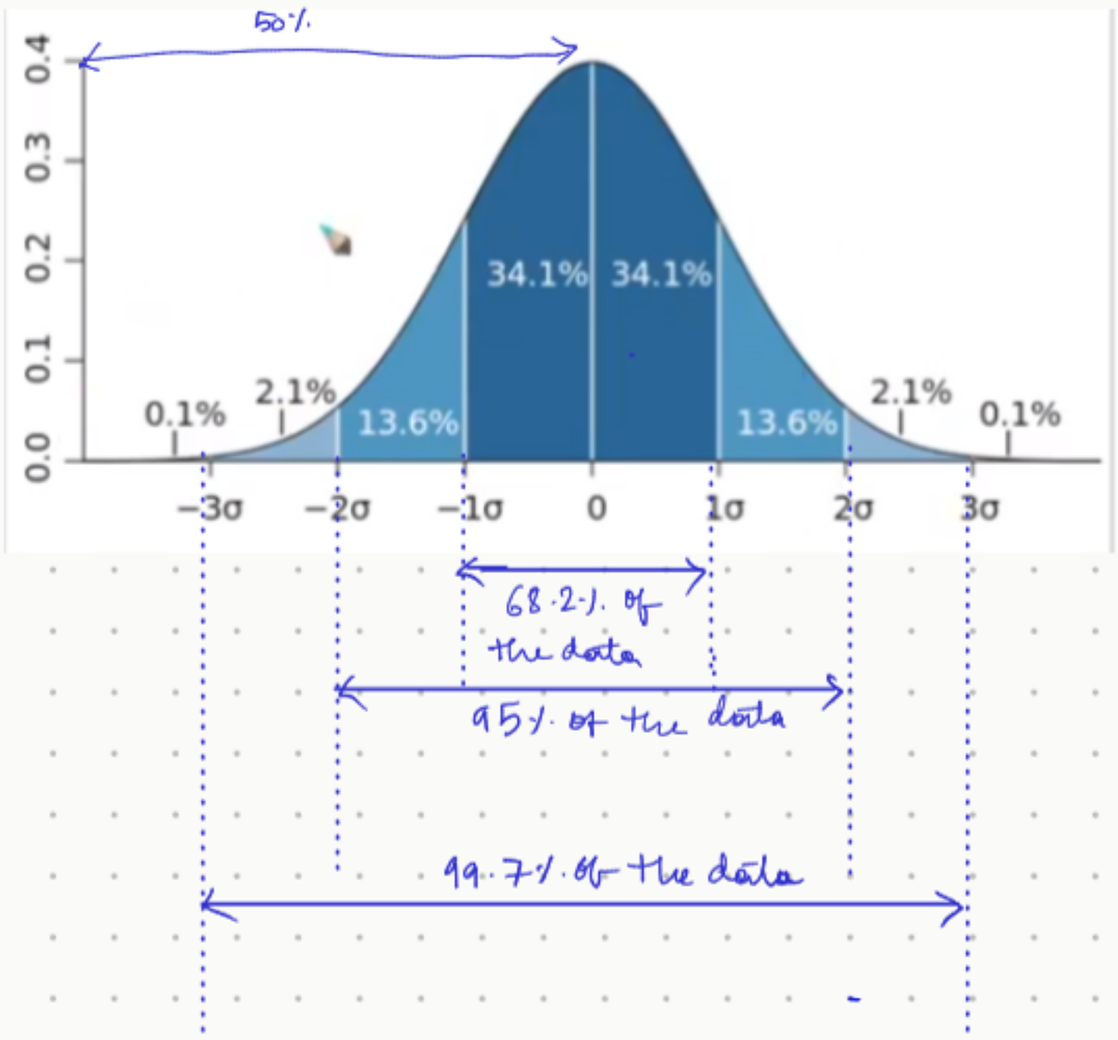
→ $\mu = 0$, center of CDF is at 0.

→ $\mu = -2$, center of CDF is at -2

$$CDF = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \longrightarrow \text{No need to memorize}$$

68-95-99.7 rule :-

if $\mu = 0, \sigma^2 = 4 \Rightarrow \sigma = 2$ $\quad X \sim N(0, 4)$



How is this useful?

ex:- if human population weight

$X \sim N(150, 25)$ — 150 → μ, 25 → σ

$\Rightarrow$ 68.2% of human population lies b/w (150 − 25, 150 + 25)

95% of people (150 − 50, 150 + 50)

