Pg No :-3

What we will Learn

3 main important thing

1. Linear Algebra
2. Statistics (Basics to advanced)
3. Different Calculas

* Applications of these 3 Topics in Data Science

1 Linear Algebra :- Scalar, Vectors, Vectors operations, matrics, matrix operations, functions, Linear Transformations, Inverse functions, Eigen Values and Eigen vectors

Neural Network :- Forward propogation → matrix operations

Applications in Data Science

2 Statistics :- → ML, Deep Learning → models => Huge Dataset

→ Tools to learn from the data

Statistics

Descriptive	Inferential
① Measure of central tendency	① Hypothesis Testing & P value
② Measure of Dispersion	② Z-test, t-test
③ Histograms, Box plot	③ Square Test
④ Types of distributions of data	④ ANNOVA Test
⑤ PDF, PMF, Normal Distribution, Log Nor	

② and ③ and ④ → Multiple problems

③ Differential Calculus :-

① Derivatives, Slope => Visual Diagrams => Deriving equations.

② Target lines

③ Polynomial Expression [Derivative of this Expression]

④ Trignometric Expression

⑤ Chain rule of derivative

⑥ Composite function

} optimizations => Chain rule

Applications of linear Algebra, Statistics & Differential Calculus in Data Science

① Simple Linear Regression, Multipule Linear Regression <= applications

② Dimensionality Reduction [principal Component Analysis] → Eigen values, Eigen Vector }

③ Neural N/w Is Framed → ANN → Multi-Layend NN }

↓

Artificial Neural Network

Linear Algebra :-

Linear algebra is a branch of mathematics that focuses on the study of Vectors, Vector Spaces [linear Spaces], linear transformations and Systems of linear equations. It provides a framework for understanding the properties and operations of these mathematical objects, which can be represented using matrics and vectors.

* It is a foundational Concepts for ML, DL, NLP & Images

* It has more applications in physis & mathematics.

* The aim is to Learn linear algebra as Computer Science Student [CSS]

Application of Linear Algebra :-

(1) Data representation & manipulation.

Dataset : House price dataset

Create a model which will be able to predict house price

Vector :- Is a list of numbers

Data is represented to model as Vectors

I/p feature

Area | no of Room's | Location | Price ← o/p features

[1200 | 2 | Bangalore | 45 lakhs INR]

* The model can Quantify the relationships like x, y

x↑ y↑
x↓ y↑
x↑ y↓
x↓ y↓

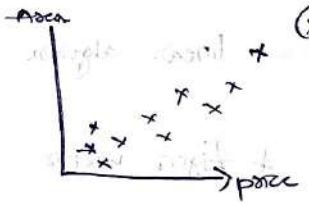
} covered in stats

↓

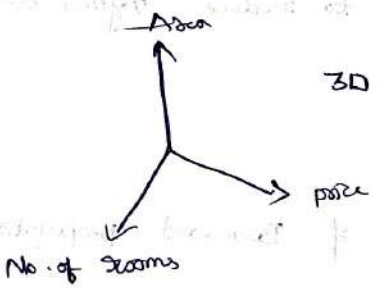
Covariance

Coordation

$\vec{v}$ [1200 45]



(x, y) => (1200, 45 lakhs) => 2D



3D

* Linear algebra works well with higher dimension data

500 dimension → PCA

↓↓ using LA concepts like Dimensionality reduction.

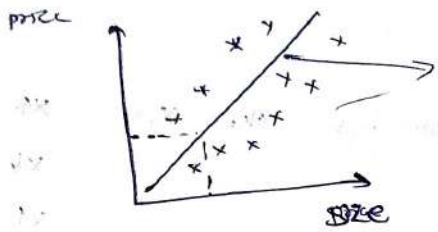
2 dimensions

(2) Machine Learning and artificial Intelligence :

* Model Train :- Uses Matrix Arithmatic operations, linear equations. etc.

to train model

Equation of a straight line $\Rightarrow ax+by+c=0 \quad \Rightarrow y=mx+c$



linear regression Create a best fit line

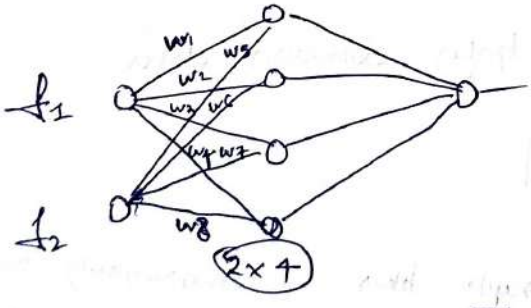
Input size & output price

* Dimensionality Reduction :- PCA uses - linear algebra concepts likes -

Eigen value & Eigen vector

With this help we will be able to reduce higher dimension to Lower dimension.

* Neural N/w :- Feaward propogation & Backward propogation



Train neural network with help of GPU → cools, GPU base fast because parallel MO & MM

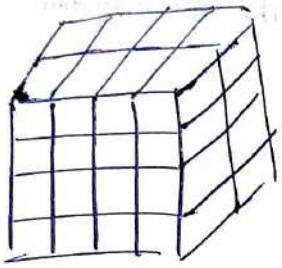
Tensorflow → Tensor ↗ works

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \begin{bmatrix} w_1 & w_2 & w_3 & w_4 \\ w_5 & w_6 & w_7 & w_8 \end{bmatrix} \Rightarrow \text{Matrix multiplication}$$

Computer Graphics

Image

0 - 255 pixels



RGB image

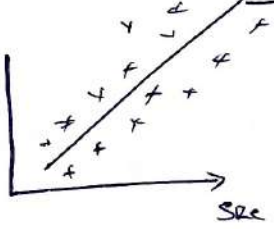
* Used to perform operations such as Scaling, rotating image, white & black image.

* Linear algebra is used to perform transformation such as scaling, rotating & translation of objects in Computer graphic

(4) Optimization

best fit line → finding right slope & intercept.

① Solving Equation :-



Error comes down is best fit line.

Gradient Descent => optimizer

(3) Scalars and Vectors

Scalar :- It is a single numerical value. It represents a magnitude or quantity and has no directions.

Ex :- (1) Car speed = 45 km/hr — Magnitude

(2) Temperature in Celsius T = 25° C — Magnitude

Applications :- [in DS]

Dataset :- 3 features + 5 specific values [Records]

f_1	f_2	f_3
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—

Count of the Total no of records = 5

average of feature f_1 = —

} Scalars Quantity

Simple linear regression => $y = mx + c$ (c: intercept)

m: slope

y: scalar value

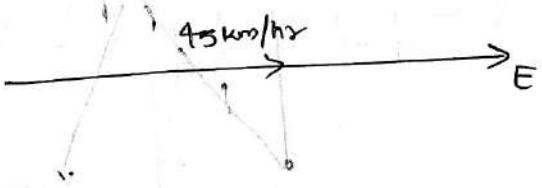
Vector :- It is a numerical value which has both magnitude & direction

In terms of Data Science,

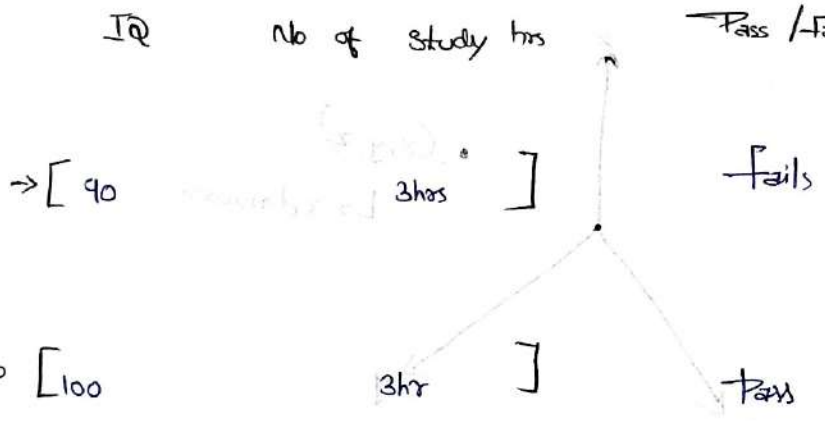
↓

A vector is a ordered list of number. It can represent a point in space or quantity with both magnitude and direction.

Example :- Speed of the car is 45 km/hr and is moving towards East Direction



Example :- Student marks.



→ A vector representing person IQ & no of study hrs = [90, 3hrs]

3 hrs ↑ units

↓

magnitude

A vector representing persons weight over time $[70, 72, 75, 73]$ <= 4d

No need to think they have specific directions

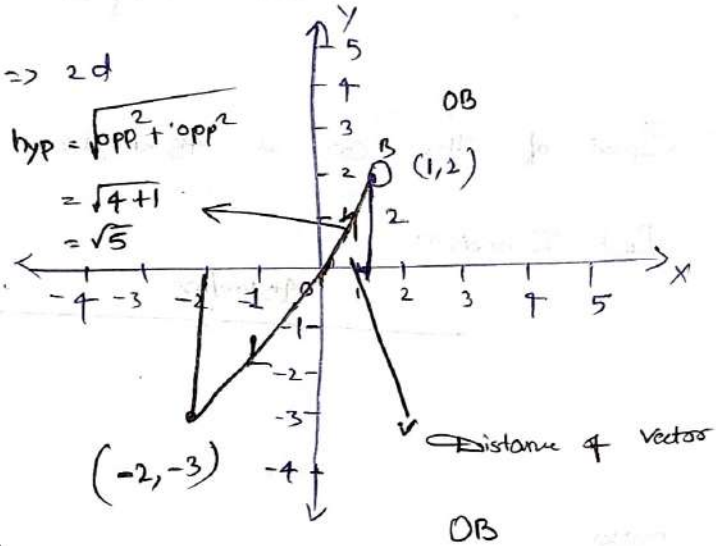
Vectors representing with respect to physics

$A = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ => 2d

$hyp = \sqrt{opp^2 + opp^2}$

$= \sqrt{4+1}$

$= \sqrt{5}$



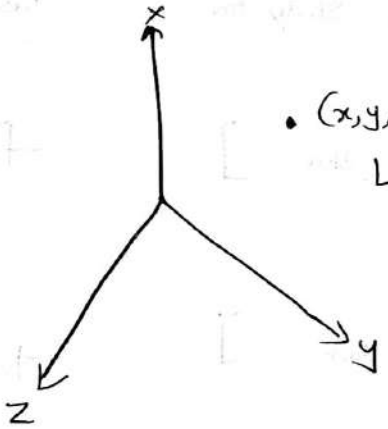
OB

Distance of vector from origin is

OB

$B = \begin{bmatrix} -2 \\ -3 \end{bmatrix}$

$C = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$



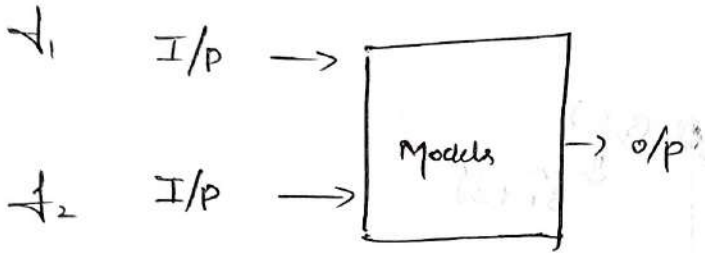
• (x, y, z)

↳ 3 dimension

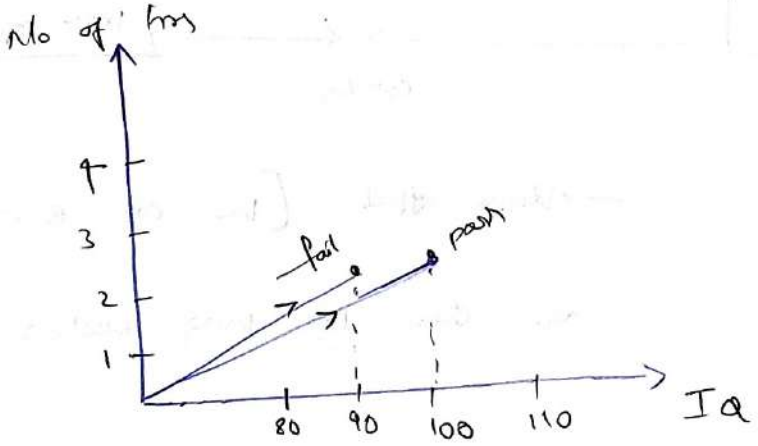
How vectors are related to Data Science. ?

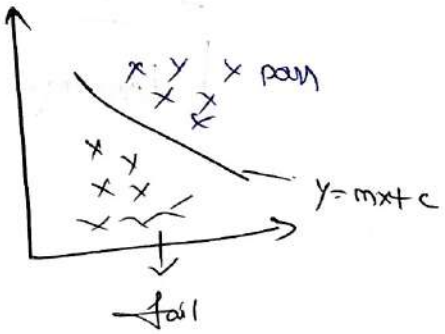
In DL, ML we Create Some kinds of models.

The aim of the model is to take input data & predict the output features



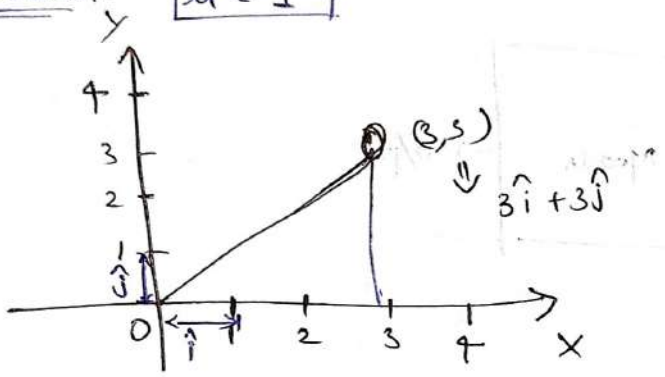
f_1	f_2	o/p ←
IQ	No. of hours	pass/fail
[90	2]	fail ⇒ 0
[100	2]	pass ⇒ 1





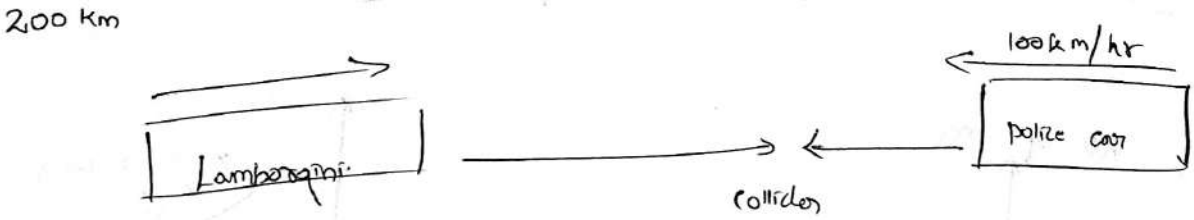
- With human eye we can see max to max 3 features but linear algebra is so amazing that [since it supports lots Matrices calculations] any number features can be represented using it.

Unit Vectors :- $\hat{u} = 1$



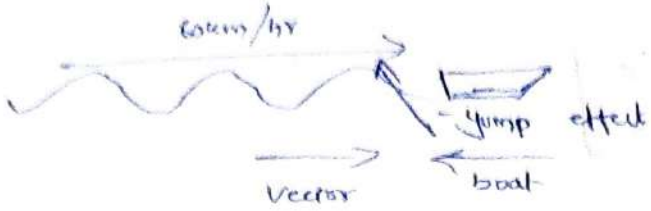
Examples of Vectors we used daily specially in gaming industry.

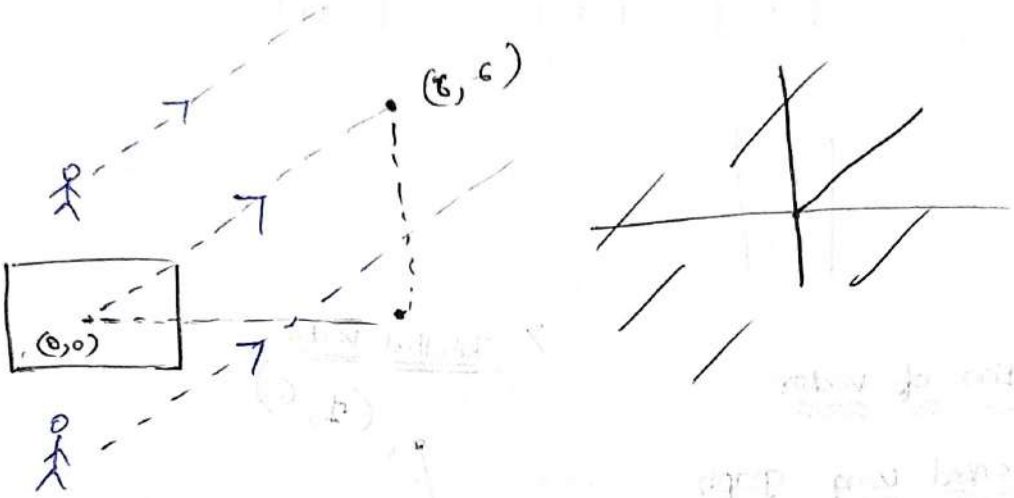
GTA 6 :-



→ Adverse effect [like car blown up, damages. etc] are done by using vectors

Boat





If (0,0) is origin or not that vector can move in same direction with different origins & it can happen in any part of Co-ordinate System

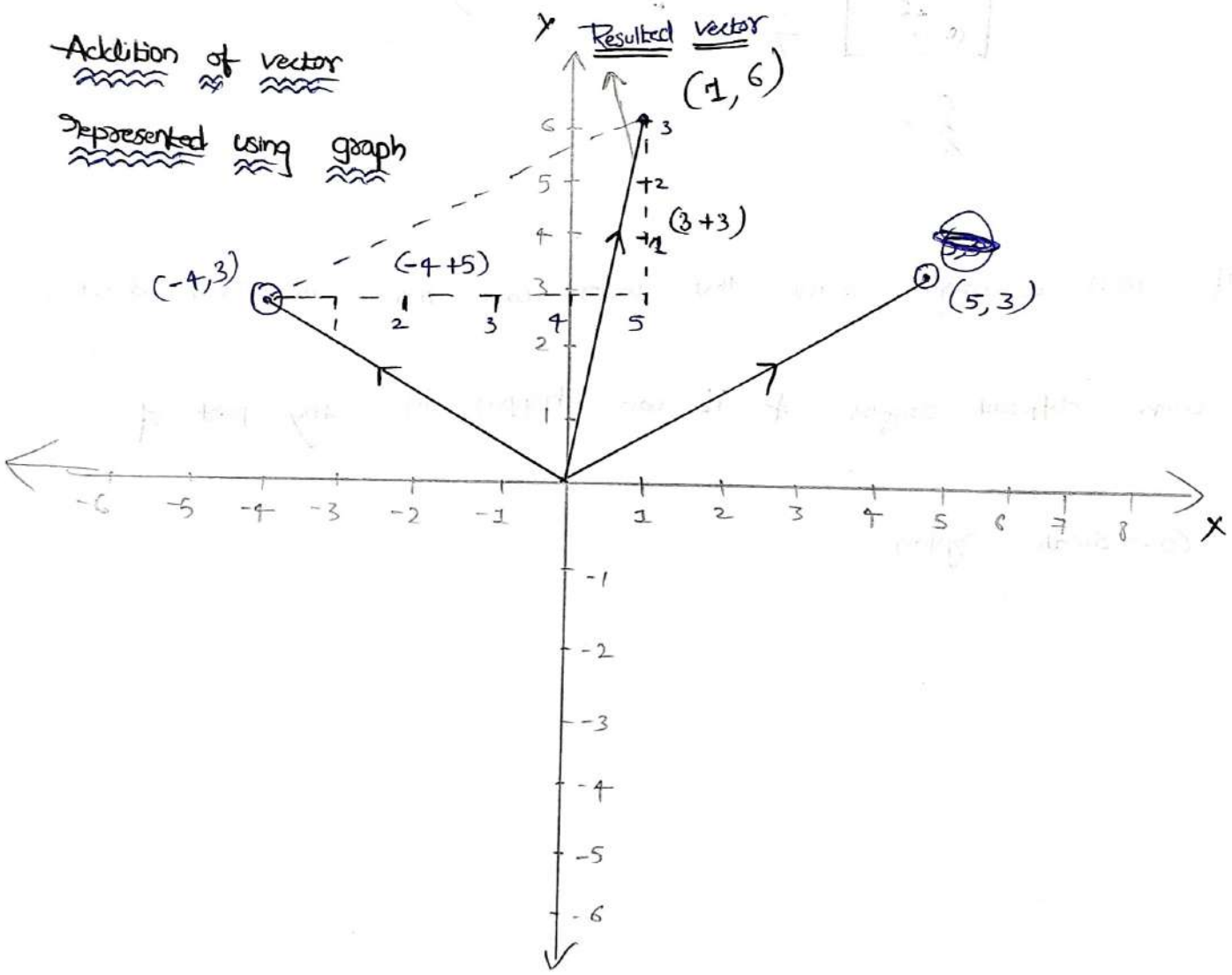
④ Addition of Vectors:

