Probability & Statistics :-

Random variable :-

ex:- rolling dice → 6 sides = {1, 2, 3, 4, 5, 6} — any one of them · equal outcome — Random Experiment

when rolled

random variable — $X = \{1,2,3,4,5,6\}$ → Sample Space ·

tossing a coin → $Y = \{H, T\}$

$P(X=1) = 1/6 \quad P(X=2) = 1/6 \ldots$

$P(X \text{ is even}) = 3/6 = 1/2$

probability of X being even

$\left(P(X=2) + P(X=4) + P(X=6) = 1/6 + 1/6 + 1/6\right)$

$P(X \text{ is odd}) = 1/2$

$P(x = x_i) \longrightarrow P(x_i)$ 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.

{1.22.2, 146.4, 132.5, ···· 12.2, 156.3, 92.7 ··· }

→ could be an outlier

outlier → could be human error (or) actual height

→ Outliers 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.

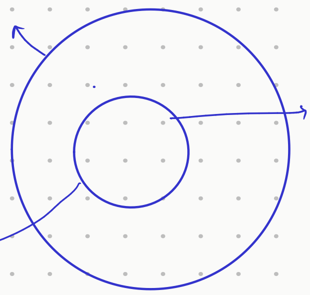
→ A random variable value depends on the outcome of a random phenomenon.

Population & Sample :-

→ Estimating the average height of human

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

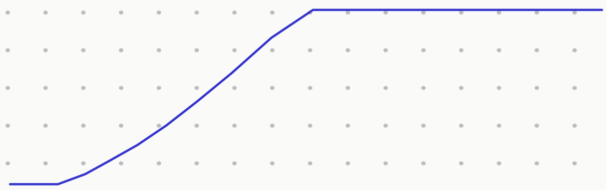
So we estimate·



often represented by $\bar{x}$ ← $\bar{x} = \frac{1}{1000} \sum_{i=1}^{1000} h_i$s

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

Sample mean ← $\bar{x}$; 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

} There 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

why learn?

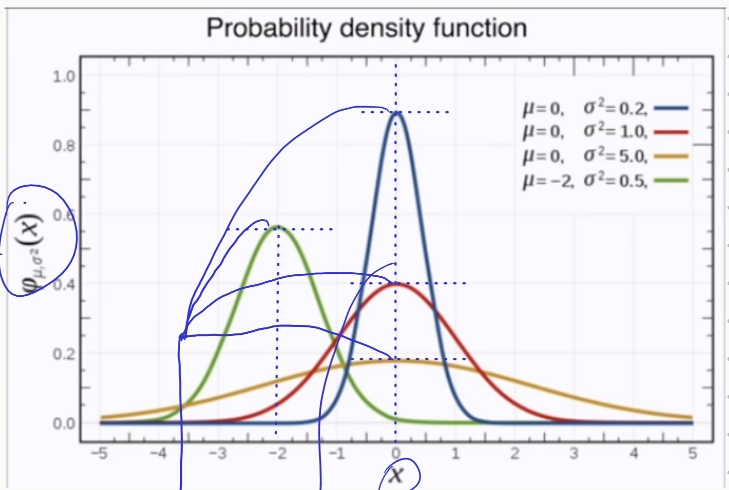
→ Stuff in nature tends to follow G.D.

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

→ They are simple models that summarize R.V.

ex:-

PDF



μ = mean, σ^2 = variance } → parameters

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

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

variance is a measure of spread.

Continuous Random Variable

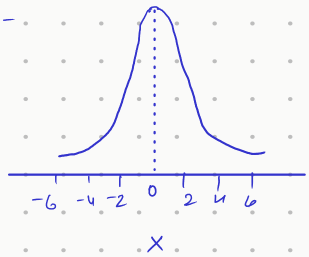
Red, Yellow, Blue have $\mu = 0$ but varying variances.

The peak of curve is usually at 'μ'.

