Probability & Statistics :-

Random variable :-

ex:- rolling dice → 6 sides = $\underbrace{\{1,2,3,4,5,6\}}_{\text{any one of these}}$ } Random Experiment

when rolled

equal outcome

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

tossing a coin → $Y = \{H, T\}$ → Sample Space.

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

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

(probability of X being even)

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

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

$P(x = x_1) \longrightarrow P(x_1)$ 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

Oultiers :-

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 disrete value is obtained by counting

→ A continuous value is obtained by measuring.

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

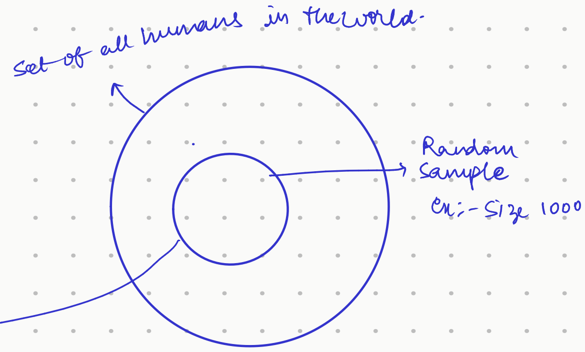
→ A random variable value depends on the outcome of a random phenomenon.

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

→ 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)

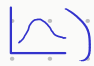
100 bikes + 150 cars.

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

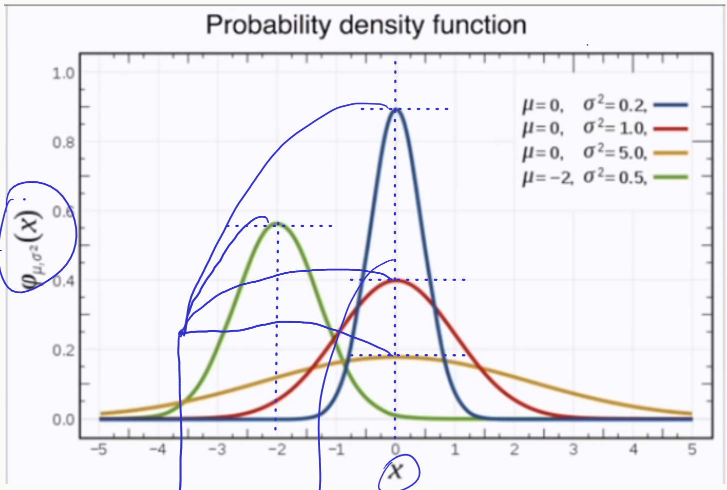
why learn?

→ Stuff in nature tends to follow G.D.

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

→ They are simple models that summarise R.V.

ex:-



PDF

x → Continuous Random Variable

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

The peak of curve is usually at 'μ'.

μ = mean

σ^2 = variance

} → parameters

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

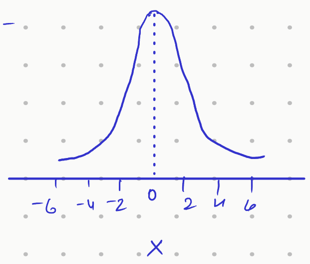
[if this is not known, then we can't]

Variance is a measure of spread.

→ 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:-



→ $X \sim N(0,2)$

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

↑

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 $\mu = 0$, $\sigma^2 = 1$, $\sigma = 1$

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

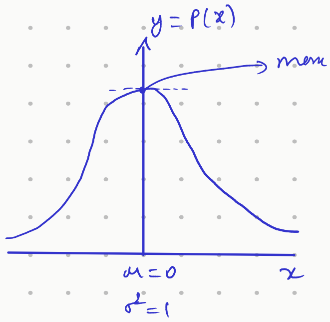
constants

further simplifying

$$\boxed{y = \exp(-x^2)}$$

↳ when plotted

as x increases
y decreases.
as x decrease
y decreases



→ 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})