① $P_1 = \begin{bmatrix} -4 \\ 3 \end{bmatrix}$ $P_2 = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$

Sol $P_1 + P_2 = \begin{bmatrix} -4 \\ 3 \end{bmatrix} + \begin{bmatrix} 5 \\ 3 \end{bmatrix} = \begin{bmatrix} -4+5 \\ 3+3 \end{bmatrix}$

$= \begin{bmatrix} 1 \\ 6 \end{bmatrix}$

Addition of vector represented using graph



$$A = \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} \quad B = \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} \quad A+B = \begin{bmatrix} x_1+x_2 \\ y_1+y_2 \\ z_1+z_2 \end{bmatrix} = \begin{bmatrix} x_3 \\ y_3 \\ z_3 \end{bmatrix}$$

Examples:

Solving a usecase in EDA & feature engineering

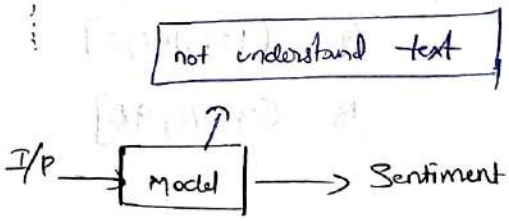
Sensor1	Sensor2	final sensor reading
[3, 5, 7]	[2, 4, 6]	[3, 5, 7] + [2, 4, 6] = [5, 9, 13]

Applications

1. Data aggregation
2. feature engineering
3. NLP
4. Word Embeddings
5. Image processing

NLP Ex: Ecommerce website

Reviews	Sentiments
The product is good	1
The product is bad	0



The text is converted to vector

text → Vector [numerical values]

OHE
TFIDF
BOW
Word2Vec

Word Embedding:

① Data : [0.2, 0.1, 0.4]

② Science : [0.3, 0.7, 0.2]

Data Science = Data + Science

= [0.2, 0.1, 0.4] + [0.3, 0.7, 0.2]

= [0.5, 0.8, 0.6] => Data Science

Image Processing

Color image [R, G, B] =

- Red channel R = [255, 128, 0]

G = [128, 156, 0]

B = [64, 78, 90]

RGB → Grey scale = R + G + B

$$= \frac{[255+128+64 \;,\; 128+156+78 \;,\; 0+0+90]}{3}$$

white + black

= [ ]

4 Multiplication of Vectors

3 Types

1. Dot product (Inner product)
2. Element wise multiplication
3. Scalar multiplication

1) Dot product:

Definition: The dot product of 2 vectors results in a scalar. It is calculated as the sum of the products of their corresponding components.

$$A = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \qquad B = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

$$A \cdot B = 2 \times 4 + 3 \times 5$$

$$= 8 + 15$$

$$= 23 = \text{Scalar}$$

$$A \cdot B^T = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 4 & 5 \end{bmatrix}$$

$$= 2 \times 4 + 3 * 5$$

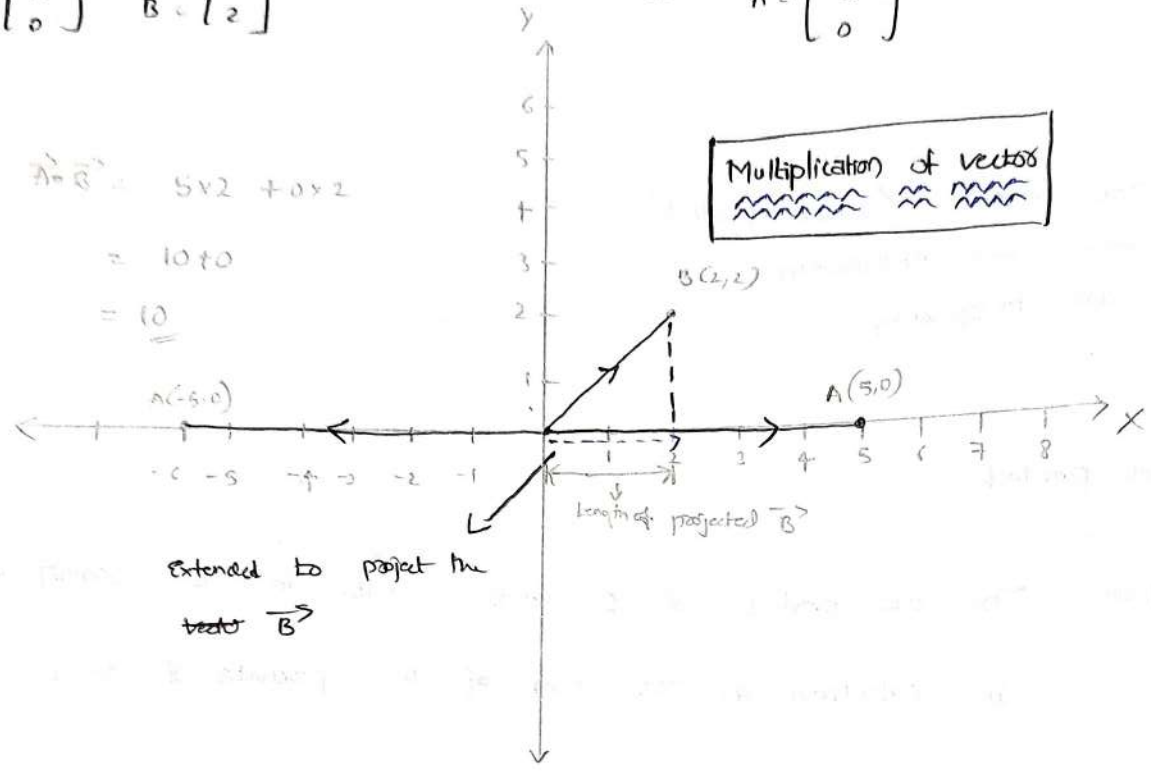
$$= 8 + 15$$

$$= 23 = \text{Scalar}$$

(1)

$A = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$ $B = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

(2) $A = \begin{bmatrix} -5 \\ 0 \end{bmatrix}$



If the length of projected vector is zero the dot product is zero

$\vec{A} \cdot \vec{B}$ = (Length of projected $\vec{B}$) . (length of the vector $\vec{A}$)

$= (2) \cdot (5) = 10$

$= +ve$

Note

$\vec{A} \cdot \vec{B} = 0$, project the vector to the origin

Application of DOT PRODUCT in Data Science:

-> Gen AI

-> App Systems like RAG

① Cosine Similarity:

It is a measure used to determine how similar 2 vectors are, it calculates cosine of the angle between 2 vectors and provide a similarity score that ranges from -1 (dissimilar) to 1 (complete similar)

$$\cos\theta = \frac{A \cdot B}{\|A\| \|B\|}$$

Recommendation System: Netflix account => Watching Action movie

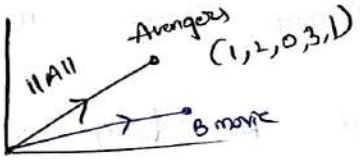
↳ Recommendation of other action movies

Avengers => [1, 2, 0, 3, 1] (Drama, story type)

(Action, Comedy, Romance)

B movie = [2, 0, 1, 1, 1]

Rough graph 5D imagination



Step1: Dot product of A·B, $A \cdot B = 1 \cdot 2 + 2 \cdot 0 + 0 \cdot 1 + 3 \cdot 1 + 1 \cdot 1 = 6$

Step2: $\|A\| = \sqrt{1^2 + 2^2 + 0^2 + 3^2 + 1^2} = \sqrt{15} = 3.872$

$\|B\| = \sqrt{2^2 + 0^2 + 1^2 + 1^2 + 1^2} = \sqrt{7} = 2.6457$

$$\cos\theta = \frac{6}{3.872 * 2.646} \simeq 0.586$$

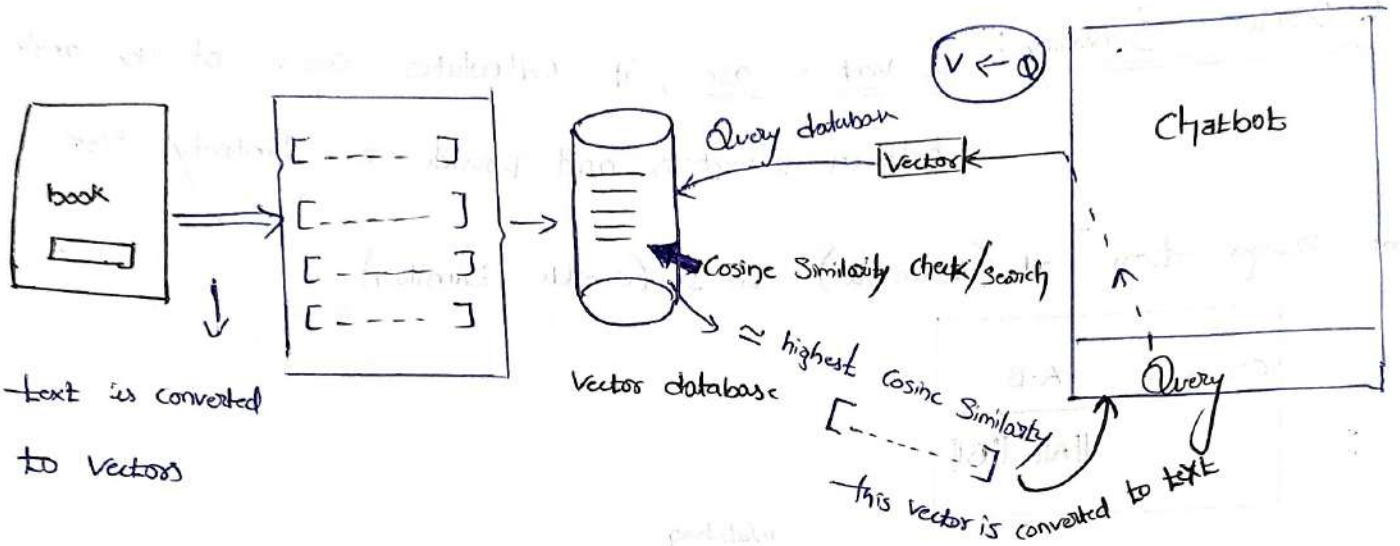
58.6% +ve similar

$$\|A\| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Vector Databases

-> Regularly used in GenAI & LLM models

=> Retrieval augmented generation (RAG) system uses Vector databases



(2) Element wise multiplication

In element wise multiplication, corresponding elements of 2 vectors are multipled to form a new vector of the same dimension

$$A = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad B = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$$

$$A \otimes B = \begin{bmatrix} 3 \\ 8 \\ 15 \end{bmatrix}$$

Application in Data Science

① Feature Engineering

Product	Cost		Discount		Discount price		Final price
A	1000	⊗	0.1	−	100	=	900
B	500		0.2		100		400
C	200		0.5		30		170

Deep learning:-

RNN , LSTM RNN , GRO RNN

⇓

⊗ ⊕ => Forget gate , input gate

$$\begin{bmatrix} 0.5 \\ 0.6 \\ 0.3 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 0 \\ 0 \\ 0.3 \end{bmatrix}$$

Gate => pass info or not

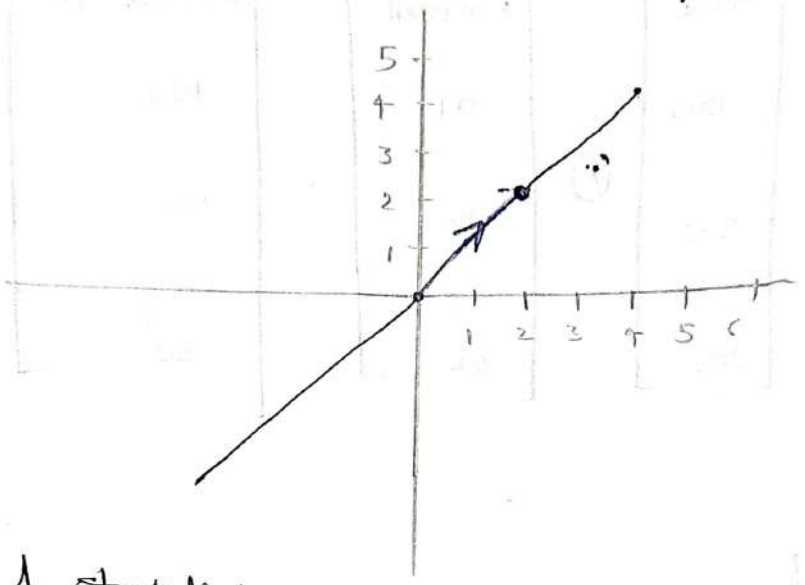
③ Scalar multiplication :-

It involves multiplying vector by a scalar, resulting in a vector where each component is scaled by the vector.

$A = \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}$ $c = 4$

$C = \begin{bmatrix} 12 \\ 20 \\ 28 \end{bmatrix}$

$A = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ $c = 2$

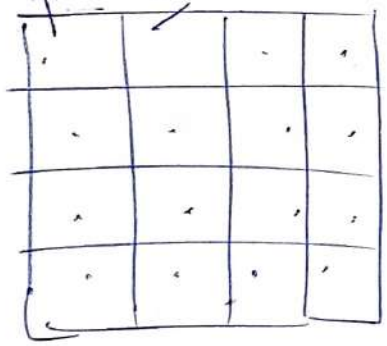


Ex: Normalization & Standardization

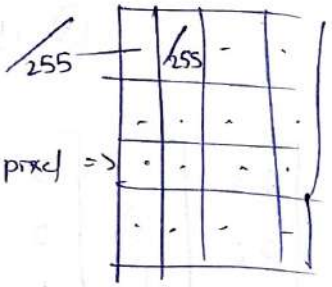
↓

Scaling data => units

0 - 255 0-255



=> Image processing => Normalize pixel => [0-1]



It require more calculation so normalization is done

Eg: feature

Height = [160, 170, 180]

Scale (to meter) = c = 0.01

1 cm = 0.01 m

Ch = 0.01 [160, 170, 180]

= [1.6, 1.7, 1.8]

Matrices: A matrix is a rectangular array of number, symbols or expressions arranged in rows and columns.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$$

Each element is represented by a_{ij}

where i denotes rows

j denotes columns

Example of matrices in Data Science:

Dataset

	Math Score	Physic Score	Biology Score
t_1	55	65	75
t_2	65	60	55
t_3	70	45	80

$$\begin{bmatrix} 55 & 65 & 75 \\ 65 & 60 & 55 \\ 70 & 45 & 80 \end{bmatrix}_{3 \times 3}$$

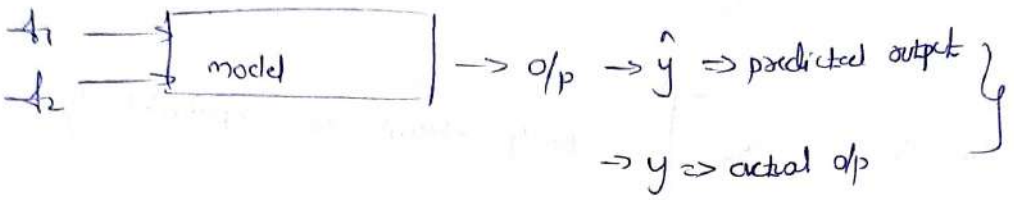
2) **Image in computer vision**

3 x 3 Grayscale image

0	128	255
255	128	0
128	255	128

$$\Rightarrow \begin{bmatrix} 0 & 128 & 255 \\ 255 & 128 & 0 \\ 128 & 255 & 128 \end{bmatrix}_{3 \times 3}$$

③ Confusion matrix : Accuracy of the model



$$\text{Confusion matrix} = \begin{bmatrix} 50 & 10 \\ 5 & 35 \end{bmatrix}$$

50 → True positive

10 → False Negative

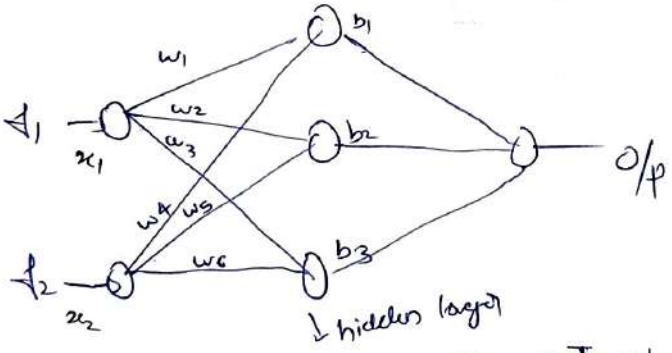
5 → False position

35 → True negative

$$\frac{TP + TN}{TP + FN + FP + TN} = \text{Accuracy}$$

④ Neural N/w :

Matrix operation :- → used in N/w & Linear regressions



Forward propagation

$$Z = W^T x + b$$

$$W = \begin{bmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \end{bmatrix}$$

$$x = [x_1, x_2]$$

$$b = [b_1, b_2, b_3]$$

Linear Regression:

x_1 no of study hours	x_2 IQ	o/p feature Score dependent feature
4	100	90
5	90	85
—	—	—

Regression

$y = mx + c$

$= m_1x_1 + m_2x_2 + c$

m_1 slope, m_2 slope or co-efficient

$\Rightarrow m^T x + c$ $\left\{ m = [m_1, m_2],\ x = [x_1, x_2] \right\}$

NLP :- Dataset

Review	Sentiment +ve/-ve
The food is good	0
The food is bad	1

$$\begin{bmatrix} [0.1 & 0.2 & 0.3] \\ [0.4 & 0.5 & 0.6] \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

we can apply different ML algorithms to train this data

Matrix operation:

To manipulate & analyze multidimension data efficiently:

1. matrix Addition & Substraction
2. Scalar Matrix Multiplication
3. Matrix Multiplication.

(1) ~~Matrix operations:~~

~~=> To manipulate & analyze mu~~

(1) Matrix Addition and Substraction:

Add or Substract corresponding elements of 2 matrices of the same dimensions

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \Rightarrow \text{STORE A} \qquad B = \begin{bmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 1 & 2 & 3 \end{bmatrix} \text{ STORE B}$$

3x3 3x3

	Product A	Product B	Product C
Day1	1	2	3
Day2	4	5	6
Day3	7	8	9

$$A + B = \begin{bmatrix} 1+4 & 2+5 & 3+6 \\ 4+7 & 5+8 & 6+9 \\ 7+1 & 8+2 & 9+3 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & 7 & 9 \\ 11 & 13 & 15 \\ 8 & 10 & 12 \end{bmatrix}$$

② Scalar multiplication:

Scalar multiplication involves multiplying every element of a matrix by a scalar value

$$B = cA$$

Eg:- Suppose we have a matrix representing product prices in dollars and we want to adjust these prices for inflation by a factor of 1.05

original $P = \begin{bmatrix} 10 & 20 & 30 \\ 15 & 25 & 35 \\ 20 & 30 & 40 \end{bmatrix}$

Scalar multiplication:

$$P_{adjusted} = 1.05 \cdot P = 1.05 \begin{bmatrix} 10 & 20 & 30 \\ 15 & 25 & 35 \\ 20 & 30 & 40 \end{bmatrix} = \begin{bmatrix} 10.5 & 21 & 31.5 \\ 15.75 & 26.25 & 36.75 \\ 21 & 31.5 & 42 \end{bmatrix}$$

Used for adjusting salary after inflation:-

s/w eng	HR	Accountant	60%
45k	30k	40K	100%
50k	35k	45k	
-	-	-	
-	-		

③ Matrix multiplication

Operation: It involves the dot product of rows of the first matrix with columns of the second matrix

For 2 matrix $A(m \times n)$ & $B(n \times p)$, the result matrix is $C(m \times p)$

$$\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \Rightarrow 1\times2 + 2\times3 + 3\times4 = 2+6+12 = [20]$$

$$A = \underset{2\times3}{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}} \qquad B = \underset{2\times3}{\begin{bmatrix} 7 & 9 & 11 \\ 8 & 10 & 12 \end{bmatrix}}$$

$m \times n$ $\qquad$ $n \times p$

$$B^T = \underset{3\times2}{\begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{bmatrix}}$$

$$C = A \cdot B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} 58 & 64 \\ 139 & 154 \end{bmatrix}$$

$c_{11} = (1\times7) + (2\times9) + (3\times11) = 58$

$c_{12} = (1\times8) + (2\times10) + (3\times12) = 64$

$c_{21} = (4\times7) + (5\times9) + (6\times11) = 139$

$c_{22} = (4\times8) + (5\times10) + 6\times12 = 154$

Functions And Linear Transformations.