→ 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:-



$\longrightarrow X \sim N(0, 2)$

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

↑

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

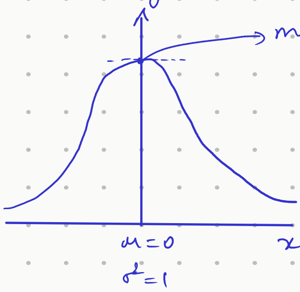
$y = \exp(-x^2)$

↳ when plotted

as x increases
y decreases.
as x decreases
y decreases

$y = P(x)$

→ max. value $= \frac{1}{\sqrt{2\pi}}$



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})

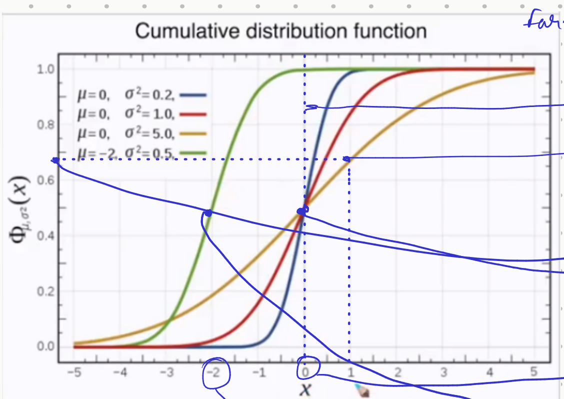
→ If $p(x)$ is the probability density at a point 'x', the probability can be obtained by computing the integral of $p(x)$ over a given interval.

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

Cumulative Distribution Function (CDF) of Gaussian Distribution/Normal Distribution :-

As σ^2 increases, CDF goes far from center line

CDF of a random variable looks like



→ $P(X \leq 1) = 0.65$

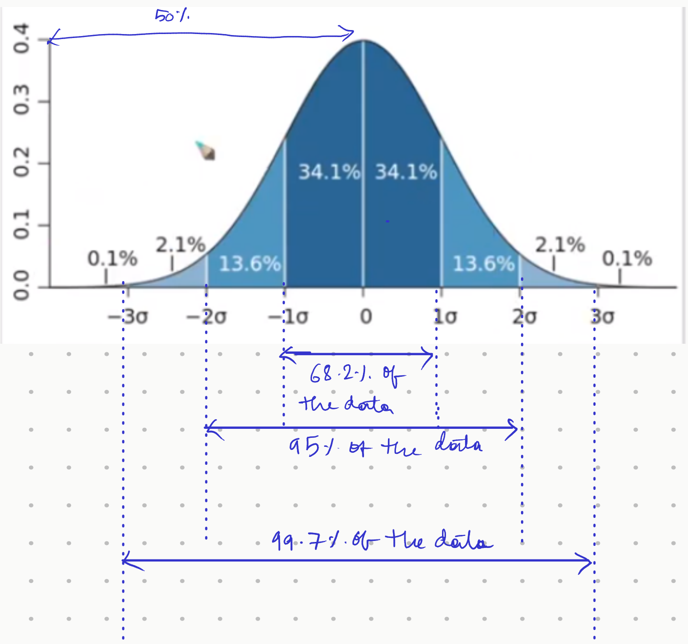
→ $\mu = 0$, center of CDF is at 0.

→ $\mu = -2$, center of CDF is at -2

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

68 – 95 – 99.7 rule :-

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



. How is this useful?

eg :- if human population height

$X \sim N(150, 25)$, where $150 = \mu$ and $25 = \sigma$

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

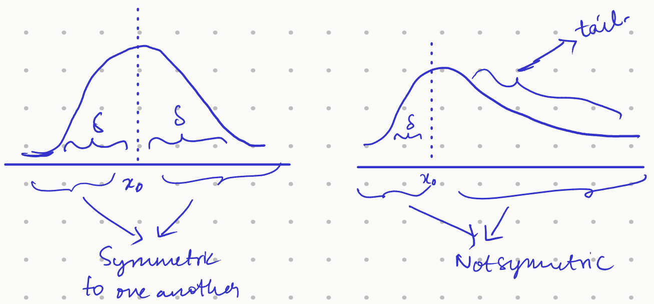
95% of people $(150-50,\ 150+50)$

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'