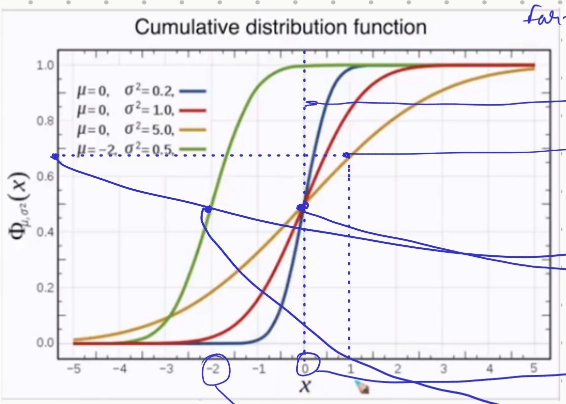
→ If p(x) is the probability density at a point 'x', the probability can be obtained by computing the integral of p(x) over a given interval.

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

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

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

CDF of a random variable looks like 


Cumulative distribution function

→ $P(X \leq 1) = 0.65$

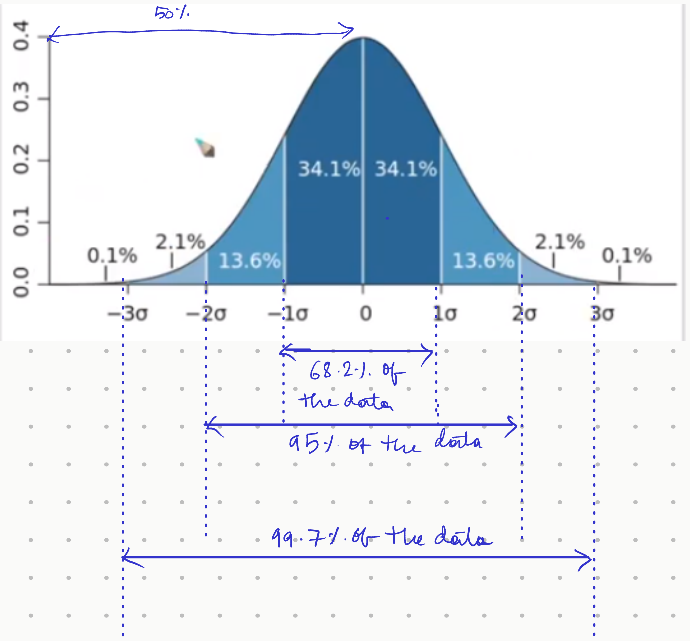
→ $\mu = 0$, center of CDF is at 0.

→ $\mu = -2$, center of CDF is at -2

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

68 – 95 – 99.7 rule :-

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



How is this useful?

eg:- if human population height

$$X \sim N(\underset{\mu}{150}, \underset{\sigma}{25})$$

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

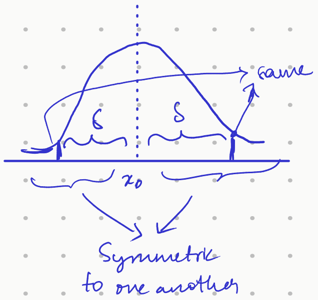
95% of people (150 − 50, 150 + 50)

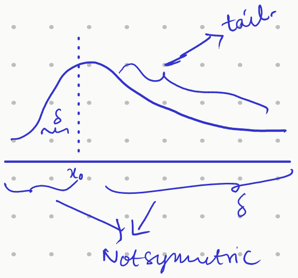
99.7% of people (150 − 75, 150 + 75)

→ 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

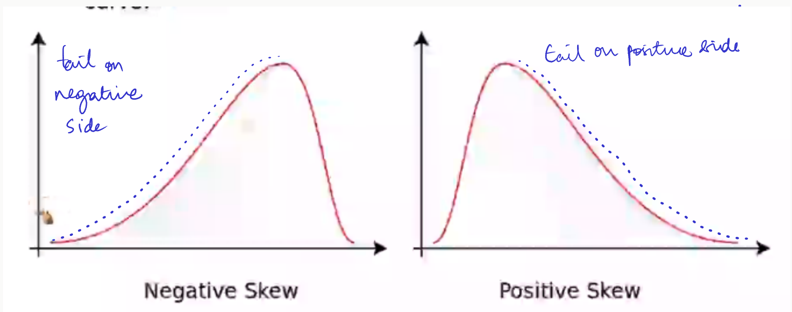
$f(x_0 - \delta) = f(x_0 + \delta)$ for all real numbers δ

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

Skewness :-

Skewness is a measure of asymmetry.