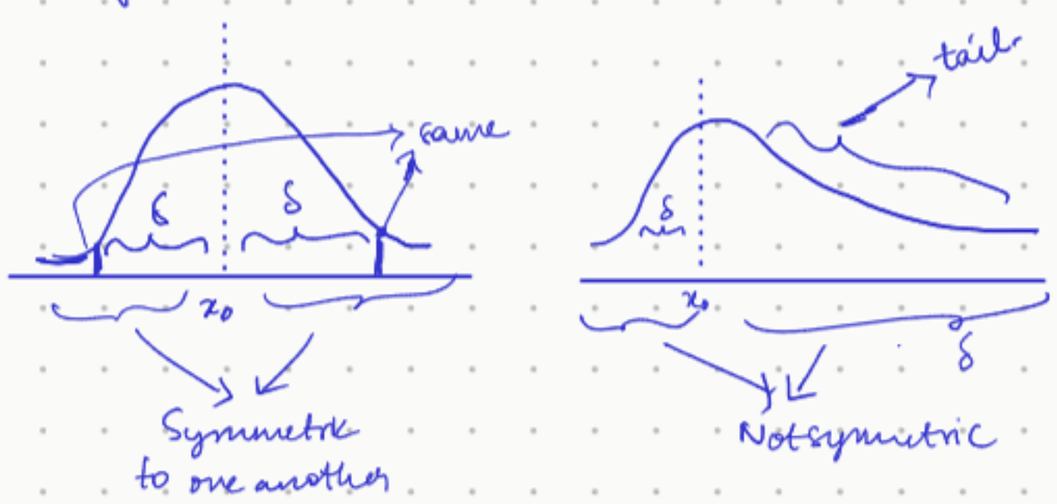
99.7% of people (150 − 75, 150 + 75)

→ A standard gaussian distribution always has a mean of 0 & variance 1.

If it has other mean & variance, it's a non standard gaussian distribution.

Symmetric Distribution, Skewness & Kurtosis :-

→ They help understand shape of PDF.



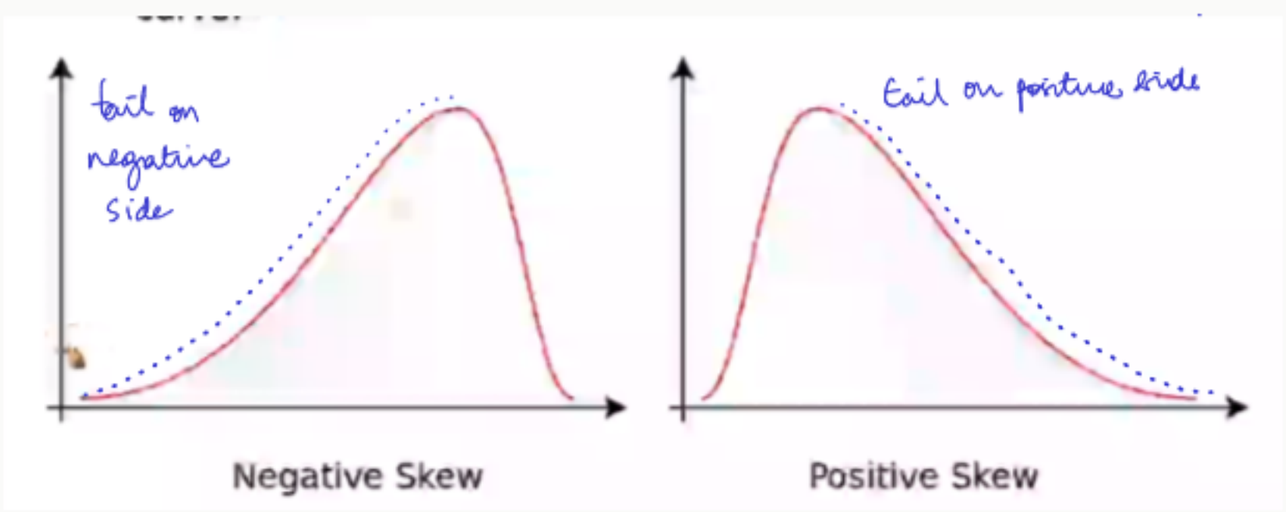
→ A probability distribution is said to be symmetric if and only if there exists a value x_0 such that

$f(x_0 - \delta) = f(x_0 + \delta)$ for all real numbers δ

$f(x)$ is the height of PDF at any point 'x'

Skewness :-

Skewness is a measure of asymmetry -



$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^{3}}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$$