Functions: A function is a mathematical relationship that uniquely associates elements of one set (called the domain) with elements of another set (called the codomain). In simpler terms, a function maps inputs to outputs in a specific way

Notation:- A function f mapping elements from set x (domain) to set y (codomain) is denoted by $f: x \to y$

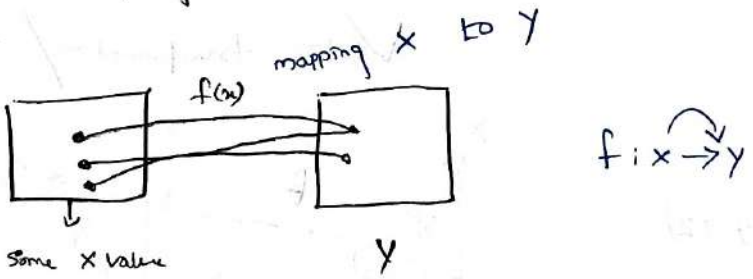
If x is an element of x, then $f(x)$ is the corresponding element in y.

Example:- $f(x) = 2x+3$ ⇒ maps each real number x to a real number

$x = 5$ $\quad x \xrightarrow{f} y$

$f(5) = 2\times5+3 = 13$

$f(x) \to$ mapping $5 \in R$ to $13 \in R$



$f: x \to y$

Vector

$$f: \begin{bmatrix} x \\ y \\ z \end{bmatrix} \in R^3 \to \in 3D \to \begin{bmatrix} x+y \\ 6z \end{bmatrix} \in R^2 \qquad f(x) = \begin{bmatrix} x+y \\ 6z \end{bmatrix}$$

$f(x)$

It mapped 3D vector to 2D vector

Eg: Dimensionality Reduction => we uses functions for Transformation to reduce the dimensions

Domain ↑ Co Domain ↑

$f: R^3 \rightarrow R^2$

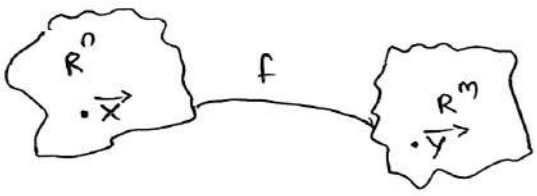
f is function which is mapping 3D real number to 2D real number

Vector Transformation:

$f: x \rightarrow y$

Both may have different set of vectors

$\vec{x} \rightarrow \vec{y}$



can be denoted as

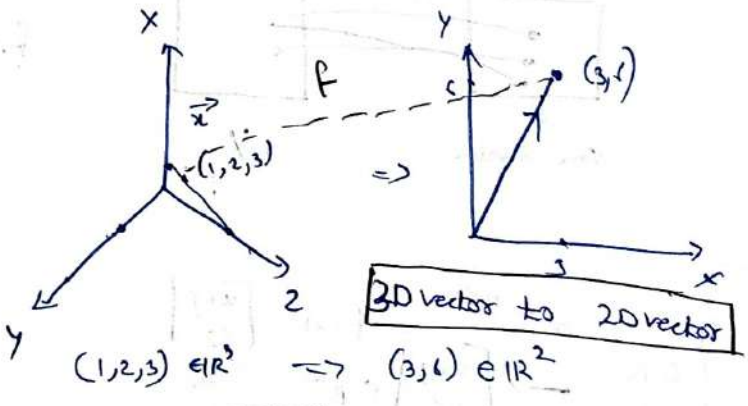
$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \quad x_1, x_2, x_3 \ldots x_n \in R$$

$f: \mathbb{R}^n \rightarrow \mathbb{R}^m$

Vector transformation / T in LA

$f(x,y,z) = (x+y, 2z)$

$f\left(\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}\right) = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$



$(1,2,3) \in \mathbb{R}^3 \Rightarrow (3,6) \in \mathbb{R}^2$

$T: \mathbb{R}^3 \rightarrow \mathbb{R}^2$

Definition:- Vector transformations refers to operations that maps vectors from one space to another, often changing their magnitude, direction, or both. These transformations are typically described using matrices & are fundamental in various fields, including physics, engineering, computer graphics, and data science

In gaming industry this kind of Transformations are heavyly used.

① Ex: Scaling:-

Scaling is a transformation that changes the magnitude of vector while keeping their direction same

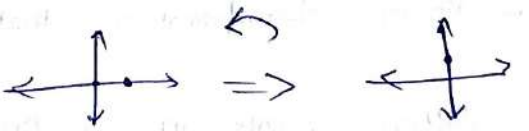
Scaling matrix = $\begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix}$

Scaling is used for Data normalization & Computer graphics to resize the image.

=> open paint → image → Resize

② Rotation:- Transformation that turns vectors around the origin

$V = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in R^2$



$V' = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ => Showing a 90 degree rotation

Used Deep learning
image processing
Robotics

Ex:- Rotation will be used in image processing => Rotating image,

Robotics => adjusting robot orientation

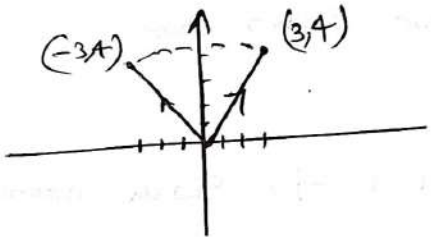
3D graphics => Rotating objects

③ Reflection [vector T is used in it]

Transformation that flips vectors over a specific axis or plane.

$V = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ => Across the y axis

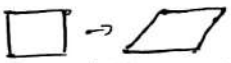
$$\begin{bmatrix} 3 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -3 \\ 4 \end{bmatrix}$$



Uses

Mirror images

Analyzing wave reflections

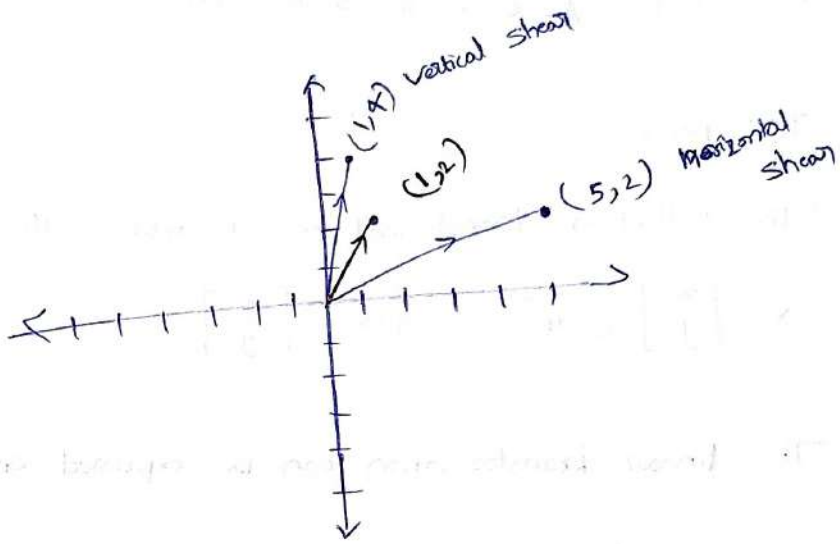
④ ~~Stretch~~: Shearing 

Shearing is a linear transformation that skews the shape of an object by shifting points in a one direction, proportional to their position on another axis. It change shape without necessarily changing area or volume

$V = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ $\quad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\quad VI = V$

Horizontal Shear: $\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$ $\quad k=2 \Rightarrow \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1+4 \\ 0+2 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$

Vertical Shear: $\begin{bmatrix} 1 & 0 \\ k & 1 \end{bmatrix}$ $\quad \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$



Uses:

cx

1. Computer graphics & Animation
2. ML & Data Augmentation
3. Structural Engineering

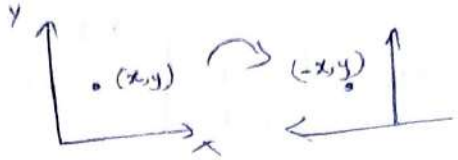
* Linear Transformation:-

A linear transformation is a function between two vector spaces that preserves the operations of vector addition and scalar multiplication. This means that if T is a linear transformation from a vector space V to a vector space W, then for any vectors

2 Important properties of LT

① Additivity $T(u+v) = T(u) + T(v)$

② Homogeneity $T(cu) = cT(u)$



$T: V \to W$ → Linear Transformation

$u, v \in V$ & c is a Scalar Value

Ex: Reflection :

The reflection transformation T across the y-axis maps a vector

$$x = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2 \qquad T(x) = \begin{bmatrix} -x \\ y \end{bmatrix}$$

The linear transformation can be expressed as → $\boxed{T(x) = AX}$

$$A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \underset{1\times 2}{[x \; y]} \underset{2\times 2}{\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}} = \underset{1\times 2}{\begin{bmatrix} -x \\ y \end{bmatrix}}$$

① Checking additivity :

Let $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ be two vector in R^2

$$T(u+v) = T(u) + T(v)$$

$$u+v = \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \end{bmatrix}$$

$$T(u+v) = A(u+v)$$

$$T(u+v) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \end{bmatrix} = \begin{bmatrix} -(u_1+v_1) \\ u_2+v_2 \end{bmatrix} \quad -(1)$$

R.H.S $\quad T(u) + T(v)$

$$T(u) = Au = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} -u_1 \\ u_2 \end{bmatrix}$$

$$T(v) = Av = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} -v_1 \\ v_2 \end{bmatrix}$$

$$R.H.S = T(u) + T(v) = \begin{bmatrix} -u_1 - v_1 \\ u_2 + v_2 \end{bmatrix} \quad -(2)$$

L.H.S = R.H.S ✓

(2) checking homogenity :-

Let $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \in R^2$ & c be a scalar

Homogenity requirement ;

$$T(cu) = cT(u)$$

L.H.S $\quad T(cu) = A(cu)$

$$= A\left(c \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}\right)$$

$$= A \begin{bmatrix} cu_1 \\ cu_2 \end{bmatrix}$$

$$= \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} cu_1 \\ cu_2 \end{bmatrix} = \begin{bmatrix} -cu_1 \\ cu_2 \end{bmatrix} \quad -0$$

R.H.S

$$cT(u) = c(A \cdot u)$$

$$= c\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = c\begin{bmatrix} -u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} -cu_1 \\ cu_2 \end{bmatrix} \quad \text{— (2)}$$

L.H.S = R.H.S ⊖

It is linear transformation.

Example that does not follow linear transformation

$$T: R^2 \to R^2$$
$$U \quad V$$

$$b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$T(x) = x + b$$

↓

fixed vector

$$T(x) = x + \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

① Check Additivity

$$T(u+v) = T(u) + T(v)$$

$$u = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \qquad v = \begin{bmatrix} 4 \\ -1 \end{bmatrix}$$

$$T(u+v) = T\left(\underset{2\times1}{\begin{bmatrix} 2 \\ 3 \end{bmatrix}}\ \underset{2\times1}{\begin{bmatrix} 4 \\ -1 \end{bmatrix}}\right) = T\left(\begin{bmatrix} 6 \\ 2 \end{bmatrix}\right)$$

$= \begin{bmatrix} 6 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 \\ 3 \end{bmatrix}$ $\Rightarrow$ R.H.S

L.H.S

$T(u) + T(v)$

$T(u) = T\left(\begin{bmatrix} 2 \\ 3 \end{bmatrix}\right) = \begin{bmatrix} 2 \\ 3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$

$T(v) = T\left(\begin{bmatrix} 4 \\ -1 \end{bmatrix}\right) = \begin{bmatrix} 4 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$

$T(u) + T(v) = \begin{bmatrix} 8 \\ 4 \end{bmatrix}$ $\Rightarrow$ L.H.S

R.H.S $\neq$ L.H.S

Checking Homogeneity:-

$T(cu) = c\,T(u)$

$u = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ $c = 2$

$T(cu) = T\begin{bmatrix} 4 \\ 6 \end{bmatrix} = \begin{bmatrix} 4 \\ 6 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}$ L.H.S

$\neq$

$cT(u) = 2\left(\begin{bmatrix} 2 \\ 3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) = 2\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \end{bmatrix}$ R.H.S

$T(x) = x + 1$ is not a linear transformation

fails both additivity & homogeneity properties.

W/by Linear Transformation:-

Linear transformations are fundamental in data science for several reasons they provide a mathematical framework for manipulating & analyzing data, which is crucial for various data processing tasks, model building & interpretation.

Here are the some key reasons why linear transformations are important in data science.

① Dimensionality Reduction:-

Principal Component Analysis (PCA):

PCA is a widely used technique for reducing the dimensionality of datasets while retaining as much variance as possible it involves finding a set of orthogonal axes (principal components) & projecting the data onto these axes. The transformation of data point in the original space to the new space defined by the principal components is a linear transformation. This helps in

While we are reducing dimension, we need to make sure that we capture a lot of variance [information that should not be lost & it talks about how data is spreaded

Reducing computational cost

Mitigating the curse of dimensions

Visualizing high-dimensional data

② Feature engineering

linear transformations can be used to create new features from existing ones. For example, interactions between features can be captured through linear combinations, which can be used in ML models to improve predictive performance.

Techniques like linear edge regression, ridge regression & linear discriminant analysis (LDA) all rely on linear transformations to find meaningful feature representations.

The fundamental step in feature engineering is Normalization & standardization

③ Data preprocessing :-

Normalization & Standardization :

linear transformations are used to scale data, making it suitable for ML models. Standardization ~~transformation~~ data to have a mean of zero & a standard deviation of one, while normalization scale data to specific range (eg.. [0,1]). These transformations are essential for ensuring that all features contribute equally to the model, especially in algorithms like gradient descent

④ Neural Networks

In Neural Network, especially deep learning model, the layers consist of linear transformation followed by non-linear activation functions. The weights in a neural network can be seen as a series of linear transformations that map input data to intermediate layers & eventually, to the output layer.

Crucial for learn complex patterns in data

⑤ Image & Signal Processing:

In image & Signal processing, linear transformations are used extensively.

Ex: Convolutional filters:- in image processing can be seen as linear transformation applied to local regions of a image. Fourier transforms, which decompose signals into sinusoidal components are linear transformations that convert time domain signals in to frequency-domain representation

⑥ understanding & interpretation:-

Linear transformations simplify complex relationships between variables into linear relationships, which are easier to understand & interpret.

For example:- linear regression provides a clear model of how each feature affects the target variable through linear co-efficients, making it easier to explain to stakeholders.

⑦ optimization & Solving Systems of Equations:

Linear transformations are used to solve systems of linear equations, which is a common problem in data analysis & optimization. Techniques like matrix inversion & the use of pseudo-inverse are essential for finding solutions in linear regression & other linear model

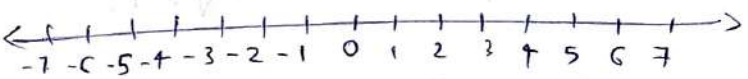
⑧ Theoretical foundations:-

Many advanced machine learning algorithms & statistical techniques have linear algebra and linear transformations at their core, understanding these fundamentals is crucial for grasping more complex topics like Support Vector Machines

Linear Transformation Visualization

$T: R \to R$ → Satisfy to properties of LT

1D

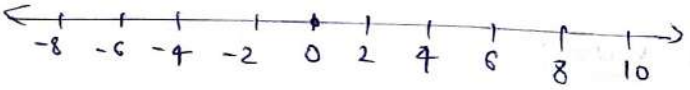


$T(x) = 2x$

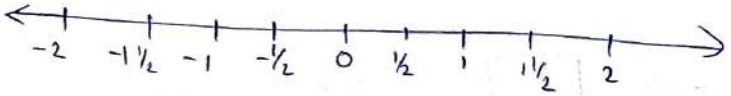
↓

$f(x) = 2x$

we can suppose it as functions for understanding



$T(x) = \frac{1}{2}x$



$T(x) = -3x$

one of the very important property of LT is origin must remain fixed

Property

① origin must be fixed

② All lines must remain lines

The better way to understand is by 2D matrix

Magnitude and unit vector :-

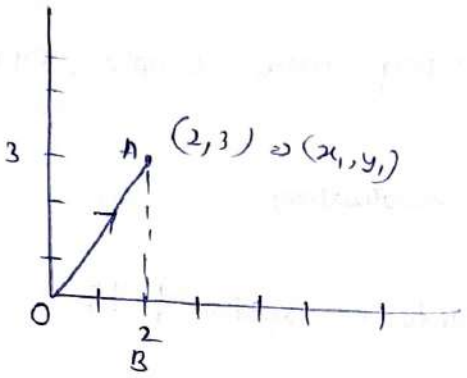
↓ Vector length

↳ vector has a length of 1

$\vec{x} \in R^n \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$\vec{A} = \begin{matrix} x \\ y \end{matrix}\begin{bmatrix} 2 \\ 3 \end{bmatrix} \in R^2$

2D



$OA = \|\vec{A}\| = \sqrt{OB^2 + OA^2} = \sqrt{2^2+3^2} = \sqrt{4+9} = \sqrt{13}$

$\|\vec{x}\| = \sqrt{x_1^2 + x_2^2 + x_3^3 \cdots x_n^2}$ magnitude / length of vector

unit vector :-

$\|\vec{u}\| = 1 \Rightarrow \hat{u}$ → cap

$\vec{V} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \rightarrow \vec{u} \Rightarrow \|\vec{u}\| = 1$ Ex: $\vec{v} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} \in R^3$

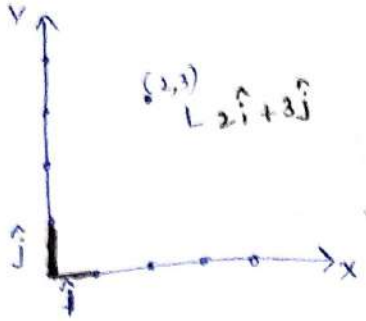
$$\vec{u} = \frac{1}{\|\vec{v}\|} \cdot \vec{V}$$

scalar multiplication

$\|\vec{v}\| = \sqrt{1^2+2^2+0^2} = \sqrt{1+4+0} = \sqrt{5}$

$\vec{u} = \frac{1}{\sqrt{5}} \cdot \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \\ 0 \end{bmatrix}$

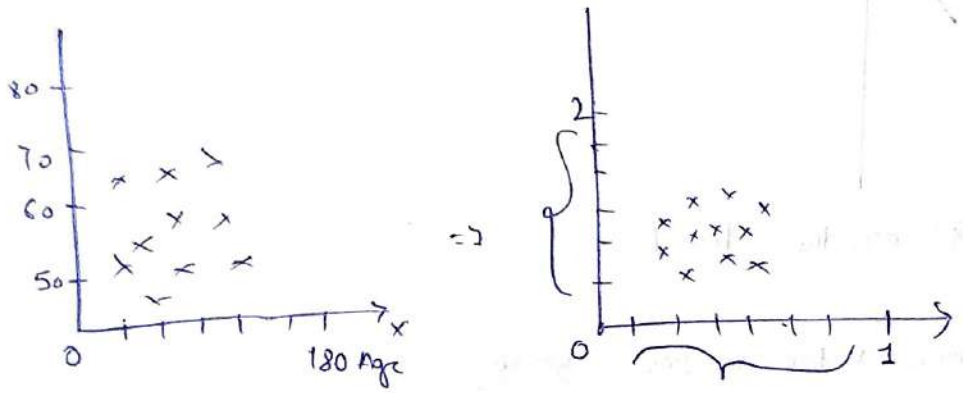
$\|\vec{u}\| = \sqrt{(1/\sqrt{5})^2 + (2/\sqrt{5})^2 + 0^2} = \sqrt{1/5 + 4/5 + 0} = \sqrt{5/5} = \sqrt{1} = 1 = \hat{u}$