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



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

→ if 2 = variance

→ sample std-deviation

Kurtosis :-

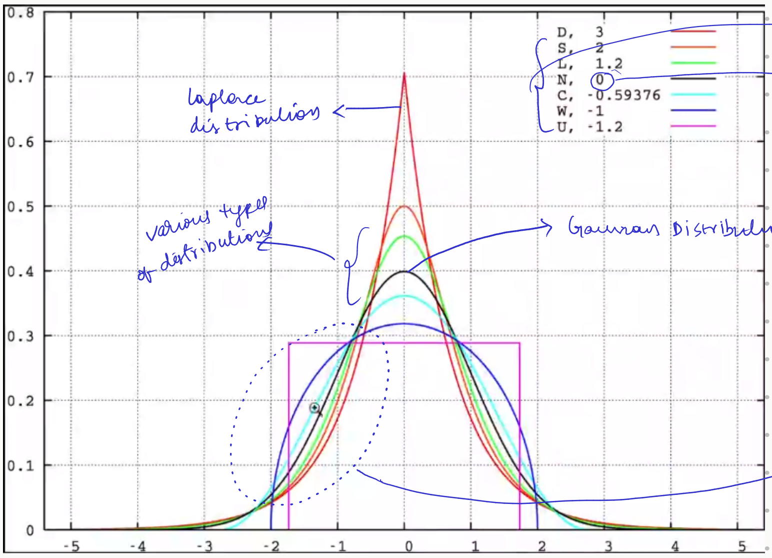
Kurtosis of Gaussian Distribution = 3

→ 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 parts

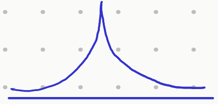
→ If there are some stocks. Some stocks fall ↓

some go up ↑

some go up enormously. ↑

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

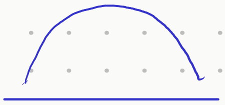
If kurtosis is > 3 ⇝



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

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

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

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

$\uparrow$ Standard Normal Variate.

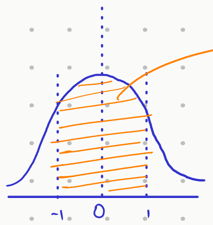
Given any random variable X, where $X \sim N(\mu, \sigma^2)$

$Z = \dfrac{X - \mu}{\sigma}$ } $\longrightarrow$ 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



$\rightarrow$ 68.2% of the data

($\because$ 68-95-99.7 rule)

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

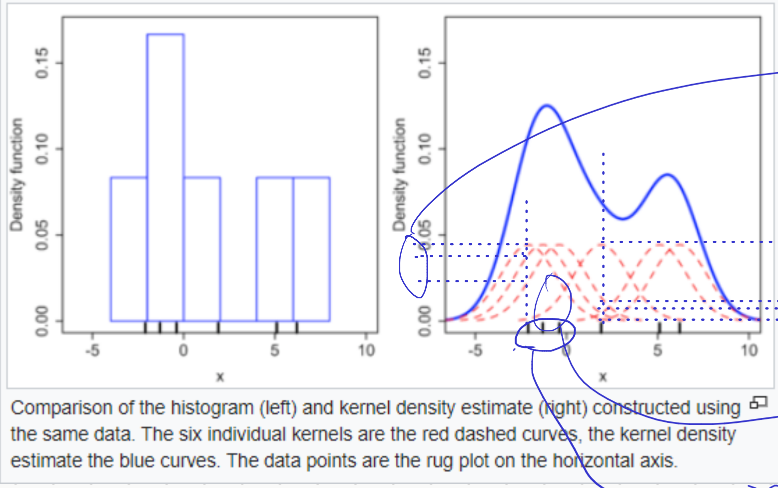
$\rightarrow$ We use StandardScalar for standardization & MinMaxScaler for Normalization.