③ → if 2 = variance

[ ]$^{3/2}$ → sample std-deviation

Kurtosis :-

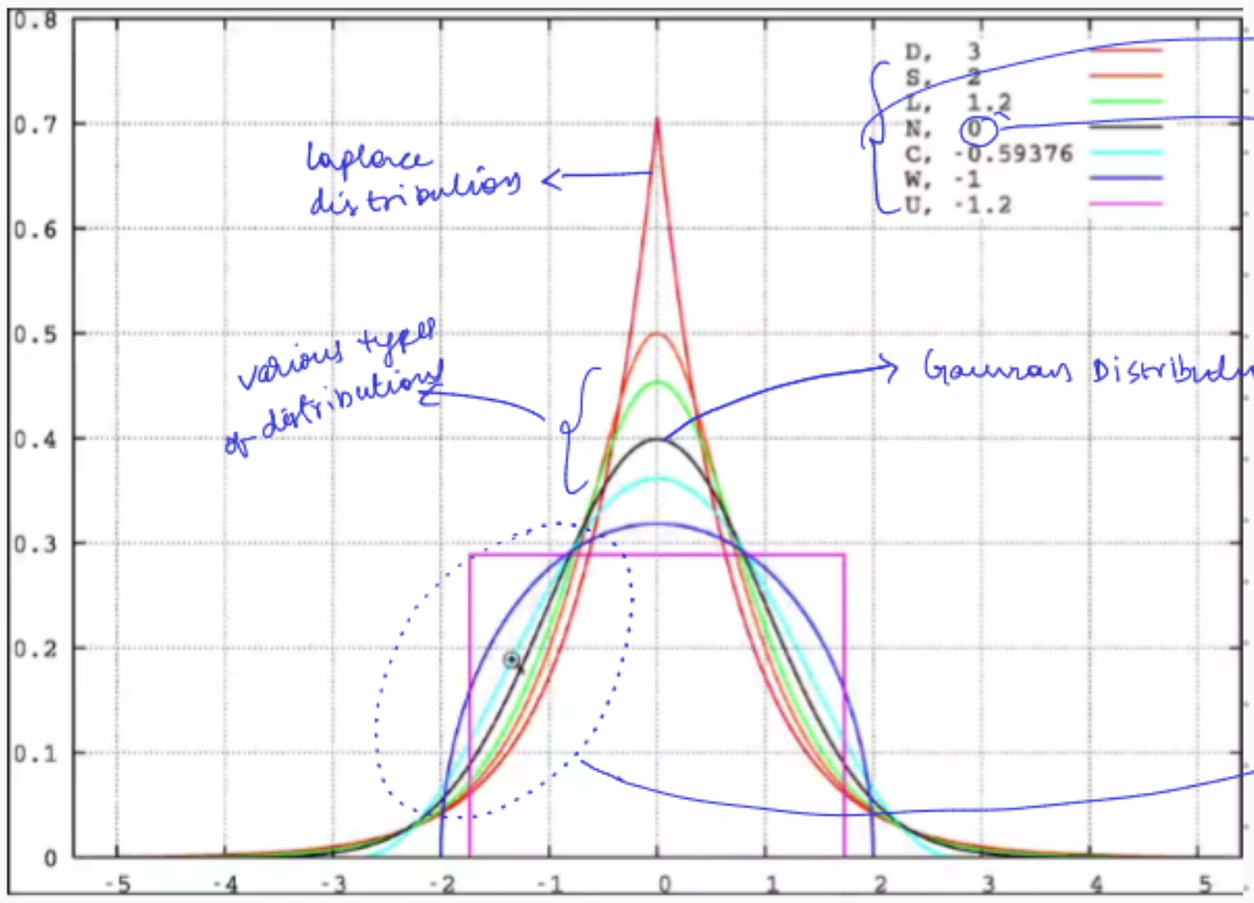
Kurtosis of Gaussian Distribution = 3

→ Measure of tailedness

excess kurtosis = kurtosis − 3

$$\text{excess kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}_{\text{variance}}\right)^2} - 3$$

→ Kurtosis is not a measure of peakedness. Might look like it -



→ excess kurtosis values.

→ Gaussian distribution, excess kurtosis = 0 ⇒ Kurtosis = 3

→ The rate at which the tails touch axis is different

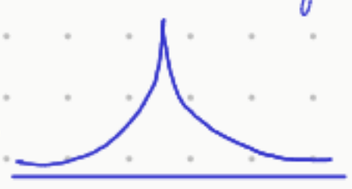
→ Kurtosis tells us if there are outliers are there in the data.