(2,3)

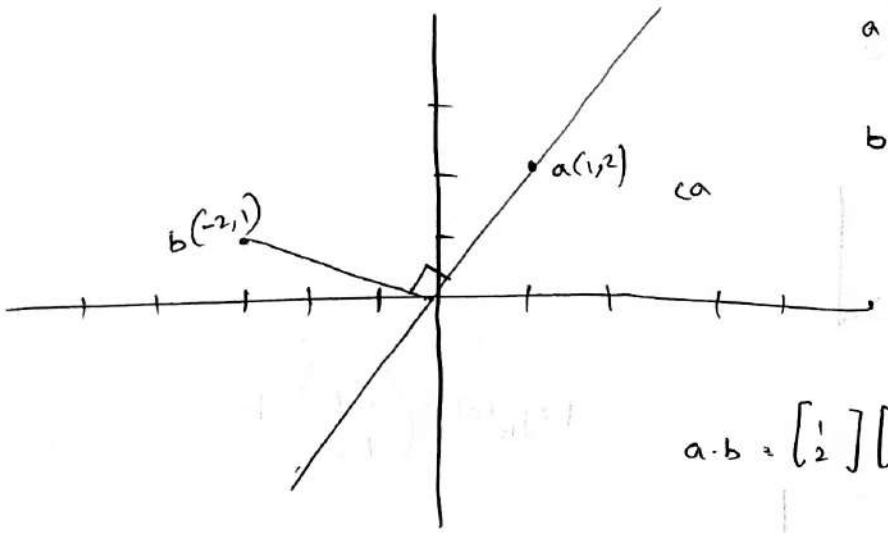
$\hat{i}, \hat{j}$ => Unit Vector

Normalization:



it improves optimization process

Introduction to projection:



$a\begin{bmatrix}1\\2\end{bmatrix}$

$b\begin{bmatrix}-2\\1\end{bmatrix}$

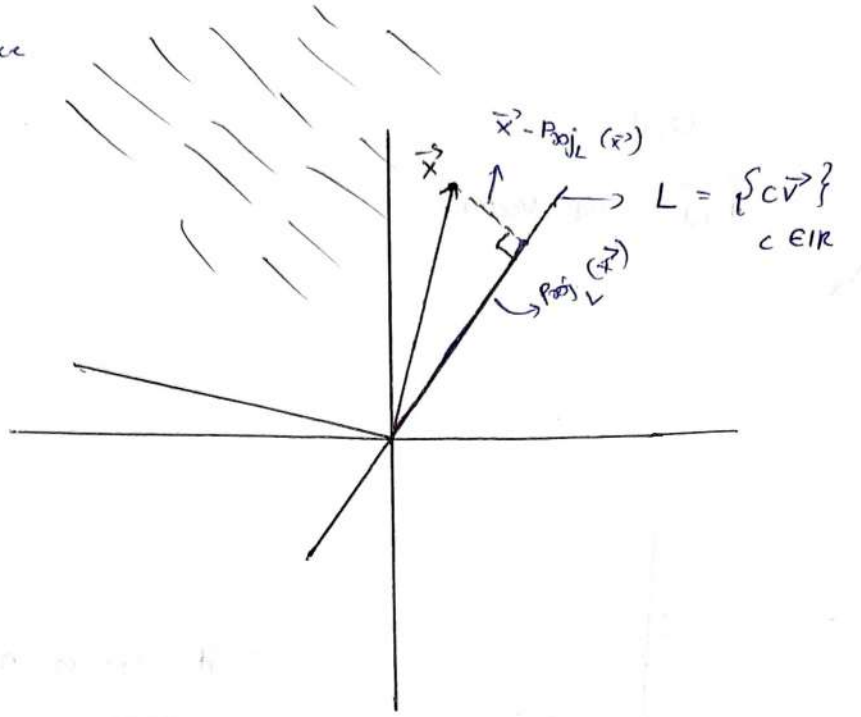
ca

$$a \cdot b = \begin{bmatrix}1\\2\end{bmatrix}\begin{bmatrix}-2\\1\end{bmatrix} = \begin{bmatrix}-2+2\end{bmatrix} = 0$$

$$a \cdot b = a_1 b_1 + a_2 b_2 \qquad a\begin{bmatrix}a_1\\a_2\end{bmatrix} \; b\begin{bmatrix}b_1\\b_2\end{bmatrix}$$

$\boxed{a \cdot b = 0}$ => Dot product is always 0.

Source



$Proj_L(\vec{x})$ => project the $\vec{x}$ on the Line L

It is a some vector is line where

$\vec{x} - proj_L(\vec{x})$ is perpendicular to L

$$(\vec{x} - c\vec{v}) \cdot \vec{v} = 0$$

$$\vec{x}\cdot\vec{v} - c\vec{v}\cdot\vec{v} = 0$$

$$\vec{x}\cdot\vec{v} = c\vec{v}\cdot\vec{v}$$

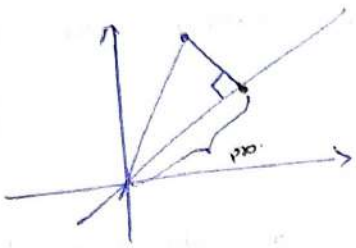
$$\boxed{c = \frac{\vec{x}\cdot\vec{v}}{\vec{v}\cdot\vec{v}}}$$

$proj_L(\vec{x}) = c\vec{v}$

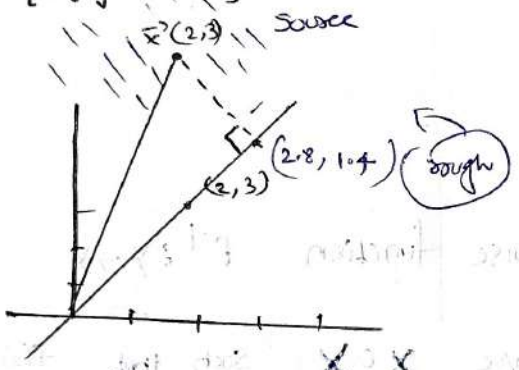
$$proj_b(a) = \left(\frac{a\cdot b}{b\cdot b}\right)\cdot b$$

$$\boxed{proj_L(\vec{x}) = \left(\frac{\vec{x}\cdot\vec{v}}{\vec{v}\cdot\vec{v}}\right)\cdot\vec{v}}$$

Problem: $\vec{v} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ & $\vec{v} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$



$L = \left\{ c \begin{bmatrix} 2 \\ 1 \end{bmatrix} \right\} \quad c \in \mathbb{R}$ $\qquad \vec{x} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$



$$\text{Proj}_L(\vec{x}) = \left(\frac{\begin{bmatrix} 2 \\ 3 \end{bmatrix}_{2\times1} \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix}_{2\times1}}{\begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}} \right) \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$= \frac{7}{5} \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 14/5 \\ 7/5 \end{bmatrix} = \begin{bmatrix} 2.8 \\ 1.4 \end{bmatrix}$$

Inverse of a function:-

A inverse of a function is a function that "reverses" the effect of the original function.

If you have a function f that maps an element x from set X to an element y in a set Y, then inverse function f^{-1} map y back to x

Defination:-

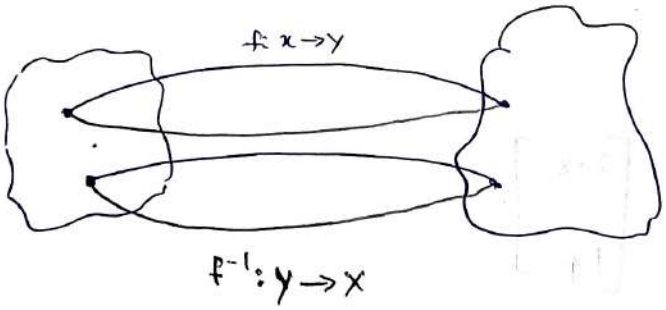
Given function $f: X \to Y$, then inverse function $f^{-1}: Y \to X$

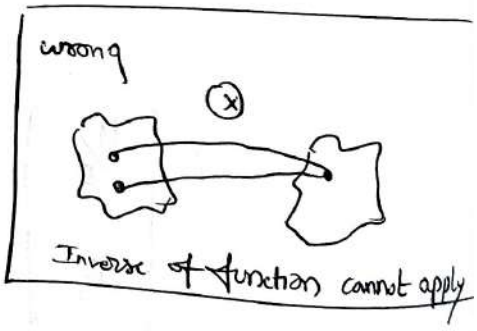
for every $y \in Y$, there is a unique $x \in X$ such that $f(x) = y$

The inverse function f^{-1} satisfies the following condition

1) For all $x \in X$: $f(f^{-1}(y)) = y$

2) for all $y \in Y$: $f^{-1}(f(x)) = x$





These conditions imply that applying the function & then its inverse will return the original value

Identity function

$I_X : X \to X \quad \Rightarrow I_X(a) = a$

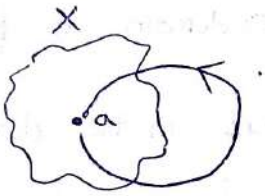
$a \in X$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad v = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$Iv = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

for a set X, the identity function I_X is defined as:

$I_X(a) = a$ for all $a \in X$



Properties of identity function

① Preservation :- Does not alter any element. If x is the domain, then the image of x under the identity f_n is x.

② Linearity :- Identity f_n is a linear transformation

* $I(u+v) = I(u) + I(v)$

* $I(cu) = c\,I(u) = cu$

③ Identity matrix : $n \times n$ => All the diagonal elements will be 1 and other's 0's

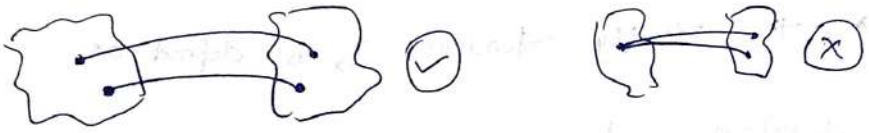
$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{3\times 3} \qquad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

4 Inverse : The identity function f_n is its own inverse

Existence & uniqueness :-

A function f has an inverse if & only if it is bijective.

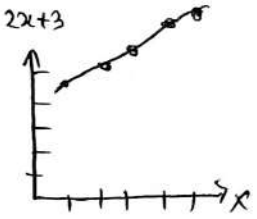
1) Injective (one to one) ; Different elements in the domain map to different element in the co domain



2) Surjective (on to) : Every element in the co domain is the image of atleast one element in the domain.

Eg : Linear function

$f(x) = 2x+3$



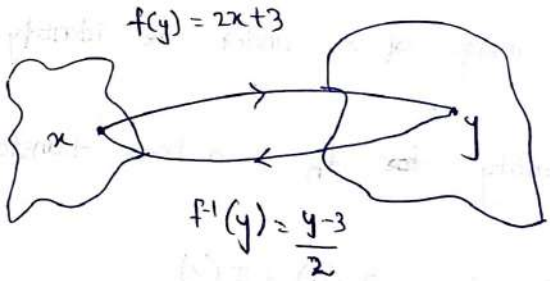
Find the inverse :-

$y = 2x + 3$ for x

$y = 2x + 3$

$y - 3 = 2x$

$x = \frac{y-3}{2}$



The inverse function $\boxed{f^{-1}(y) = \frac{y-3}{2}}$

Verification:

① $f(f^{-1}(y)) = f\left(\frac{y-3}{2}\right) = 2\left(\frac{y-3}{2}\right)+3 = y$

② $f^{-1}(f(x)) = f^{-1}(2x+3) = \frac{(2x+3)-3}{2} = x$

Application of Inverse function In Data Science :-

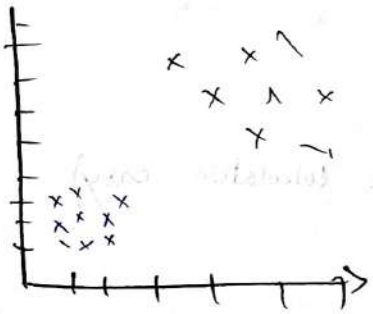
① Normalization And Standardization :

Ex: House Price Dataset

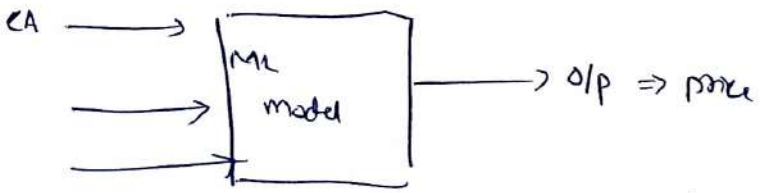
Independent feature ↑ — Dependent feature ↑

Carpet Area	No. of Room	Area (sqft)	price
1500	4	1800	50 lakhs
1800	3	2100	60 lakhs

Some Scaling is done to the vectors in a Same Scale



instead of having huge magnitude we try to reduce it



[0 - 1]

i) mathematical operation } => Quickly => optimization
operations

Standardization

Eve take a Data feature [entire] which will transform to the new feature of having $\mu = 0$ & $\sigma = 1$

No of Rooms

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \in R \Rightarrow \text{Standardization Transformation} \Rightarrow Z = \frac{x_i - \mu}{\sigma}$$

No. of Rooms

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \in IR \xrightarrow{f(x)} \begin{bmatrix} -1.5/2 \\ -0.5/2 \\ 0.5/2 \\ 1.5/2 \end{bmatrix}$$

$f^{-1}(x)$

No. of Rooms^1

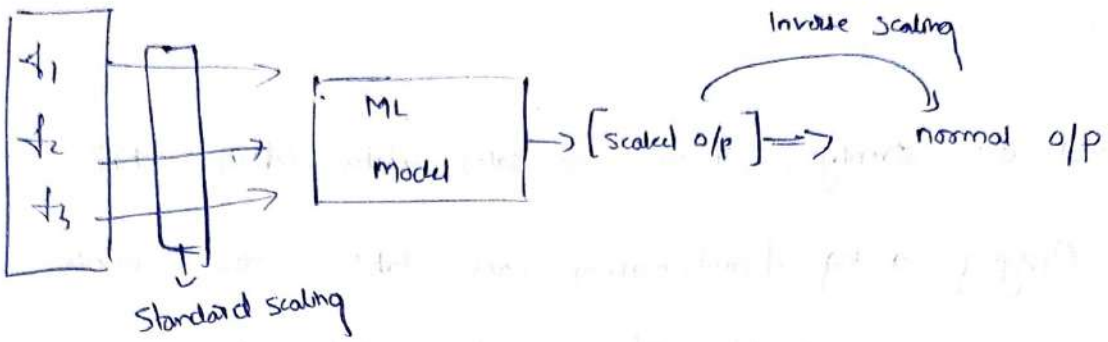
$$Z_1 = \frac{1 - \mu}{\sigma}$$

$$\mu = \frac{1+2+3+4}{4} = \frac{10}{4} = 2.5$$

$\sigma = 2$ (not correct answer, just to make calculation easy)

Original Transformation $\quad Z = \frac{x - \mu}{\sigma} \qquad x = \frac{-1.5}{2} \times 2 + 2.5 = 1$

Inverse Transformation $\quad x = z\sigma + \mu$



Use case :-

After training a ML model on standardized data, the predictions are often rescaled back to the original scale to interpret the result in a meaningful way. For instance, if house prices were standardized, the inverse transformation would convert the standardized predictions back to the original price scale.

* Normalization :-

Feature Scaling with Min Max Normalization

→ original Transformation : $z = \frac{x - \min(x)}{\max(x) - \min(x)}$ → T : x → y

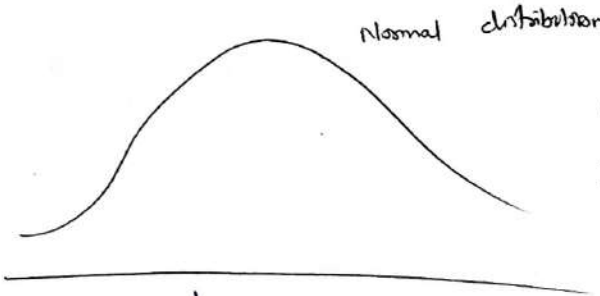
→ Inverse Transformation : $x = \frac{z\,(\max(x) - \min(x))}{x - \min(x)}$ → T^{-1} : Y → X

* Distribution of Data

* logarithmic distribution

=>

=>



OT :- $y = \log(x)$

IT :- $x = e^y$

↓

Because ML model need ND

Uses case :-

In financial data analysis, income or sales data often exhibit skewness. Applying a log transformation can stabilize the variance & make patterns more visible. After model prediction, the inverse log transformation is applied to interpret the result on the original scale

box - cox transformation.

⑤ Data Encryption & Decryption :-

Encryption function: $E(p) = c$ (where p is plain text & c is a cipher text)

Decryption function : $D(c) = p$

Use case :-

Sensitive data like personal information, financial records, & medical data are encrypted before storage or transformation.

Decryption is applied to retrieve the original information.

How to find inverse of a matrix

① Determinant

② How to Inverse

Eg: 2×2 matrix

find A^{-1} & verify using a transformation

$$A = \begin{bmatrix} 4 & 7 \\ 2 & 6 \end{bmatrix}$$

$T: R \to R$

$\Downarrow$

$A \cdot x = y$

$\Downarrow$

A^{-1}

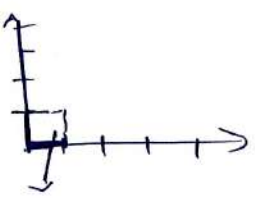
Find the inverse of A

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$\Rightarrow$

The determinant is a scalar value that can be computed from a square matrix. It provides important information about the matrix, such as whether the matrix is invertible (i.e has an inverse), & it also has geometric interpretations, such as describing the scaling factor of linear transformations represented by the matrix.



Area = Determinant

Determinant $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \Rightarrow \boxed{ad-cb} \Rightarrow$ scalar $\Rightarrow$ determinant

$$A = \begin{bmatrix} 4 & 7 \\ 2 & 6 \end{bmatrix}$$

$$\det(A) = 24 - 14 = 10$$

Step 1 :- $\det(A) = 10 \neq 0$

non zero $\Rightarrow$ inverse of matrix exists since $\det(A) \neq 0$ ∴

matrix A is inversible

step ② find the inverse

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \Rightarrow \text{adjacent matrix}$$

$$= \frac{1}{10} \begin{bmatrix} 6 & -7 \\ -2 & 4 \end{bmatrix} = \begin{bmatrix} 0.6 & -0.7 \\ -0.2 & 0.4 \end{bmatrix}$$

step3 Verification using a vector

$X = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow x$ using A then use A^{-1} to recover the original vector

Transformation using A;

$$Y = AX = \begin{bmatrix} 4 & 7 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4+7 \\ 2+6 \end{bmatrix} = \begin{bmatrix} 11 \\ 8 \end{bmatrix}$$

Recovering x using A^{-1}

$$x = A^{-1}y = \begin{bmatrix} 0.6 & 0.7 \\ -0.2 & 0.4 \end{bmatrix} \begin{bmatrix} 11 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

This A^{-1} Successfully recover the original vector X

Eigen Vectors And Eigen Values

Eigenvalues & Eigen vectors are fundamental concepts in linear algebra that have numerous applications in various fields such as physics, computer science, and data science. They provide insights into the properties of linear transformations represented by matrices

Definition

Eigen value (λ) :- A scalar that indicates how much an eign vector is stretched or compressed during linear transformation

Eigen vector : A non zero vector that only changes in scale (not direction) when a linear transformation is applied.