Two techniques of feature scaling.

Normalization $\longrightarrow$ $x_i = \dfrac{x_i - x_{min}}{(x_{max} - x_{min})}$

Kernel Density Estimation :-

$\rightarrow$ Used for smoothing histograms to obtain PDFs.



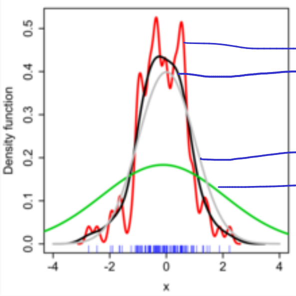
Comparison of the histogram (left) and kernel density estimate (right) constructed using the same data. The six individual kernels are the red dashed curves, the kernel density estimate the blue curves. The data points are the rug plot on the horizontal axis.

$\rightarrow$ give height at that point

$\rightarrow$ Variance of the guassian Kernel is known as Bandwidth.

$\rightarrow$ Gaussian Kernel.

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



$\rightarrow$ Narrow Bandwidth

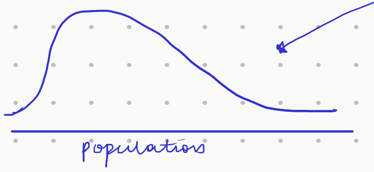
$\rightarrow$ Medium Bandwidth

$\rightarrow$ Actual Distribution.

$\rightarrow$ Wide Bandwidth

Sampling Distribution & Central limit Theorem :-

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



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

〃 S_2 〃 〃 〃 → mean $= \bar{x}_2$

〃 〃

〃 〃

→ mean $= \bar{x}_n$

These will also have a distribution

The distribution $\bar{x}_i$ = Sampling distribution 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 distribution is called sampling distribution of sampling mean, central limit theorem states that

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

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

mean is same as original as that of distribution

Gaussian Distribution

Why? → By using just m×n datapoints, we are able to estimate μ & σ^2 of any distribution, 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 → more powerful.

Q-Q-Plot → graphical method

How? :- ① Sort X & calculate percentiles.

$x_1', x_2', x_3', \ldots\ldots, x_{500}'$

$x_5' \longrightarrow$ 1st percentile $\longrightarrow x^{(1)}$

$x_{10}' \longrightarrow$ 2nd 〃 $\longrightarrow x^{(2)}$

⋮

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

② take $Y \sim N(0,1)$

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

↓

$y_1', y_2', y_3', \ldots\ldots y_{1000}'$

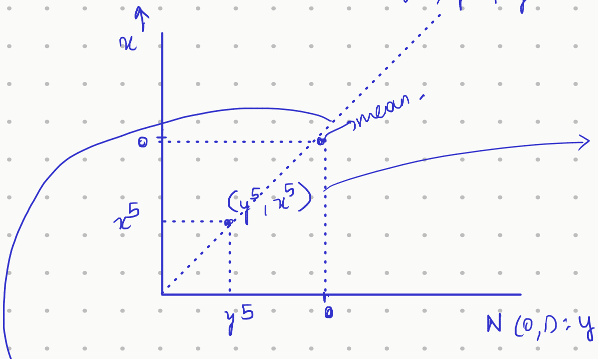
↓

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

} These are called as Theoretical Quantiles

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

$y^{(1)}, y^{(2)}, y^{(3)}$



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) ∀ $i \rightarrow 100$, lie roughly on a straight line, they y & x have similar distributions.

⇒ X also has gaussian distribution.

$\boxed{\text{→ Stats.probplot( )}}$

→ 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, are X & Y have the same distribution.