→ Kurtosis is used in analysis of trading, finance etc-)

→ Kurtosis measures how much density/weight is in tails compared to middle part

→ If there are some stocks. Some stocks fall ↓
some go up ↑
some go up enormously ↑

To estimate risk after buying some stocks, we model data - Its generally assumed that they follow gaussian distributions for ease of calculations (Pre 2008 financial crisis). Acc to gaussian distributions chances of losses are less & has good chance of high profits But now, instead of assuming the data is a gaussian distributions, we look at the actual data & calculate the kurtosis

If kurtosis is > 3 ⇝  ⇒ chances of extreme profits are high & chances of losses are also high-

If kurtosis = 3 ⇒ Follows gaussian distribution

If kurtosis < 3 ⇒  ⇒ chances of extreme profits are low & chances of losses are also fairly low.

→ Kurtosis has a degree of 4. So if there is large deviations, then it will lead to higher kurtosis value.

→ High kurtosis is caused by infrequent extreme deviations rather than frequent modestly sized deviations.

→ In order to compare kurtosis b/w two curves, they need to have the same variance.

Standard Normal Variate:- (z)

$Z \sim N(0,1)$

$\mu = 0$

$\sigma^2 = 1$

Let $X \sim N(\mu, \sigma^2)$

$X = [x_1, x_2, x_3, \ldots, x_{50}]$

Standardization:- $x_i' = \frac{x_i - \mu}{\sigma}$

$\Rightarrow x_i' \sim N(0,1)$

↑ Standard Normal Variate.

Given any random variable X, where $X \sim N(\mu, \sigma^2)$

$$z = \frac{x - \mu}{\sigma}$$ → This is also called as basic z-score formula. This basically tells us how many σ is x away from μ.

$\Rightarrow z \sim N(0,1)$

Why?:-

① After standardization, PDF becomes

→ 68.2% of the data

(∵ 68-95-99.7 rule)

② If we are comparing multiple GDs with different μ & σ^2, doing this helps understand better & interpret better

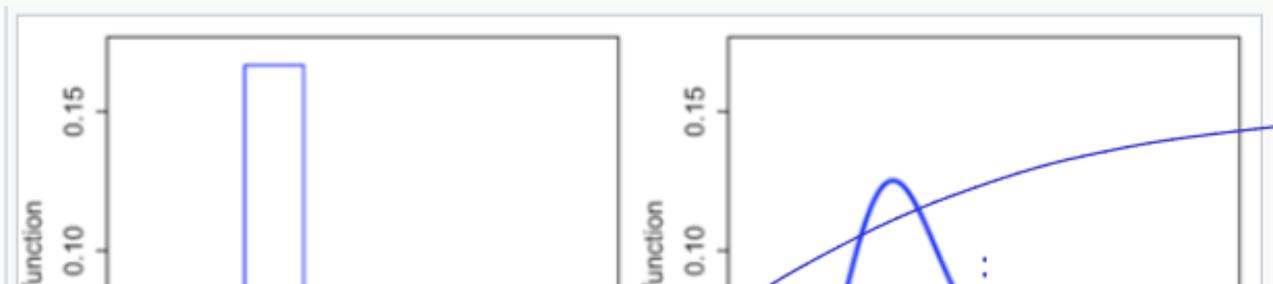
→ We use StandardScaler for standardization & MinMaxScaler for Normalization.

Two techniques of feature scaling.

Normalization → $x_i = \frac{x_i - x_{min}}{(x_{max} - x_{min})}$

Kernel Density Estimation:-

→ Used for smoothing histograms to obtain PDFs.



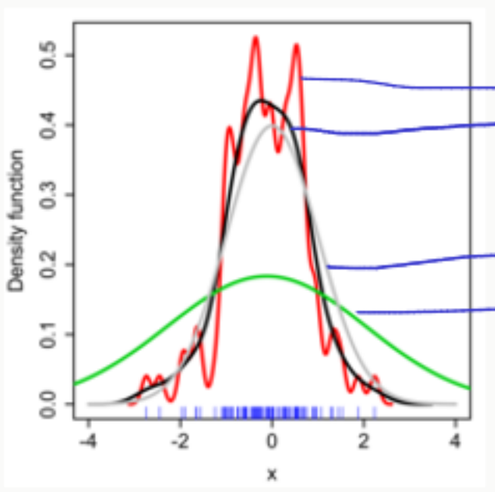
Comparison of the histogram (left) and kernel density estimate (right) constructed using the same data. The six individual kernels are the red dashed curves, the kernel density estimate the blue curves. The data points are the rug plot on the horizontal axis.

→ give height at that point

→ Variance of the gaussian kernel is known as Bandwidth.

→ Gaussian Kernel.