* To calculate eigen vector given by $AV = \lambda V$

For Square matrix A, an eigen vector and its corresponding eigen value λ satisfy the above equation

$$A = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$$

① Find Eigen values

$$\det(A - \lambda I) = 0$$

$$A - \lambda I = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix}$$

$$d\begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix} = 0 \Rightarrow (4-\lambda)(3-\lambda) - 2 = 0$$

$$\Rightarrow 12 - 4\lambda - 3\lambda + \lambda^2 - 2 = 0$$

$$\Rightarrow \lambda^2 - 7\lambda + 10 = 0$$

② Solve the equation

$$\lambda^2 - 7\lambda + 10 = 0$$

$$\boxed{\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}}$$

$a = 1$, $b = -7$ & $c = 10$

$$\lambda = \frac{7 \pm \sqrt{49 - 40}}{2} = \frac{7 \pm \sqrt{9}}{2} = \frac{7 \pm 3}{2}$$

$$\lambda_1 = \frac{10}{2} = 5 \quad (\text{or}) \quad \lambda_2 = \frac{4}{2} = 2$$

Eigen values of A $\lambda_1 = 5$ & $\lambda_2 = 2$

③ Find the eigen vectors:

$$(A - \lambda I)\, V = 0$$

for $\lambda_1 = 5$

$$A - 5I = \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} - \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} = \begin{bmatrix} 1-5 & 1-0 \\ 2-0 & 3-5 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}$$

$$(A - \lambda I)\, V = 0$$

$$\begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 0$$

$2 \times 2 \quad 2 \times 1$

$$-x + y = 0$$

$$\begin{cases} -x + y = 0 \\ 2x - 2y = 0 \end{cases}$$

$\boxed{y = x}$ $\boxed{x = y}$ → Eigen vector corresponding to Eigen value $\lambda_1 = 5$

for $\lambda_2 = 2$

$$A - 2I = \begin{bmatrix} 1-2 & 1 \\ 2 & 3-2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}$$

$$(A - \lambda I)\, V = 0$$

$$\begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$$

$$2x + y = 0$$

$\boxed{y = -2x}$ → Eigen vector corresponding to $\lambda_2 = 5$

$$A = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} \qquad \lambda_1 = 5 \qquad \lambda_2 = 2$$

for $\lambda_1 = 5$ $\quad V_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

for $\lambda_2 = 2$ $\quad V_1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$

$$A\vec{V} = \lambda\vec{V}$$

↳ how much eigen value that the vector $\vec{V}$ is stretched

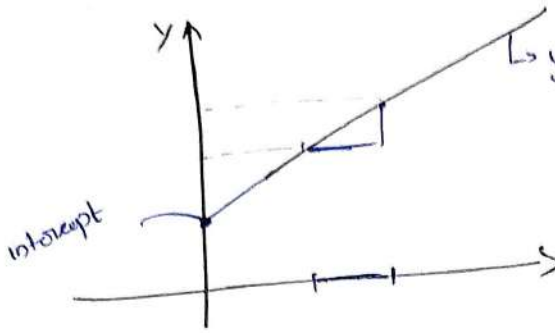
These eigenvectors & eigenvalues describe how the matrix A A scales and rotates vectors in its transformation. Eigen values indicates the factor by which the eigenvectors are stretched or compressed, and eigenvectors provide the directions in which this stretching or compression occurs

—Applications:

—Principal component analysis → Dimensionality reduction

Equation of line, 3d plane and hyperplane (in Dimension)

(2D)



$y = mx + c$ — intercept
slope

$y = \beta_1 x + \beta_0$

$ax + by + c = 0$

$by + c = -ax$

$by = -ax - c$

~~$y = \left(\frac{a}{b}\right)x - \frac{c}{b}$~~

$y = \frac{-a}{b}x - \frac{c}{b}$

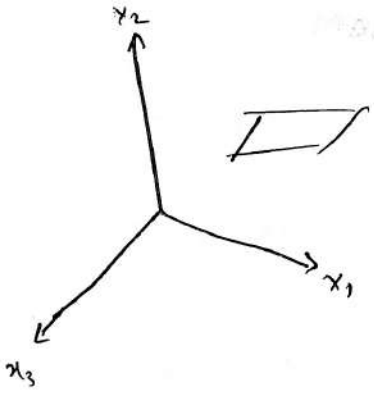
m $\quad$ c

$w_1 x_1 + w_2 x_2 + b = 0$

$\boxed{w^T x + b = 0}$

eq of straight line

(3D)



$w_1 x_1 + w_2 x_2 + w_3 x_3 + b = 0$

$\boxed{w^T x + b = 0}$

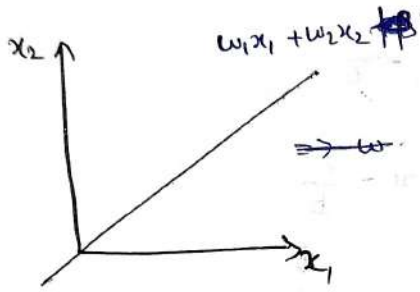
$$w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \cdot x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

n-Dimension plane

$$w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n + b = 0$$

$$\boxed{w^Tx + b = 0}$$

If eqn pass through origin $\Rightarrow$ $\boxed{\beta = 0}$



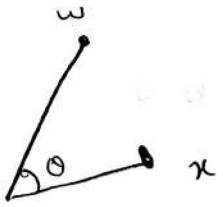
Equation of a straight line passing through origin

$$w^Tx = 0$$

3.142

Equation of a plane = Π_n : $w^Tx = 0$

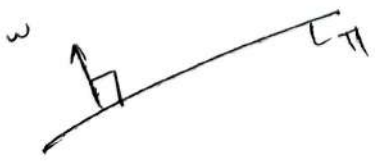
$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$

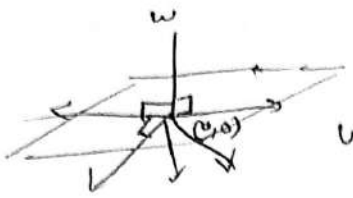


$$w.x = w^Tx = \|w\| \|x\| \cos\theta = 0$$

$\theta = 90$

$\cos\theta = 0$





$w \perp \pi$

intercept - 0

Statistics :-

Def: Statistics is the Science of Collecting, organizing and analyzing data.

Decision making

Data :- facts or pieces of information => measured, collected, Analyzed

Ex :- weight of student in the class * IQ of the student of class

{ 60kg, 50kg, ... } { 100, 90, 99 ... }

city	Area	No. of rooms	price
Hyd	1200	2	45 lakhs
New York	1250	2.5	50 lakhs

Data Analyst → Report → Visualization → Meaning Decision

project

Analysis can be done to the data with the help of statistics

Application: Imp statistics

① Data exploration & summarize

② Model build & validation

③ Statistical Analysis ⇒ Different analysis that we does on sample data to make condition about popular

④ Hypothesis Testing

⑤ optimization & efficiency

⑥ Reporting

Types of statistics

① Descriptive Statistics
② Inferential Statistics

① Descriptive Statistics :-

Descriptive statistics involves methods for summarizing & organizing data to make it understandable. This type of statistics helps to describe the basic features of the data in a study

1] Measure of central Tendency

[mean, median, mode]

2] Measure of Dispersion

[variance, standard, deviation]

3] Data Distribution

i] Histograms [to understand probability distribution function & probability density function]

ii] Box plot

iii] pie chart

iv] PDF, PMF

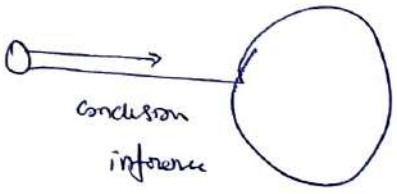
4] Summary statistics

i] five number Summary

Q_1, Q_2, Q_3, Maximum

[2] Inferential Statistics:-

Inferential statistics involves methods for making predictions or inferences about a population based on a sample of data. It allows for hypothesis testing, estimation & drawing conclusion



① Hypothesis Testing

② P value

③ Confidence interval

④ statistical Analysis Test ① Z test

② t test

③ ANOVA → F test

④ Chi square

Type of Statistics	Key Concepts	Examples
Descriptive Statistics	Measures of central tendency (mean, median, mode), Measure of Dispersion (Range, variance, Standard deviation), Data Distribution (Histograms, Box plots), Summary statistics (five-number Summary)	Mean score of the Students, Range of temperature, Histogram of ages
Inferential Statistics	Hypothesis Testing (Null & Alternative hypothesis, P-value), Confidence intervals, Regression Analysis (Simple & multiple Linear Regression), ANNOA, Chi-square Test	P-value in test ~~Gauge~~ scores Comparision, 95% confidence interval for average height, predicting house price, Comparing test scores of different Schools, Association between gender & product perference.

Eg:- Let Say there are 20 Statistics class in your college, & you have Collected the height of students in the class

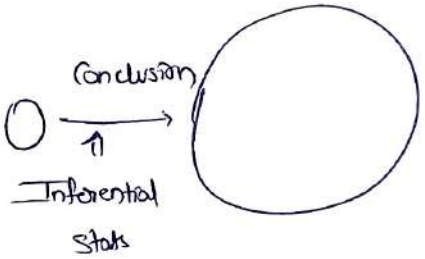
→ Heights are recorded [175cm, 180 cm, 140cm, 135 cm, 160cm, 120 cm]

Descriptive question:-

"What is the average height of the entire classroom" ← Measure of Central Tendency

Inferential Question

"Are the higher of the Sample students in classroom similar to what you Expect in the entire College"?



(3) Population & Sampling Data :-

Population :- A Population is the entire set on individuals or objects of interest in a particular study. It includes all members of a defined group that we are Studing or Collecting information on.

Characteristics:-

① Complete Set: Contains all the observation of interest

② Parameter: A numerical value Summarizing the entire population

Ex: ① population mean (μ)

② population Variance (σ^2)

Example:

① population in a School study

* All student enrolled in a School

Use case

Determine the avg height of student,

population mean

② Population in market Research

* All Consumers in a city

Use cases

To understand the purchasing behaviour of all consumers

③ population in a medical study

* All the patients with a Specific disease

Use case

To study the effective of a drug.

Sample data :-

* A Sample is a subset of the population that is used to represent the entire group. Sampling involves selecting a group of individuals or observations from the population to draw conclusions about the whole population.

Characteristics :-

① Subset : Represents a portion of the population.

② Statistic : A numerical value summarizing the sample data

[Sample mean, Sample variance]

③ Random Sampling : Sampling should be randomly selected to avoid bias

Example :-

① Sample in a School Study

(i) A group of 50 students from School

(ii) Usecase : Estimate the average height of students in a school

② Sample in Market research :-

* 500 consumer from the city
* Behaviour → population

⇒ ③ Sample in a Medical study

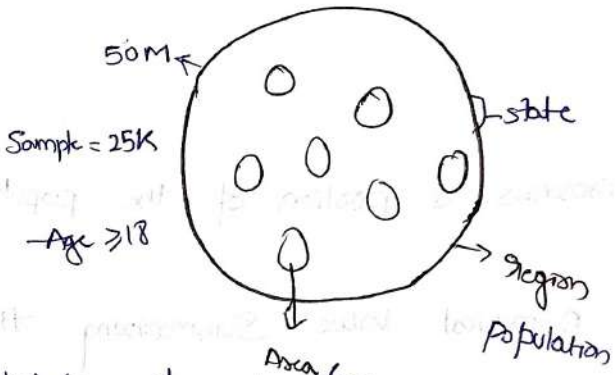
* 200 patients
* Tests the effectiveness of the drug

Types of Sampling Techniques :-

① Probability Sampling

② Non probability Sampling

Eg :- Exit poll A, B, C, D



① Probability Sampling :- All individual have chance to being selected

(a) Sample Random Sampling :

The Sampling data is collected randomly

⇒ Every no of the population has an equal chance of being selected

Ex :
- Selecting people randomly
- Draw names random from a class of student

(b) Systematic Sampling

Select every n^{th} member of the population after a random starting point.

Ex : Airport → credit card → 5^{th} person, 10^{th}, 15^{th}

- Feedback Survey → Selected every 11^{th} number → feedback Survey

© Stratified Sampling:

Divide the population into strata (group) based on specific characteristics & then randomly sampling from each strata

Eg:- Divide employees by department & then randomly select a proportional number from each department to form a survey sample

Eg:- Age → <12 | 12-18 | >18 & politics

(d) Cluster Sampling:-

Divide the population in to clusters, randomly selecting clusters, then Sampling all the members from the selected clusters

Eg:- Randomly selecting several schools from a district & surveying all teaching within these schools.

(e) Multi stage Sampling:-

Combining several Sampling method usually involved selecting clusters, then randomly sampling with in those cluster

Eg:- Randomly selecting Citical, each selecting city randomly selecting households to survey

(2) Non Probability Sampling:-

Not all the members have equal chance of being Selected

(a) Convenience:-

Selecting individual who are easy to reach

Ex: Surveying students in a classroom because they are readily available

* Consecutive sampling :- Similar to convenience sampling with slight variation

The researcher picks a single person or group of people for sample

(b) Judgmental (purposive) Sampling:-

Select individual based on the research's judgement ⇒ useful or representation

↓

Eg:- choose experts in a field to participate. ~~Select~~ focused on specific characteristic / condition

(c) Snowball Sampling:-

Existing participants recruit new participants creating a chain of referrals - used for hidden or hard to reach population

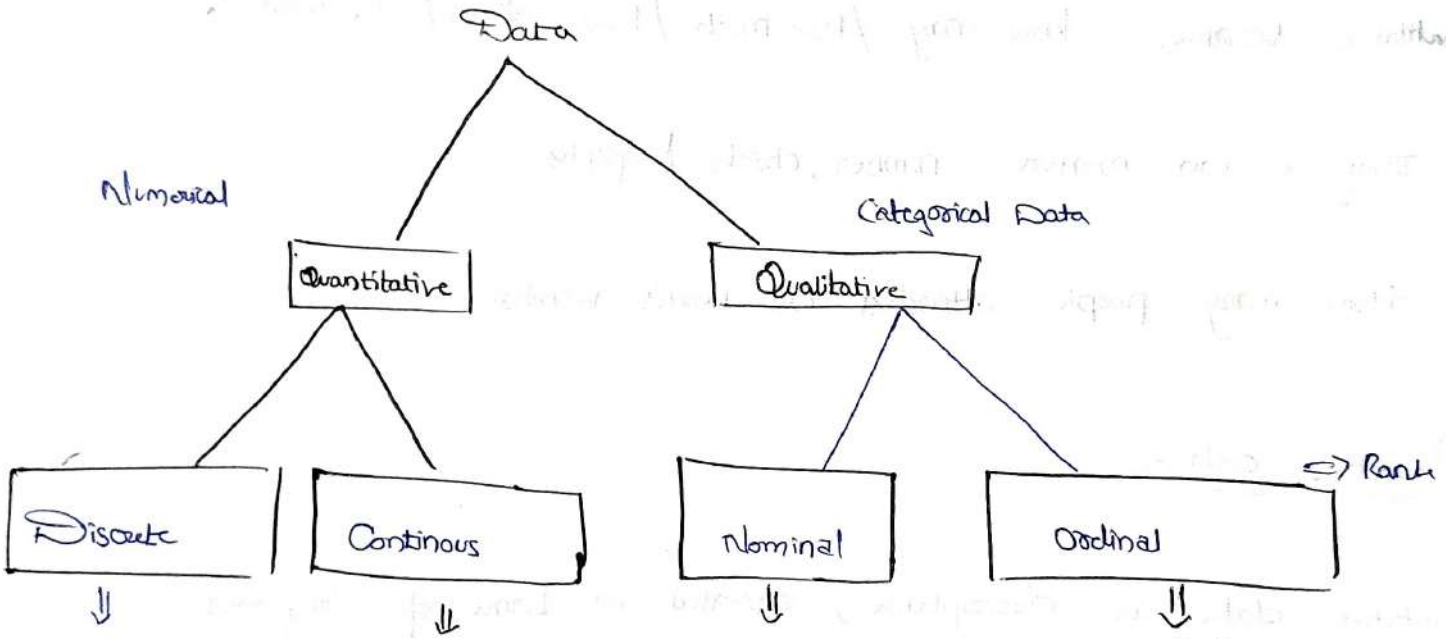
Eg:- Survey members of a rare distance.

(d) Data Sampling: Sample represents certain characteristics in % like

- Age, group, gender, caste

Eg: Choosing 60% woman & 40% man

Types of data:-



Whole number

Eg:

1) No of childrens in a family
2) Marks obtained by students of a class10

Any value

Eg:

Weight
Height
Temperature
Speed
=> It can be negative / positive decimal

No Rank

Categorical data

Eg:- Gender

M, F

- Blood group
- Pincode

Rank

Categorical data

Eg:- Customer feedback

Good bad better
1 3 2

Quantitative data :

It refer to any information that can be quantified - that is, numbers. If it can be counted or measured & given a numerical value, it's quantitive in nature.

Specific & objective measure of numerical facts

Quantitative variable :- how may / how much / how often / The what

Thing u can measure number, charts & graphs

Ex How many people attended last week's webinar

Qualitative data :-

Qualitative data is descriptive, expressed in terms of languages rather than numerical values

Subjective explanatory measures of qualities & characteristics

why or how

How do customers feel about their customer service experience,

* Scales of Measurement of Data :-

The Scales of measurement describe the nature of information with in the value assigned to variable

↳ primary scales of measurement

1. Nominal Scale
2. Ordinal Scale
3. Interval
4. Ratio

1 Nominal Scale :

This Scale classifies data into distinct categories that do not have an intrinsic order

Qualitative data / categorical data

Characteristics :-

(i) Data is categorized based on label, names or qualities

(ii) Categories are mutually exclusive

(iii) No logical order among categories [No rank]

any one will be true

Ex:

- Gender
  - → M
  - → F
- Colors
  - Red → 5 — 50%
  - Blue → 4 — 40%
  - Pink → 1 — 10%

② ordinal scale :-

This scale classifies the data into categories that can be ranked or order

Characteristics

(i) Data is categorized & Ranked in a specific order

(ii) The interval between rank are not necessarily equal

Example :-

Education level	Ranks
High School	1
Bachelor	2
Masters	3
Doctorate	4

Customer feedback	
Satisfied	1
Very satisfied	2
Not satisfied	0

Socio Economic status :-

Low	2
Middle	1
High	0

③ Interval scale

The interval scale not only categorizes & orders but also specify the exact different between intervals. It lacks a true zero point

Characteristics:-

① Data is ordered with consists interval between values

② Allows for meaningful comparision of differentes [ratio cannot be measured]

③ No true zero point

Ex: IQ Score:

90, 10, 100,

IQ # 0

4 Ratio Scale :-

* The order matters
* Differentes are measurable
* Contains a 0 starting point

Ex:- Student marks in a ~~Gate~~ class

0, 90, 60, 30, 45

Ratio = $\frac{90}{30} = \frac{3}{1}$

3:1

Ex: Weight

10, 20, 30, 40

Income: 10,000 1p

20000

30000 4p I = 4X II

40000 4p

Assignment:-

1) Length of different rivers in the world? { Ratio }

2) Favorite food based on Gender? { Nominal }

3) Marital status? { Nominal }

5) IQ Measurement { Ratio }

Centre
Measure of ↑ tendancy:

Measure of central tendancy are statistical metrics that describes the Centre point or type value of a data set. They provide a single value that summarizes a set of data by identifying the central position within that dataset

1 Mean or Average

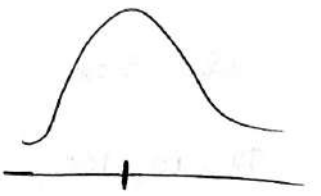
2 Median

3 Mode

Age = [24, 28, 15, 20, 30]

↓

Centre point



1 Mean:- mean is the sum of all values divided by the number of values

- Population mean (μ) || Sample mean ($\bar{x}$)
- Population size (N) || Sample size (n)

$n \leq N$

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N} \qquad \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

* Here X is a random variable

$$x = \{5, 8, 12, 15, 20\}$$

$N = 5$

$$\mu = \frac{5+8+15+12+20}{5} = \frac{60}{5} = 12$$

* Characteristics:

- Affected by the extreme outliers
- Used for interval & ratio data

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

$\mu = \frac{1+2+3+4+5}{5} = 3$

$x = \{1,2,3,4,5,100\}$

$x = \frac{1+2+3+4+5+100}{6} = 23$

Median:

The median is the middle value in a dataset when the values are arranged is ascending or descending order

$x = \{1,2,3,4,5\}$ $\quad x = \{3,4,1,5,2,100\} \Rightarrow [1,2,3,4,5,100]$

No. of element = 5 [odd] $\quad$ No of element = 6 [even]

$\Rightarrow \left(\frac{n+1}{2}\right)^{th}$ term $\quad \Rightarrow \frac{\left(\frac{n}{2}\right)^{th}\text{ term} + \left(\frac{n+1}{2}\right)^{th}\text{ term}}{2}$

Median $= \left(\frac{5+1}{2}\right)^{th}$ term $\quad \Rightarrow$ median $= \frac{3+4}{2}$

$= 3^{th}$ term $\quad = 3.5$

$= 3$

Characteristics:

* Not affected by extreme outliers
* Used for ordinal, interval & ration data

③ Mode

The mode is the value that appears most frequently in a dataset

Dataset :- 2, 4, 4, 9, 9, 8, 7, 4, 10

mode = 4

=> 6, 6, 3, 17, 4, 9, 3

mode = 6, 3 (bimodal) Similarly multimodal

* Characteristic :-

① Not affected by extreme values

② used for Nominal, ordinal, interval & ratio data

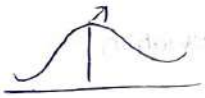
Choosing the Appropriate Measures :-

1 Mean :- Best used when data is Symmetrically distributed without outliers provides a mathematical average, which is useful for further Statistical calculates

2 Median: Best used when data is skewed or contains outliers provides the middle value, which better represents the center of a skewed data

③ Mode: Best used for categorical data to identify the most common category. Also useful for identifying the most frequent value in, ordinal, interval or ratio data

This missing values are replaced with the mean / median to get accurate output



Real world Application:

Used in feature engineering

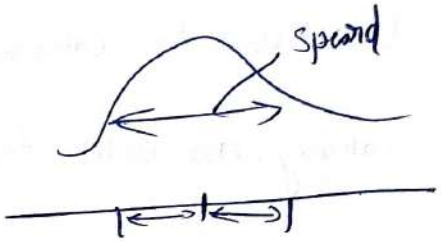
Nominal data

mode

Age	Weight	Salary	Gender	Degree	
24	70	40k	M	BE	Handeling values
25	80	70k	F	—	Nominal + ordinal
27	95	80k	F	—	
24	—	55k	M	PhD	
32	—	45k	F	B.E	
—	60	—	M	Master	
—	75	—	M	BSc	
40	90	35k	—	BE	

Measure of Dispersion:-

Measures of dispersion describe the spread or variability of a dataset. They indicate how much the values in a dataset differ from the Central tendancy



Common measure of Dispression:

1. Range
2. Variance
3. Standard Deviation
4. Interquartile range (IQR)

1 Range: Range is the difference between the maximum & minimum value in a dataset

Range = Maximum value - Minimum Value

Ex: Ages {14, 13, 10, 20, 25, 75, 15}

Range = 75 - 10 = 65

Characteristics:-

1. Simple to calculate
2. Sensitive to outliers
3. Rough measure of dispersion

Weight = {34, 40, 45, 30}, 100}

Range = 45 - 30 = 15

Range = 100 - 30 = 70

2 Variance:-

Variance measures the average squared deviation of each value from the mean. It provides a sense of how much the values in a dataset vary

Population variance	Sample Variance
$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$	$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$
$x_i \Rightarrow$ Data points	$x_i \rightarrow$ Data points
$\mu \Rightarrow$ population mean	$\bar{x} \rightarrow$ Sample mean
$N \rightarrow$ population size	$n \rightarrow$ Sample size

Example: Size of a flower petals

$\{5, 8, 12, 15, 20\} \Rightarrow$ Variance of this distribution

$N = 5$

$$\mu = \frac{5+8+12+15+20}{5} = \frac{60}{5} = 12$$

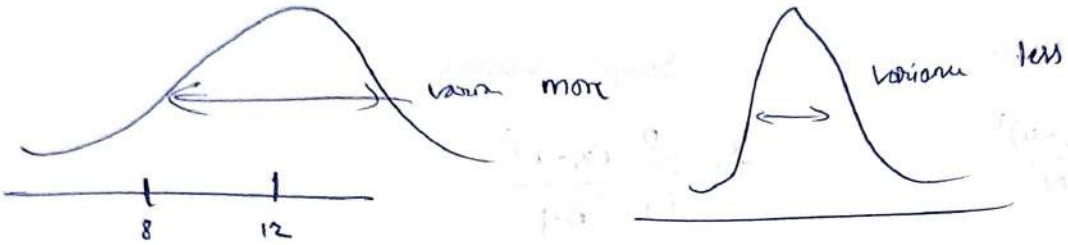
$$\text{Variance} = \frac{(12-5)^2 + (12-8)^2 + (12-12)^2 + (12-15)^2 + (12-20)^2}{5}$$

$$= \frac{49 + 16 + 0 + 9 + 64}{5}$$

$$= 27.6$$

Characteristics:-

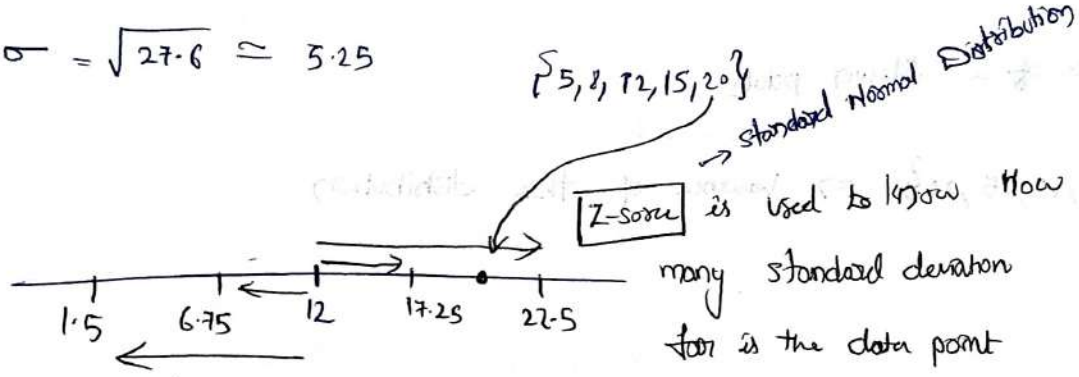
* provide a precise measure of variability
* Units are squared of the original data units
* More sensitive to outliers than the range



③ Standard Deviation

The standard Deviation is the square root of the variance

$$\sigma = \sqrt{27.6} \simeq 5.25$$



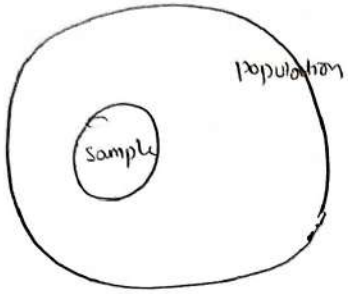
Characteristics:-

① provide a clear measure of spread in the same units as the data

② Sensitive to outliers.

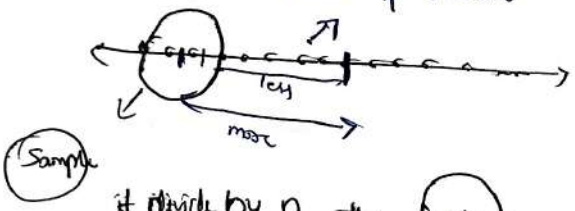
Sample Variance :-

if we use Sample variance formula as $S^2 = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n}$



Ages : $\{x_1, x_2, x_3 \text{ ---- } x_n\}$

$\bar{x} \ll \mu$ & $s^2 \ll \sigma^2$



if divide by n the

Sample mean will in middle & it will be far away from the population mean / variance

variance

* So the have population variance is underestimated.

- thus we use: $s^2 = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n-1}$ & It is called a

* Bessel's correction (using n-1) compensates for the estimation

** "loose one degree of freedom" when we estimate $\bar{x}$

you are using information(n) from sample itself, not the true population mean. This makes the Sample variance a biased estimator of the population variance if you divide by n

using n-1 corrects that bais & gives you an unbiased estimate of the population variance.

Random Variables (X) :-

A random variables is a mathematical function that assigns numerical values to the outcomes of a random experiment

$y = 5x + 5$

$x = 1, 2, 3, 4 \ldots$

Tossing coin

$x = \begin{cases} 0 & H \\ 1 & T \end{cases}$

Rolling a die

$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{cases}$

rolling a pair dice

Types of Random variable → example

① Discrete random variable :- A Discreate random variable is used to quantify the outcome of a random experiment. It takes a countable number of possible outcomes. It also called as stochastic variable

Generally counted as 0, 1, 2, 3 —

def:- It is a type of a variable whose value depends upon the numerical outcomes of a certain random experiment

Continous variable:-

It is a variable that takes any real value with in a given range or interval, including fractions & decimals. It is possible values are uncountably infinite

Ex:- Temperature on a given day

26.4°C, 26.45°C, -- etc

Percentiles and Quartiles:-

Percentage = {1, 2, 3, 4, 5, 6}

No of odd numbers = 3

percentage of odd numbers in a group = $\frac{3}{6} \times 100 = 50\%$

Percentile:- A percentile is a value below which certain percentage of observations lie

{ 2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 9, 9, 10 } n = 9

$\frac{3+4}{2}$

$$\text{Percentile} = \frac{\text{number of value below } x}{n} \times 100$$

$$= \frac{11}{14} \times 100 = 78\% \text{ of value } 9$$

25% is 3.75

$$\Rightarrow \text{Value} = \frac{\text{Percentile} \times (n+1)}{100}$$

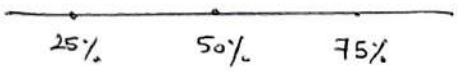
$$= \frac{25}{100} \times (15)$$

$$= \frac{15}{4}$$

$= 3.75 \neq 3.5$

$\frac{3+4}{2} = 3.5$

Quartiles



25% = first Quartile

50% = 2nd Quartile

75% = 3rd Quartile

5 Number Summary

Used in F.E 4 to remove outliers

① Minimum

② 1st Quartile (25%)

③ Median

④ 3rd Quartile (75%)

⑤ Maximum

Eg:- 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27

27 — Outlier.

* The ~~first~~ step to remove outlier is to define

lower fence & higher fence value

lower fence ~~&~~ higher fence value

- Any thing lower than the lower fence value is removed as outlier.
- Any thing higher (higher) than the fence value will be remove as outliers

Lower fence = Q1 − 1.5 (IQR)

Higher fence = Q3 + 1.5 (IQR)

$$Q1 = \frac{\text{Percentile}}{100} \times (n+1) = \frac{25}{100} \times (19+1) = 5^{th} \text{ position}$$

$$Q3 = \frac{15}{\cancel{75}/\cancel{100}} \times \frac{1}{\cancel{20}} = 15^{th} \text{ position}$$

$$\frac{5}{1}$$

IQR = $Q_3 - Q_1 = 7 - 3 = 4$

Lower fence = $Q_1 - 1.5\ (IQR)$ higher fence = $Q_3 + 1.5(IQR)$

$= 3 - 1.5(4)$ $= 7 + 1.5(4)$

$= 3 - 6 = \boxed{-3}$ $= 7 + 6$

$\boxed{= 13}$

After removing outliers

Eg:- 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9

Minimum = 1

1st Quartile = 3

Median = 5

3rd Quartile = 7

Maximum = 9

} This information indicates the 5 number Summary

* Based on the particular value we can define a box plot. (import)

* What kind of point you can actually use to see or visualize the outliers ? box plot

* One more plot is Whisker plot:

Histogram And skewness

A histogram is a graphical representation of the numerical data. It is an estimate of the probability distribution of a continuous variable and is used to visualize the shape, central tendency & variability of a dataset

Ages = {11, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50}

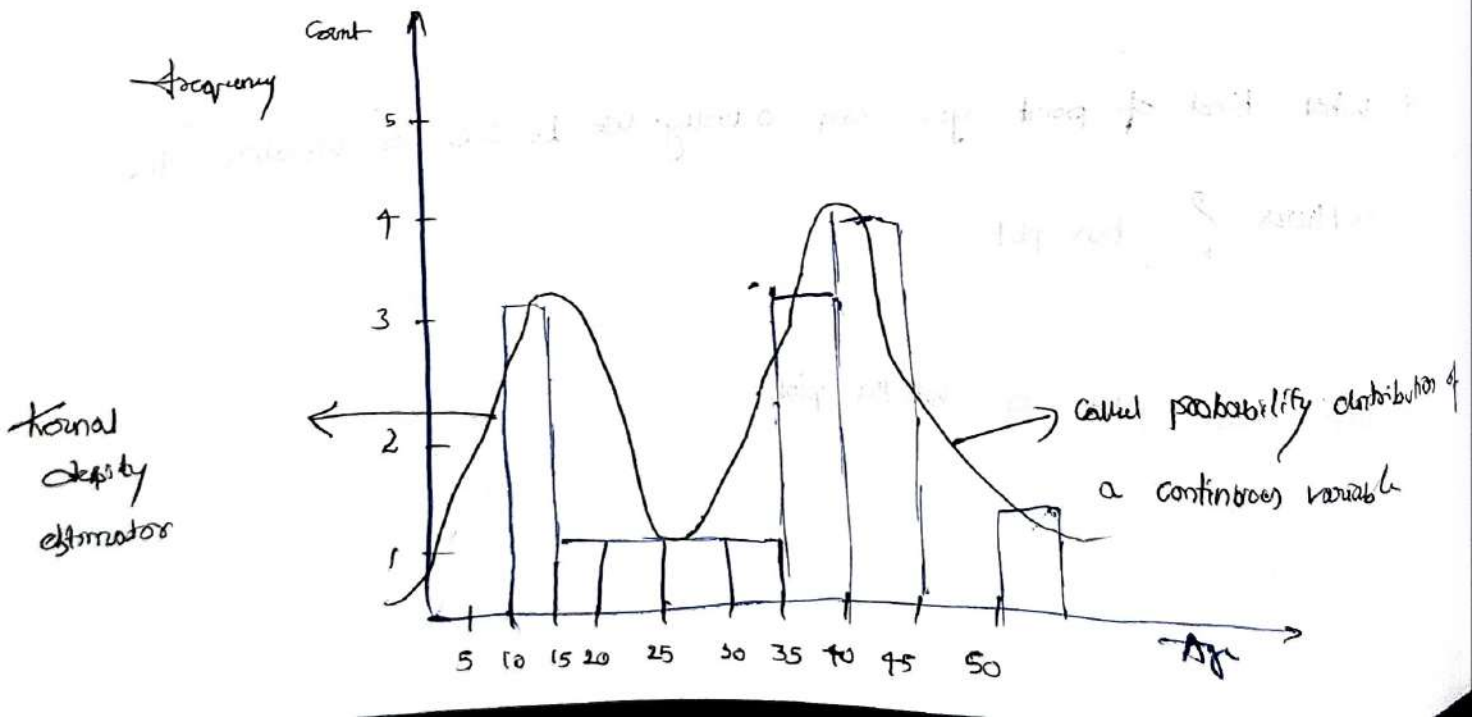
⇒ Histograms

① Bin refers to a interval or range of values in to which data points are grouped

0 - 50

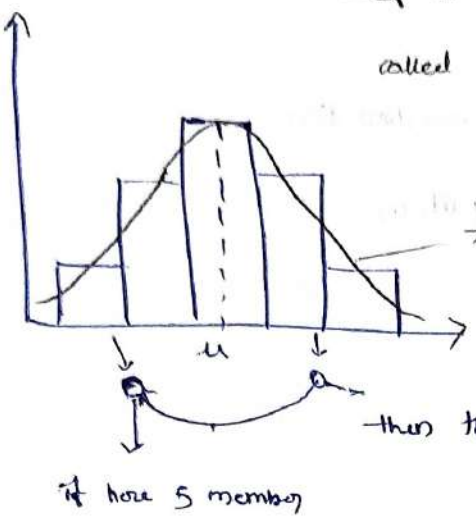
① No. of bins = 10 ⇒ $\frac{0-50}{10} = 5$ bin size.

Bin → [0-5, 5-10, 10-15 . ---- 45 - 50]



Skewness:-

* If a distribution is having perfect bell curve than it is called as Normal or Gaussian distribution.



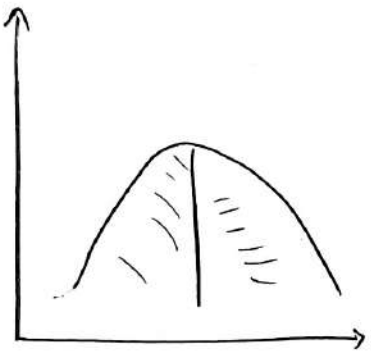
→ perfect bell curve

then there 5 members

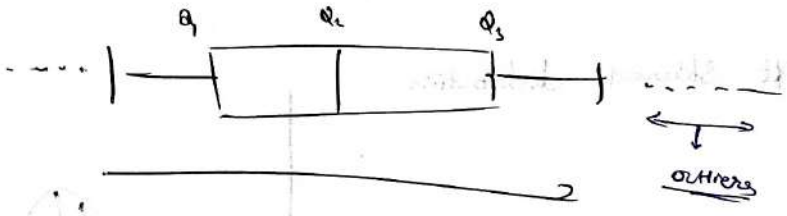
If here 5 members

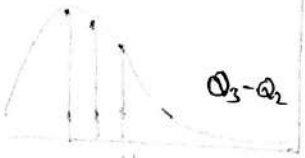
It is Symmetrical distribution & it has no skewness, perfect in shape.