→ Since there are many points here, the number of means will be high. So the PDF will be high. In normal histogram,



→ Narrow Bandwidth

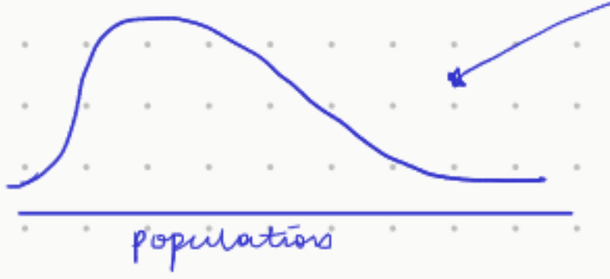
→ Medium Bandwidth

→ Actual Distribution.

→ Wide Bandwidth

Sampling Distributions & Central Limit Theorem :-

Let X be distribution of incomes over populations [Not necessarily Gaussian]



Let S_1 be random sample of size n (Let n = 30) → mean = $\bar{x}_1$ → Sample means.

" S_2 " " " " → mean = $\bar{x}_2$

" " " " " "

" " " " " "

" " " " " → mean = $\bar{x}_m$

These will also have a distribution

The distribution $\bar{x}_i$ = Sampling distributions of sample means.

Central Limit Theorem :- If original distribution 'X' has finite mean (there can be infinite mean ex:- pareto) & variance & 'm' samples are created of size 'n' whose sample means are $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots\ldots$ whose distributions is called sampling distributions of sampling mean, central limit theorem states that

$$\bar{x}_i \longrightarrow N\left(\mu, \frac{\sigma^2}{n}\right) \text{ as } n \to \infty$$

(But IRL if $n \geq 30$, then it becomes gaussian. Rule of thumb)

mean is same as original or that of distributions

Gaussian Distributions

Why? → By using just m x n datapoints, we are able to estimate μ & σ^2 of any distributions, if we just know that they are finite.

Quantile Quantile Plot (Q-Q-Plot) :-

Let X be a random variable

$X : x_1, x_2, x_3, \ldots\ldots, x_{500}$

Question :- Is X Gaussian Distributed. Q-Q-Plot helps us identify. Other techniques such as statistical testing (KS testing) etc, also exist.

Q-Q-Plot → graphical method

KS testing → more powerful.

How? :- ① Sort X & calculate percentiles.

$x'_1, x'_2, x'_3, \ldots\ldots, x'_{500}$

$x'_5 \longrightarrow 1^{st}$ percentile $\longrightarrow x^{(1)}$

$x'_{10} \longrightarrow 2^{nd}$ " $\longrightarrow x^{(2)}$

$\vdots$

$x'_{500} \longrightarrow 100^{th}$ percentile $\longrightarrow x^{(100)}$

② take $Y \sim N(0,1)$

$Y : y_1, y_2, y_3, \ldots\ldots, y_{1000}$

$\downarrow$

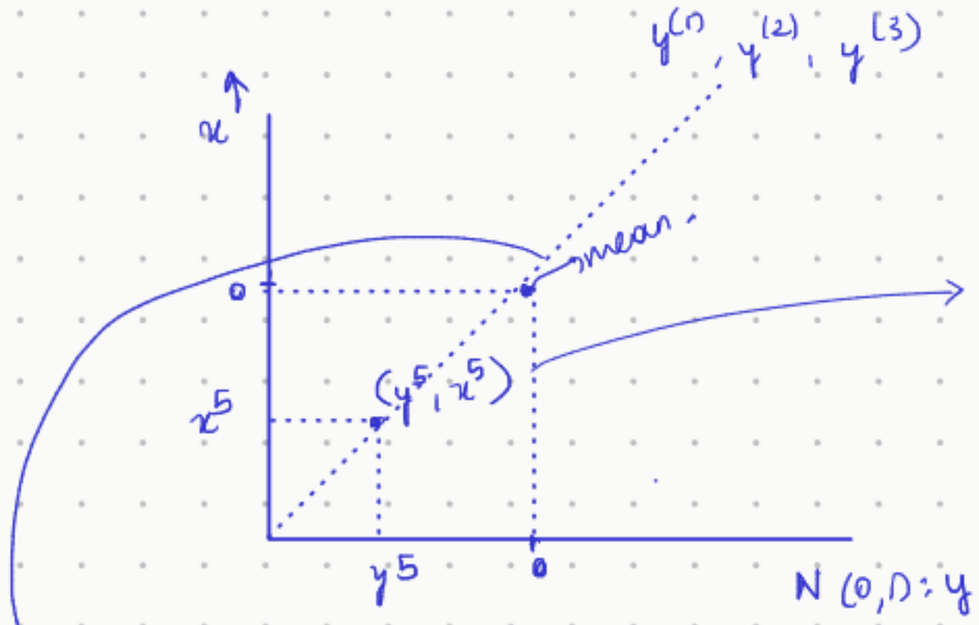
$y'_1, y'_2, y'_3, \ldots\ldots, y'_{1000}$

$\downarrow$

$y^{(1)}, y^{(2)}, y^{(3)}, \ldots\ldots, y^{(100)}$

These are called as Theoretical Quantiles

③ Plot Q-Q-Plot using $x^{(1)}, x^{(2)}, x^{(3)}, \ldots$



we get 100 such pairs (y^1, x^1), (y^2, x^2), ⋮, (y^{100}, x^{100}) — All are plotted.

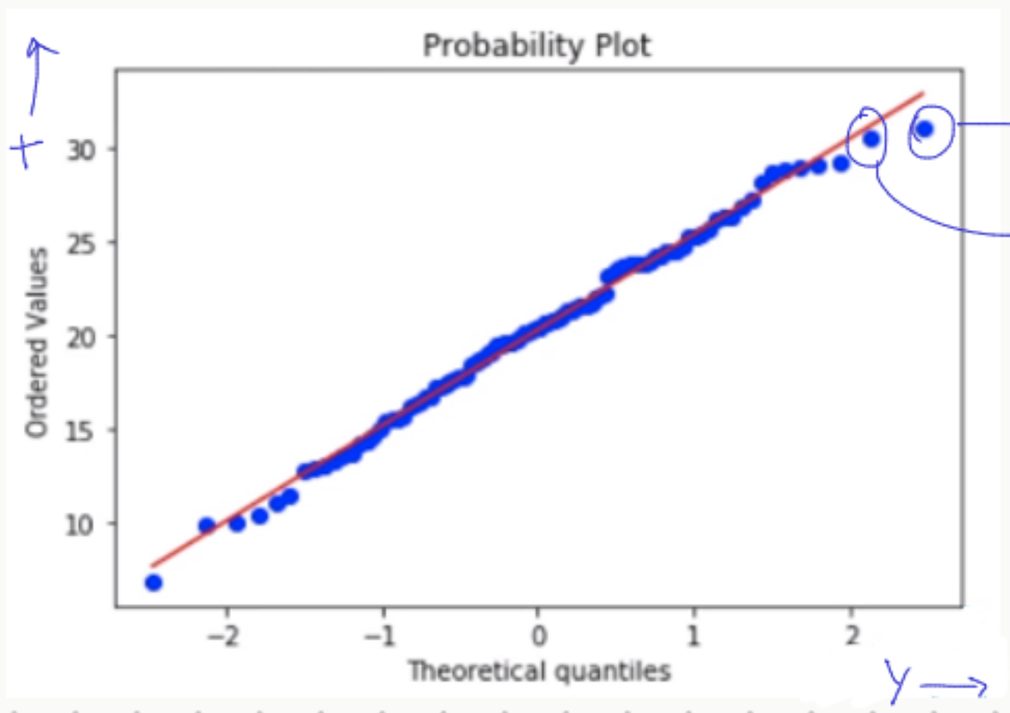
→ If (y^i, x^i) $\forall$ i → 100, lie roughly on a straight line, they y & x have similar distributions.

⇒ X also has gaussian distributions.

→ stats.probplot( )

→ If size of X is small, it is hard to interpret Q-Q plot. Won't be a straight line.

→ Another use of Q-Q plot is: given X, Y, does X & Y have the same distribution.



Generating Y :- (theoretical Quantiles)

```
#Q-Q plot
import numpy as np
import pylab
import scipy.stats as stats

# N(0,1)
std_normal = np.random.normal(loc = 0, scale = 1, size=1000)
```

mean, std-dev

Generating X :-

```
# generate 100 sanples from N(20,5)
measurements = np.random.normal(loc = 20, scale = 5, size=100)
#try size=1000


stats.probplot(measurements, dist="norm", plot=pylab)
pylab.show()
```

How & where to use distributions :-

→ All probability concepts are used for EDA most of the time.

→ If data is Gaussian Distributed, many assumptions can be made by plotting PDF & CDF

Margin of error :- How much deviations from original numbers is allowed in samples.

Chebyshev's Inequality :-

→ If we know that data is Gaussian Distributed, we can apply 68-95-99.7 rule.

→ But if we don't know the distribution, but we know μ, σ can we make assumptions like this ?