50% of PD oneside & 50% of PD otherside



Box plot



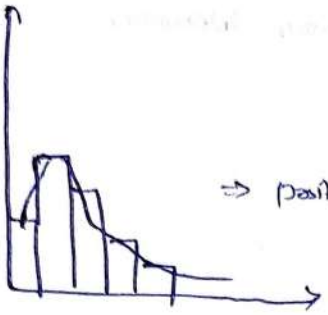


$Q_3 - Q_2 \approx Q_2 - Q_1$

(x) The mean, median, mode will are all perfectly at the centre

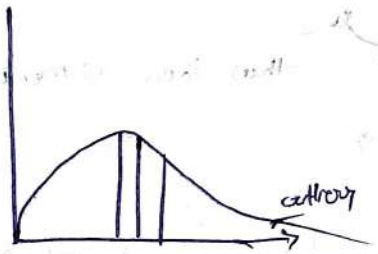
Mean = Median = Mode

② Right Skewed :-



⇒ positive

Log Normal Distribution have same pattern

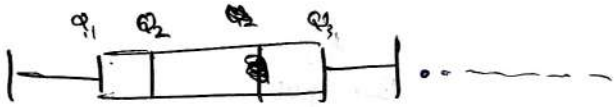


mean is mostly impacted

$$\text{mean} \geq \text{median} \geq \text{mode}$$

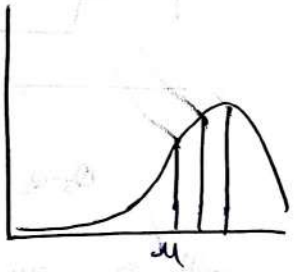
$$Q_3 - Q_2 \geq Q_2 - Q_1$$

Box plot



=

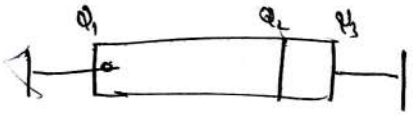
③ left skewed distribution



⇒ negative skewed

Relation

$$\text{mean} \leq \text{median} \leq \text{mode}$$



$$Q_2 - Q_1 \geq Q_3 - Q_2$$

Covariance and Correlation :-

Covariance & Correlation are two statistical measures used to determine the relationship between two variables. Both are used to understand how changes in one variable are associated with change in another variable

Covariance :-

Covariance is a measure of how much random variables change together. If the variables tend to increase & decrease together, the covariance is positive. If one tends to increase when the other decreases, the covariance is negative

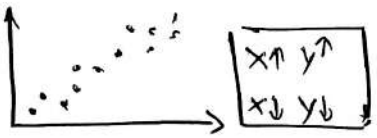
x	y
2	3
4	5
6	7
8	9

[Quantify the relationship between x & y]

like

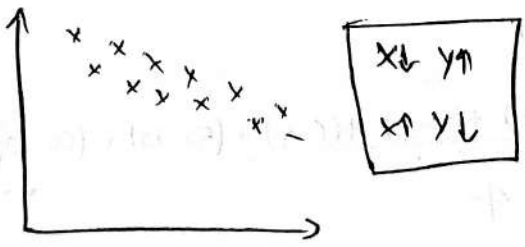
x↑	y↑
x↓	y↓
x↑	y↓
x↓	y↑

↑↓ Size of House	price ↑↓
1200	45
1300	50
1500	50



=> positive covariance

x	y
7	10
6	12
5	14
4	16



Negative covariance

Covariance

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad cov(x,x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

$$\boxed{cov(x,x) = Var(x)} \qquad \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

x_i → Datapoint of RV X

$\bar{x}$ → Sample mean of X

y_i → Datapoint of RV y

$\bar{y}$ → Sample mean of y

Student:

Hour Studied (x)	Exam Score (y)
2	50
3	60
4	70
5	80
6	90

↑ ↑ ↓ ↓ +ve

$$\bar{x} = \frac{2+3+4+5+6}{5} = 4$$

$$\bar{y} = \frac{50+60+70+80+90}{5} = 70,$$

$$cov(x,y) = \frac{(2-4)+(3-4)+(4-4)+(5-4)+(6-4)+(50-70)+(60-70)+(70-70)+(80-70)+(90-70)}{4}$$

$$= 20$$

This +ve covariance indicate the no of hours studied increased the exam score

x	y
7	10
6	12
5	14

x↑ y↓ => -ve

Advantages :

1) Quantify the relationship between x & y

Disadvantages

1) Covariance does not have a specific limit value

$Cov(x,y) \Rightarrow -\infty$ to ∞

-1 to 1 limits 0.96 & 0.98

2) Correlation :
- Pearson Correlation Co-efficient
- Spearman Rank Correlation

1) pearson Correlation co-efficient => [-1 to 1]

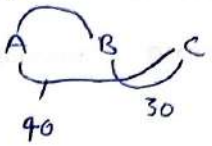
$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

=> it limits to -1 to 1 & we can say which is Strong

- The more the value towards +1 the more +ve correlated x & y
- The more the value towards -1 to more -ve correlated x & y

20 Strong Cov(A,B)

30 Cov(B,C) if not have limit



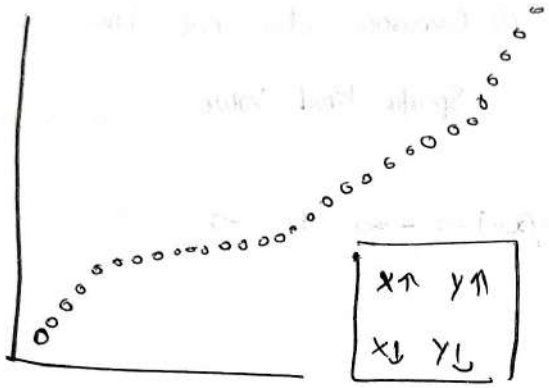
we cannot say which is

+300 or -300

(ii) Spearman rank Correlation:-

Spearman correlation = 1

Pearson correlation = 0.88 (because it is non linear)



$$r_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$

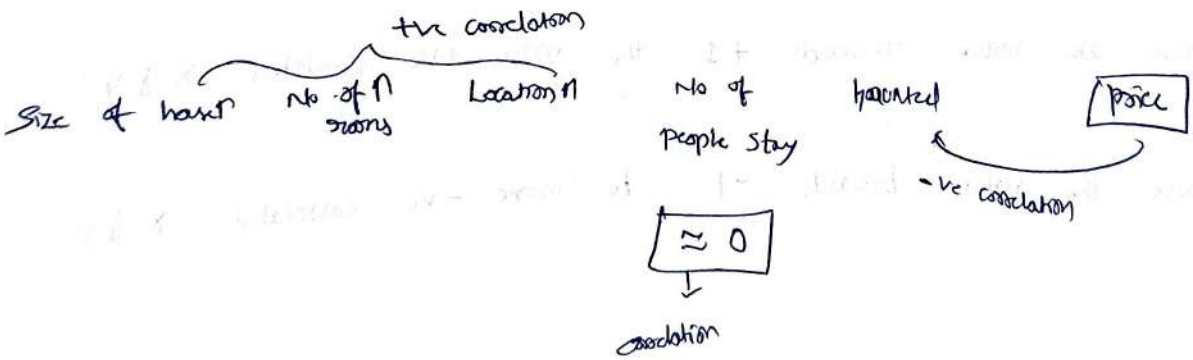
x	y	R(x)	R(y)
1	2	2	1
3	4	3	2
5	6	4	3
7	8	5	5

Spearman correlation of 1 results when 2 variables being compared are Monotonically related, even if relation is not linear.

⇒ x↑ → y↑

x↓ → y↓

Used in feature selection:-



Probability

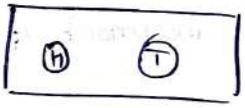
1. Introduction
2. Addition Rule (For mutually exclusive event)
3. Addition Rule (For non mutually exclusive event)
4. Multiplication rule (Independent & Dependent Event)

1. **Probability:-** It is about determining the likelihood of an event

Eg:
- Toss a coin
  $P(H) = 1/2$ $\{H,T\}$
  $P(T) = 1/2$
- Rolling a dice $\{1,2,3,4,5,6\}$
  $P(x=1) = 1/6$

Mutual exclusive Event:-

Two event are mutual exclusive if they cannot occur at the same time



$P(H \text{ or } T) = P(H) + P(T)$ {Addition rule for mutual exclusive event}

$= 1/2 + 1/2 = 1$

Eg: Rolling a die $\{1,2,3,4,5,6\}$

$P(1 \text{ or } 5) = P(1) + P(5)$

$= 1/6 + 1/6 = 2/6 = 1/3$

* Non Mutual Exclusive Events:-

Two event are mutually exclusive if they can occur at the same time

Eg:- Taking a card from deck

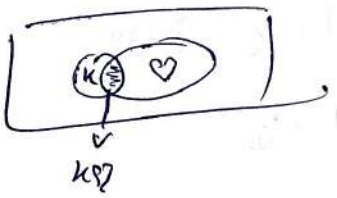
[k] again [k] can get

Non mutual exclusive event

$P(k \text{ or } \heartsuit) = P(k) + P(\heartsuit) - P(k \cap \heartsuit)$

$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$

$= \frac{16}{52}$



Multiplication rule :- {independent & dependent events}

2 Events are independent if they do not affect one another

Eg:- Tossing a coin {H & Tails}

$P(H) = 1/2$ & $P(T) = 1/2$

Eg:- Rolling a dice

Dependent events :

2 events are Dependent if they effect each other

Eg:- Take a ring from a clock & then Taking a Queen card

$$P(K) = \frac{1}{52} \qquad P(Q) = \frac{1}{51}$$

Multiplication rule:

① Indepent event {Tossing a coin}

$$P(H \text{ and } T) = P(H) * P(T)$$

$$= 1/2 * 1/2$$

$$= 1/4$$

② Dependent Event

$$P(K \text{ and } Q) = P(K) * P(Q/K)$$

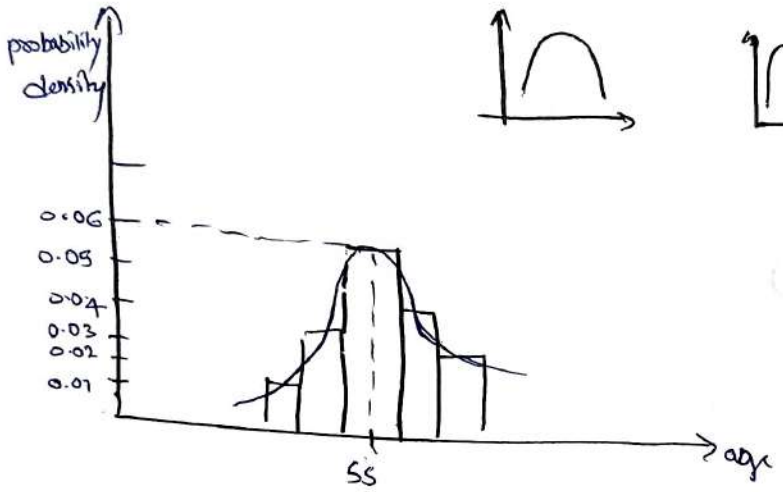
(-> Conditional probability)

$$= \frac{4}{52} * \frac{4}{51}$$

Probability Distribution function :-

Probability distribution function → helps us to understand How describe how the probabilities are distributed over the values of a random variable

Ages = {... } → continuous random variable



Two main types of probability distribution function

① Probability Mass function (PMF) :- used for discrete

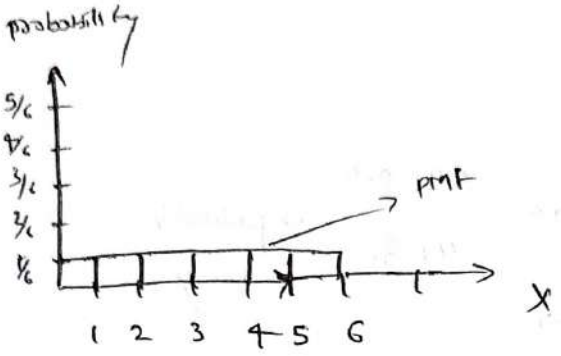
② Probability Density function (PDF) :- used for continuous

① Probability Mass function : [Discrete Random variable]

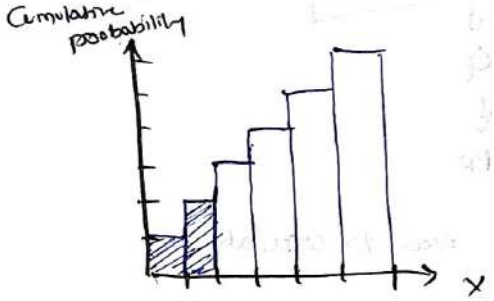
Eg :- Rating a dice {1, 2, 3, 4, 5, 6} => fair dice

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

PMF:



Cumulative Density function:-



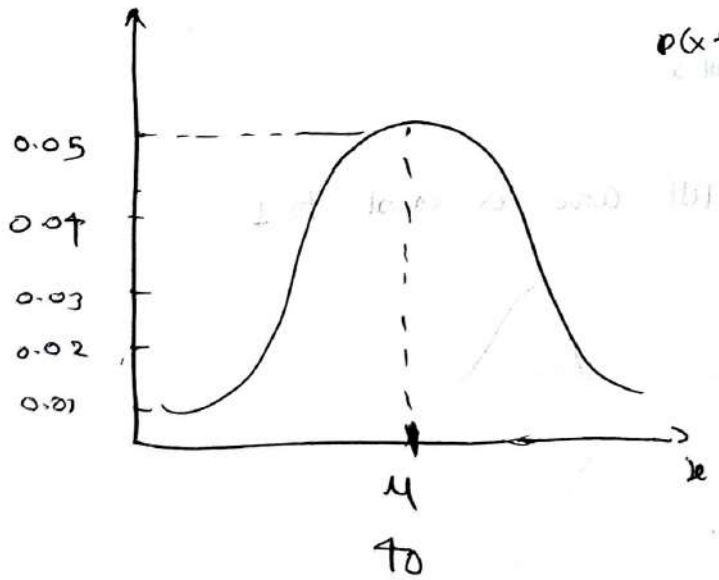
$P(X \leq 2) = P(X=1) + P(X=2)$

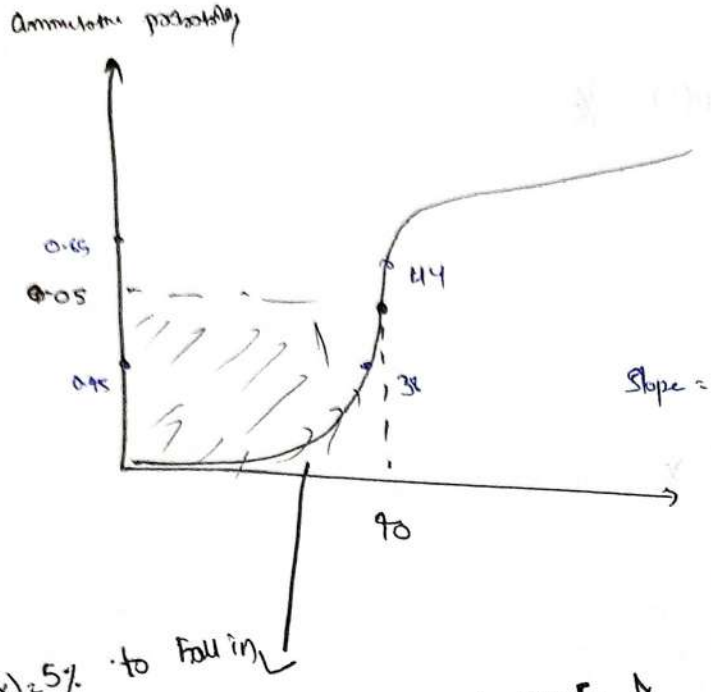
$= 1/6 + 1/6 = 1/3$

Cumulative density function we need to combine all the probabilities as we go from 1 to 6

② Probability Density function :- [Distribution of continuous random variables]

$P(X = 40) = 0.05$ (X)





$$\text{Slope} = \frac{0.65 - 0.45}{44 - 38} \Rightarrow \text{probability density}$$

$P(x) = 5\%$ to fall in this area. This area is given by PDF

$$\text{Slope} = \frac{x_2 - x_1}{y_2 - y_1}$$

$\downarrow$ G

$\downarrow$ PD

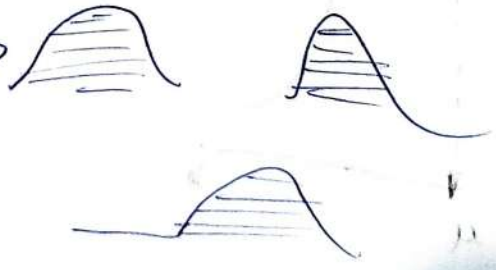
In order to calculate probability density function we need to calculate Gradient of CPF / slope of that point

Probability density function = Gradient of Cummulative Density function

Properties

1. Non negative $f(x) \geq 0$ for all x
2. The total area under the Pdf curve is equal to 1

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



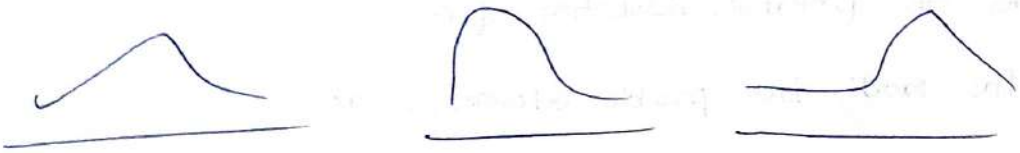
With respect to different distribution

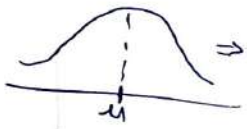
f(x) function is going to change

Different types of distribution:-

Types of Probability Distribution (pdf, pmf, cdf)

Age, weight, Salary



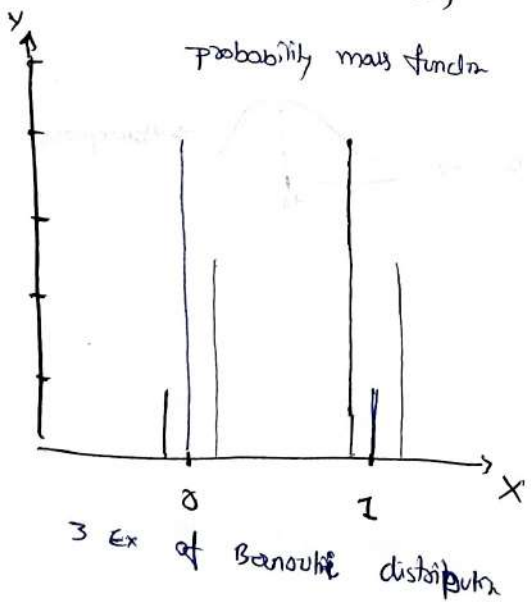
1. Bernoulli Distribution:- outcomes are binary (pmf) => Discrete random variable
2. Binomial Distribution → pmf
3. Normal / Gaussian Distribution → pdf ⇒  ⇒ Assumptions
4. Poisson Distribution pmf
5. Log Normal Distribution pdf
6. Uniform Distribution pmf

Dataset → HPPD

Size of hous	No. of rooms	Location	Floor	sea side	price
CRV	Discrete		3 Discrete pmf	0 & 1 pmf	Continous pdf

Bernoulli Distribution :-

The Bernoulli's distribution is the simplest discrete probability distribution. It represents the probability distribution of a random variable that has exactly two possible outcomes: Success (with probability p) and failure (with probability 1-p). It is used to model binary outcomes, such as a coin flip or yes/no question.



3 Ex of Bernoulli distribution

$P(x=0) = 0.2$ & $P(x=1) = 0.8$

$P(x=0) = 0.8$ & $P(x=1) = 0.2$

$P(x=0) = 0.5$ & $P(x=1) = 0.5$

① Discrete Random Variable

② Outcomes are Binary

Ex: Student will pass/fail

$P(x = \text{pass}) = 0.4$

$P(x = \text{fail}) = 1 - 0.4 = 0.6$

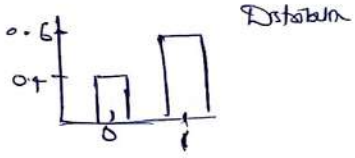
Parameters: $0 \leq p \leq 1$

$q = 1 - p$

Support $k \in \{0, 1\}$ => 2 outcomes

PMF $k = 1$ pass = 40% $\Rightarrow p$

fail = 60% $\Rightarrow q$



Distribution

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

$PMF = p^k * (1-p)^{1-k}$

if $k = 1$

$P(k=1) = p^1 (1-p)^{1-1} = p$

if $k = 0$

$P(k=0) = p^0 (1-p)^{1-0} = 1 - p = q$

Mean:-

The expected value of a Bernoulli random variable X is

$E[x] = p$

$k = \{0, 1\}$

$p = 0.6$

$q = 0.4$

$$E[x] = \sum_{i=0}^{1} k \cdot p(k)$$

$= 0 * 0.40 + 1 \times 0.60$

$= 0 + 0.6$

$= 0.6 = q$

$= p$

~~Mode~~

~~P>q~~ ~~p will~~

Median of Bernoulli

$$\text{Median} \begin{cases} 0 & \text{if } p < 1/2 \\ (0,1) & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases} \qquad \begin{cases} \text{median} = 0 & \text{if } q > p \\ = 0.5 & q = p \\ = 1 & q < p \end{cases}$$

Mode

$p > q \Rightarrow p$ will be the mode

~~q~~ q will be the mode

* Variance

(K = 0, 1)

$E[X^2] = P(X=K) \cdot K^2$ ⟶

$= P(X=1) \cdot 1^2 + P(X=0) \cdot 0^2$

$= p \cdot 1^2 + q \cdot 0^2$

$= p = E[K]$

$\text{Var}[X] = E[X^2] - E[X^2]$

$= E[X] - E[X^2]$

$= p - p^2$

$= p(1-p) = pq$

$\sigma^2 = 0.40 \times (0-0.6)^2 + 0.6(1-0.6)^2$

$= 0.40 \times 0.36 + 0.6(0.16$

$= 0.24$

$= pq$

$\sigma^2 = pq$

$\sigma = \sqrt{pq}$

Binomial Distribution :- ⟶ It is basis for Binomial test of statistical significance