(μ: finite; σ: non zero & finite)

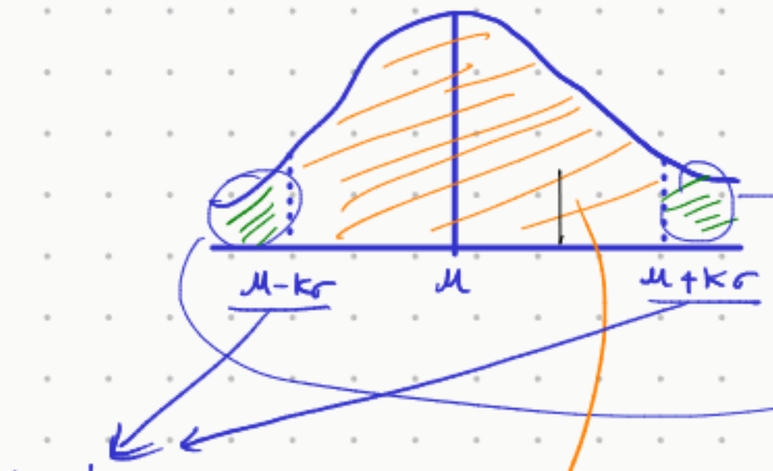
→ Chebyshev's Inequality stats that

'if X is a random variable with finite mean μ and non zero & finite σ

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

k std deviations

68 → 1 std dev away
95 → 2 std dev away
99.7 → 3 std dev away



Probability of being in this regions

$$\Rightarrow P(\mu-k\sigma < X < \mu+k\sigma) \geq 1 - \frac{1}{k^2}$$

→ From Chebyshev's inequality we can conclude that 75% of any distributions lies within $[(\mu-2*\sigma)$ & $(\mu+2*\sigma)]$ and 90% within $[(\mu-3*\sigma)$ and $(\mu+3*\sigma)]$

→ In case the range is not symmetric, there are asymmetric variations of chebyshev as well.

Uniform Distributions:-

- Discrete
- Continuous.

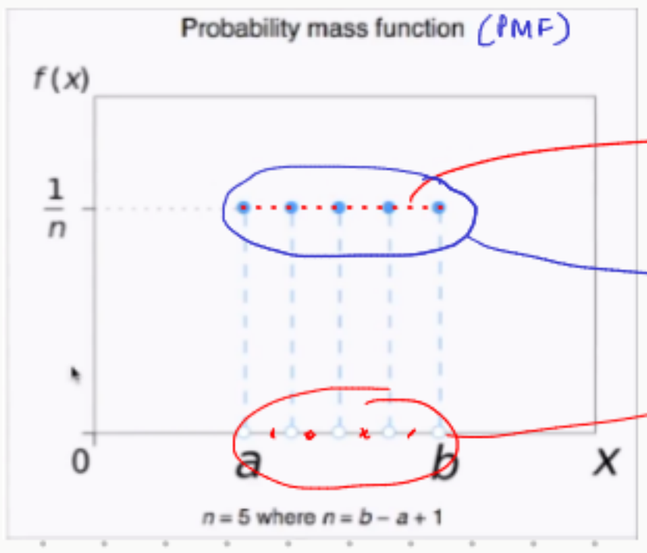
↳ Just like random variables

PDF (Probability Density Functions) are drawn for continuous random variables.
PMF (Probability Mass Functions) are drawn for discrete random variables.

→ Probabity of getting a value in dice is same for all values. This is called equiprobable

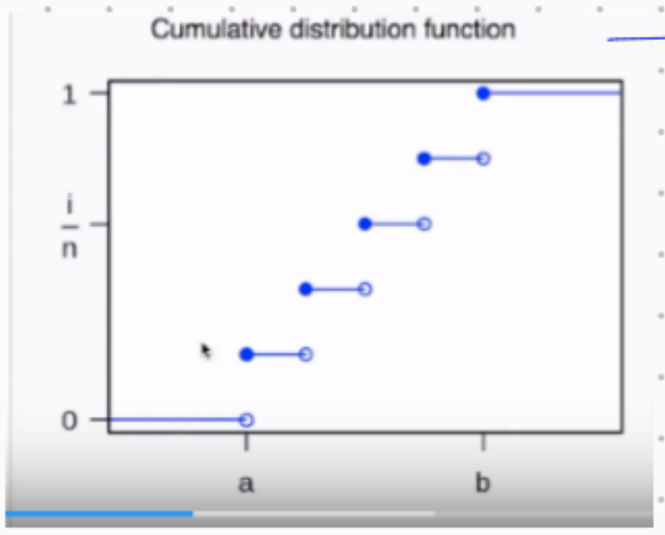
→ CRV has $N(\mu, \sigma)$ as parameters.
DRV has (a, b) as parameters.
↳ both are numbers.
and
$n = b-a+1$. Another formula for calculating $n = \frac{(b-a+k)}{k}$ where k is the common difference.
↳ number of outcomes.

→ In uniform distributions all the variables are equiprobable. ie, $\frac{1}{n}$



This line can't be drawn because these points don't exist in uniform distribution

∴ Only these does exist in PMF



→ CDF

→ This is called a non smooth function

Props of URV :-

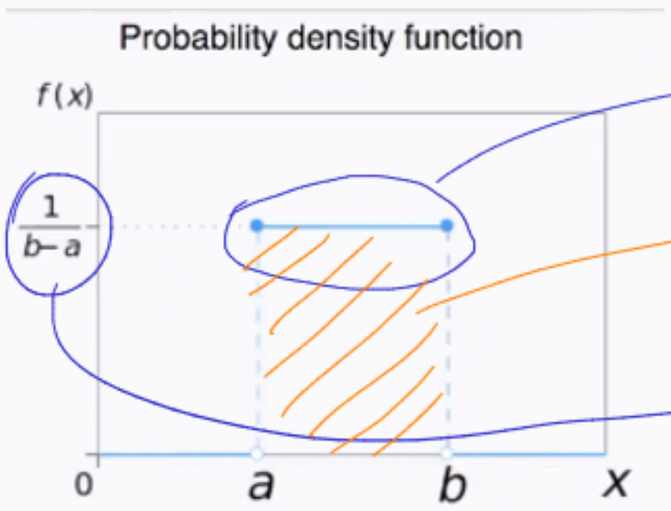
Mean = $\frac{a+b}{2}$

Median = $\frac{a+b}{2}$

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

Skewness = 0

Continuous Random Variable :-

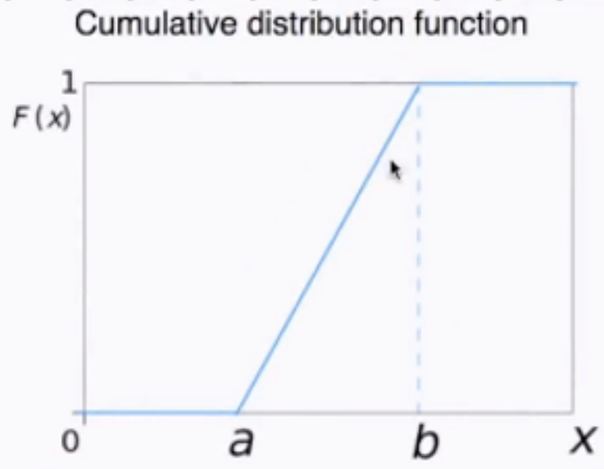


→ Line can exist here because it's continuous.

$\frac{1}{b-a}$ because this area has to be 1. The length is $b-a$. So the height is $\frac{1}{b-a}$.

Using maximum convention

CDF :-



Property of CURV :-

Parms: a, b

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

Mean = $\frac{a+b}{2}$

Median = $\frac{a+b}{2}$

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

PDF → Density of data at the point

PMF → Probability of finding that point there

Random Number generators :-

→ Most random number generators generate uniform random variables unless explicitly specified.

→ python's random.random() picks a number b/w 0 & 1 with uniform probability.

ex :-

→ Picking 30 random values from iris

```
n=150   // Length of his DS
m=30    // Sample size
p= m/n      // 0.2

Sample_d = []
for i in range(0, n):
    if random.random() ≤ P:
        Sample_d append (d[i,:])

len(sample_d)
```

random.random() picks value b/w 1 & n with equal probability.
Chances of getting val ≤ 0.2 = 20%

20% of 150 vals = 30

≈ 30
Not exactly 30

→ This is an application of continuous uniform random variables.

Bernoulli & Binomial Distribution :-

→ Discrete Distributions.

→ Bernoulli has only two outcomes. One has a probability of p while the other has a prob of 1-p

value 1 ↓ , value 0 ↓

ex:-
$X \sim Bernoulli(p=0.5)$

$$PMF = \begin{cases} q=(1-p) & \text{for } k=0 \\ p & \text{for } k=1 \end{cases}$$

Mean = p

variance = pq

Binomial Random Variable :-

Let X be tossing coins

$X \sim Bernoulli(p=0.5)$

Let Y = {Number of heads when coin is toss n times (let n be 10)}

$Y \in \{0,1,2,3,4,\cdots,10\}$

Now $Y \sim Binomial(n, P)$ → probability of getting 1.
number of trials

Params :-

$$PMF = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } P(Y=k)$$

Not used often in Machine Learning.