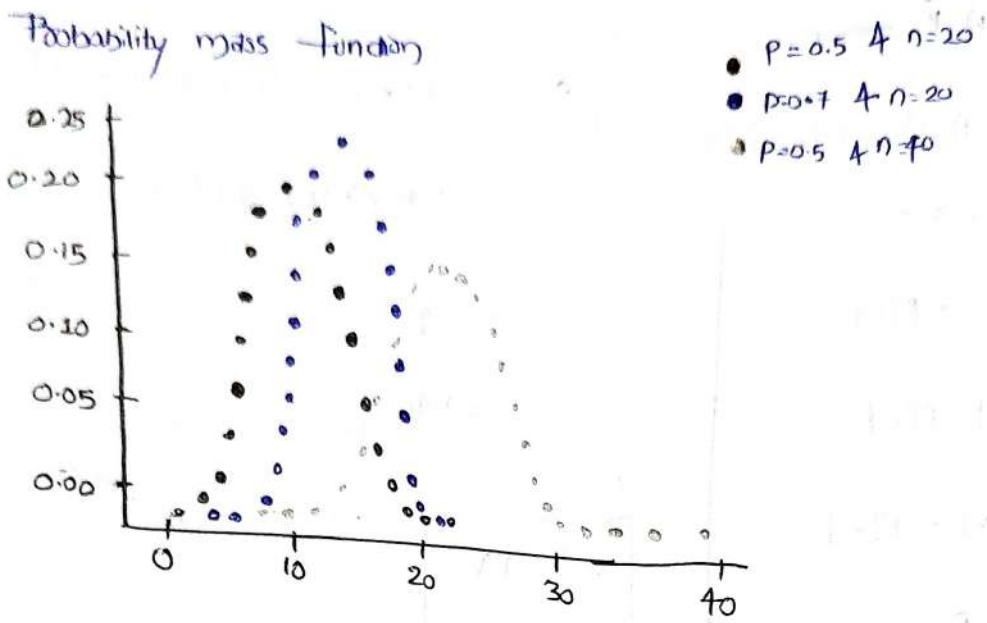
In probability theory & statistics, the binomial distribution with parameter n & p is the discrete probability distribution of no of success in a sequence of n independent experiment, each asking Yes/No Questions and each with its own Boolean-valued outcomes: success (p) or failure (q)

⇒ A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment & a sequence of outcomes are called a Bernoulli process

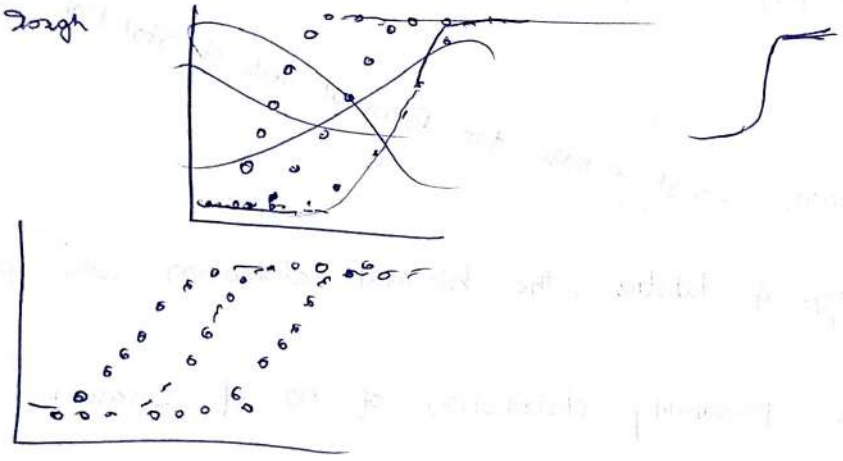
⇒ for single trial Binomial distribution is a bernoulli distribution.

Binomial distribution

Probability mass function



Cumulative distributive function.



* Discrete random variable

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

Eg: Tossing a coin 10 times $n = 10$

{H, T}

Notation

$B(n,p)$

parameters :- $n \in \{0,1,2,\ldots\}$ => no of trails or Experiment

$p \in [0,1]$ -> Success probability for each trial

$q = 1-p$

Support :- $k \in \{0,1,2,3\ldots n\}$ => no of Successes

PMF $P(k,n,p) = {}^{n}C_{k}\, p^{k} (1-p)^{n-k}$

For $k = 0,1,2,\ldots n$ where

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

=> Binomial Co-efficient

Mean :- $n \cdot p$

Variance :- npq

σ : $\sqrt{npq}$

Eg:- ① Coin flip

No. of trials (n) = 5

probability of success (p) = 0.5

No. of success (k) = varies from 0 to 5

② what is the probability of getting exactly 3 heads in 5 flips?

$n = 5 \quad k = 3$

$P(X=3) = {}^5C_3 (0.5)^3 (1-0.5)^{5-3}$

$= 0.3125$

Example:-

Scenario:- Inspecting 10 items in a factory where each item has a 10% chance of being defective

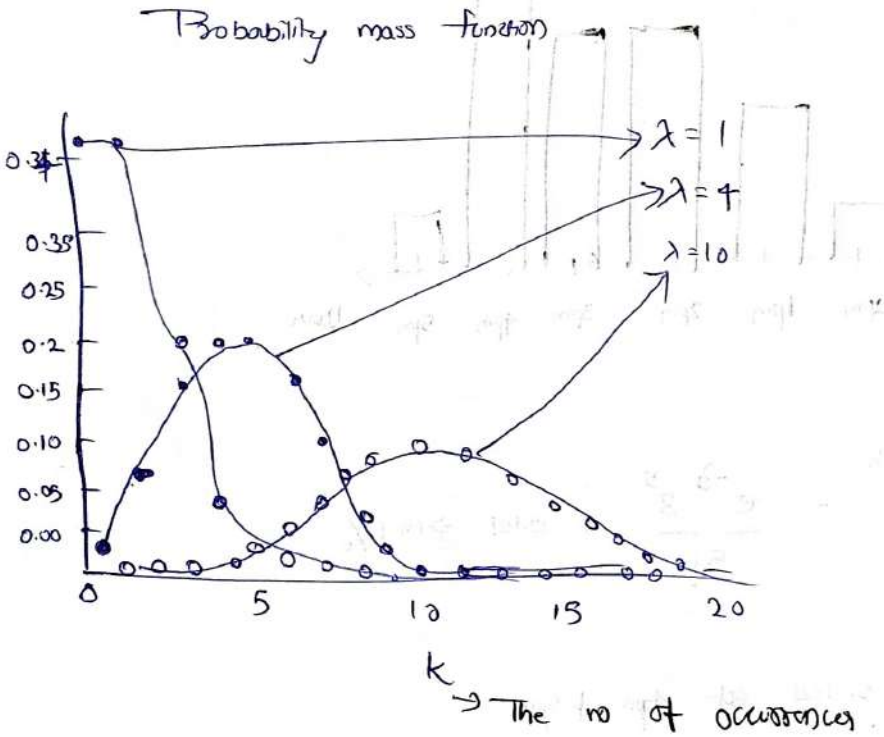
* No. of Trails (n) = 10
* Probability of Success (p) = 0.1 (defective item)
* No of Successes (k) = varies from 0 to 10

Q. what is the probability of find exactly 2 defective items in a sample of 10?

$P(X=2) = {}^{10}C_2 (0.1)^2 (1-0.1)^{10-2} \approx 0.1937$

Poisson distribution :-

In probability theory & statistics, the poisson distribution is a discrete probability distribution that expresses the probability of a given no of events occuring in a fixed interval of time if these events occurs with a know constant mean rate & independently of the time since the last event



λ is the expected rate of occurences

y-axis is probability of k occurences given λ.

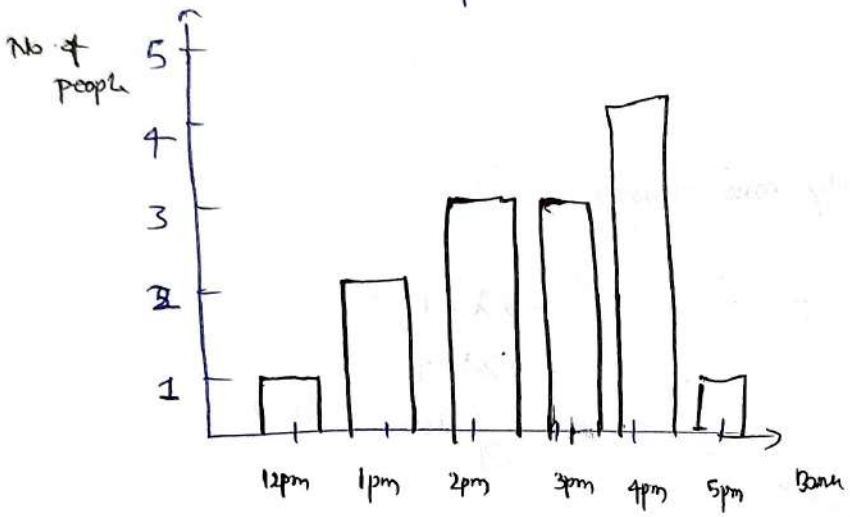
The function is defined only at integer values of k.

① Discrete random variable (PMF)

② Describe the no of events occuring in a fixed time intervals

Eg: No of people visiting hospital every hour

No of people visiting banks every hour



$$P(x=5) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-3}3^{5}}{5!} = 0.101 \Rightarrow 10.1\%$$

How many people visited at 4pm & 5pm

$$P(x=4) + P(x=5)$$

Mean of poisson Distribution :-

mean Absolute deviation $= E[x-\lambda] = \frac{2\lambda^{[\lambda]+1}e^{-\lambda}}{[\lambda]!}$

mean & variance $= \lambda$

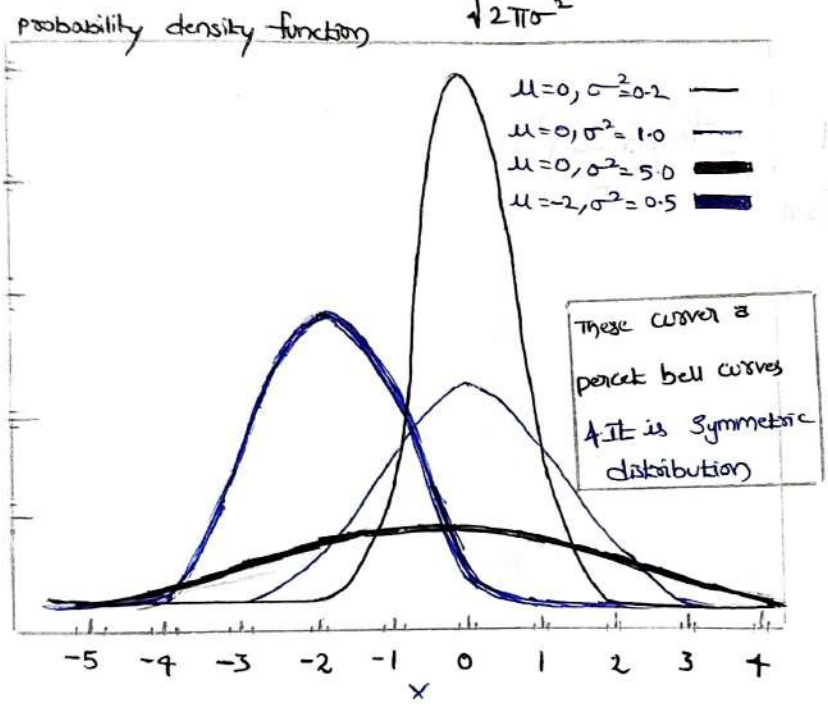
$\lambda \times t$?

Normal / Gaussian Distribution

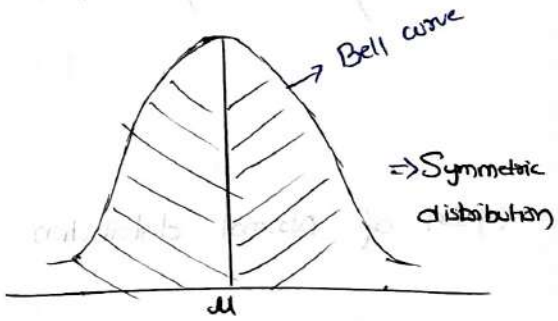
In probability theory and statistics, a normal distribution or Gaussian Distribution is a type of continous probability distribution for a real-valued random variable. The general form of its probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

* Continous random values (PDF)



* Symmetric distribution



mean (μ) = median = mode

Eg: IRIS Dataset

It usually follows normal distribution

4 imp feature it has → Petal Length, Sepal length

Petal width, Sepal width

Based on this 4 features, What flower it is

Ex (2) Weight of Student in a class

Ex (3) Height of Students in a class

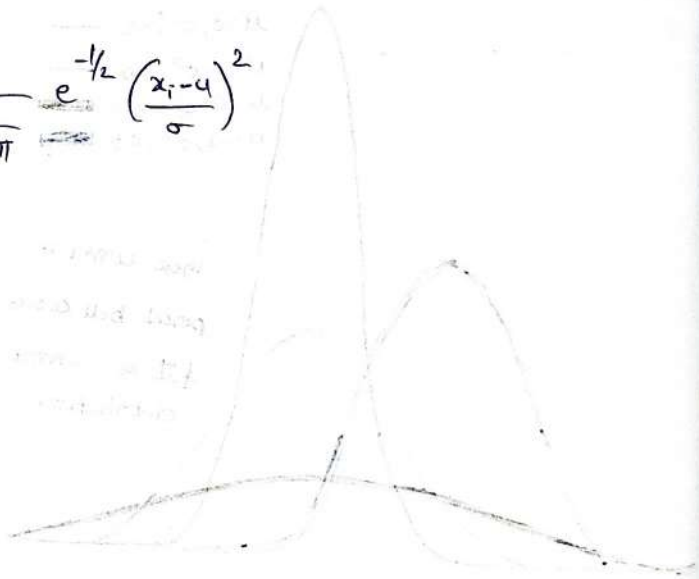
Notations :- $N(\mu, \sigma^2)$

parameters :- $\mu \in R$ = mean

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

$x \in R$

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



Mean of Normal distribution

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

Variance :-

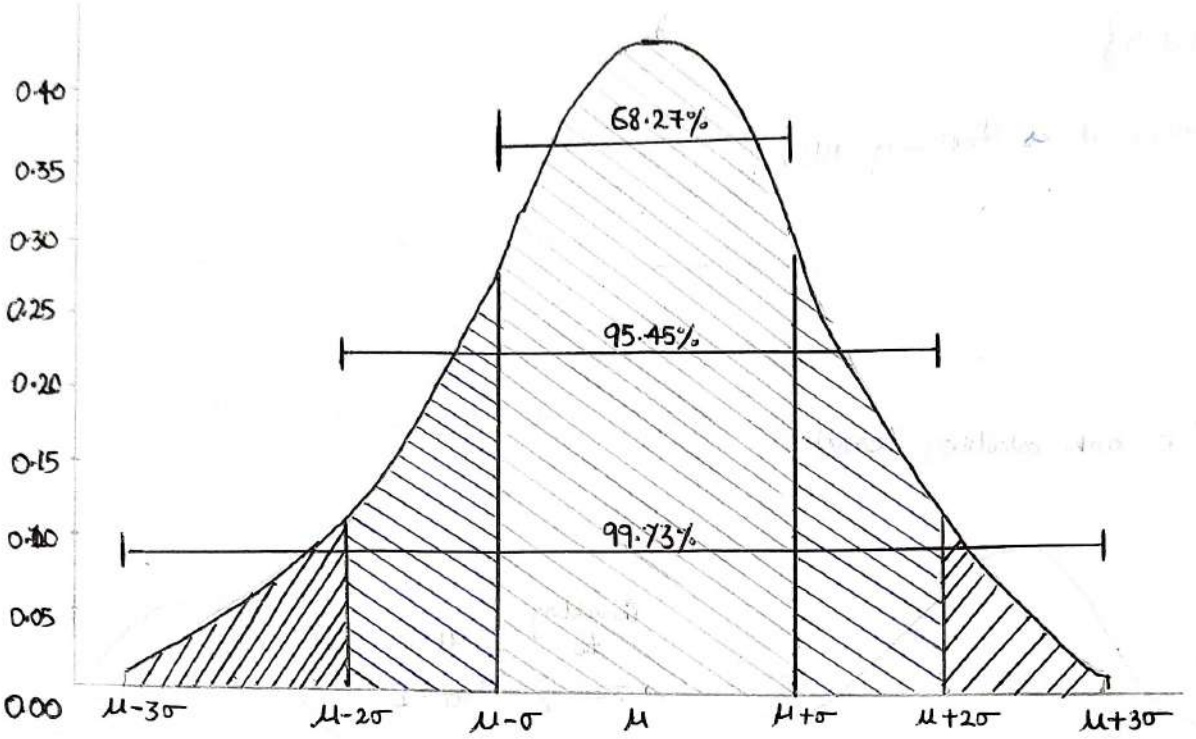
$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{n}$$

Standard deviation (σ) = $\sqrt{\text{variance}}$

$= \sqrt{\sigma^2}$

$= \sigma$

Emperical Rule of Normal / Gaussian Distribution



$X = \{x_1, x_2, x_3 \dots x_n\}$ Assumpe it follows Gausian distribution

* QQ plot is used to determine wheather a random variable X will follow a Normal / Gausian distribution or not.

Probability :-

$P(\mu-\sigma \leq X \leq \mu+\sigma) \approx 68\%$

$P(\mu-2\sigma \leq X \leq \mu+2\sigma) \approx 95\%$

$P(\mu-3\sigma \leq X \leq \mu+2\sigma) \approx$ 99.7%

1 Standard normal Distribution { Any distribution having $\mu = 0$ & $\sigma = 1$ }

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

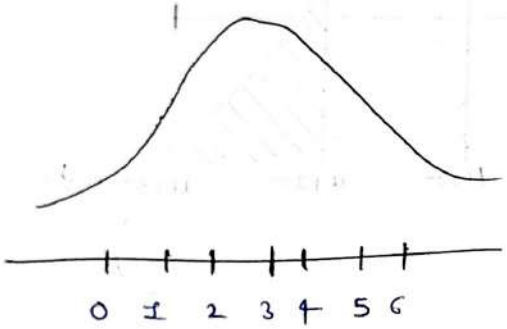
↓

Consider it as following ND

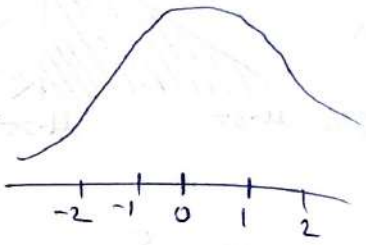
$\mu = 3$

$\sigma = 1.414$

≈ 1 (To make calculation easy)



Converting to ⟹ $\mu = 0$, $\sigma = 1$



* Normal Distribution is converted such that $\mu = 0$ & $\sigma = 1$ then the distribution is called standard normal distribution

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

* We uses $Z\text{-score} = \frac{x_i - \mu}{\sigma}$ for converting

↓

$y = \{-2, -1, 0, 1, 2\}$

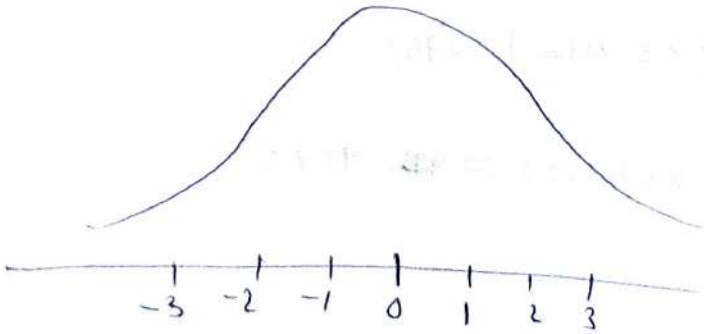
① $\frac{1-3}{1} = -2$

② $\frac{2-3}{1} = -1$

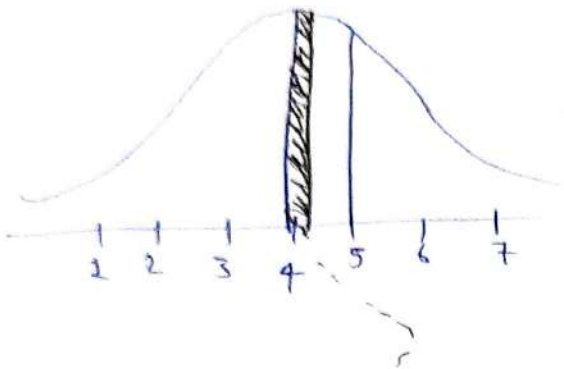
③ $\frac{3-3}{1} = 0$

④ $\frac{4-3}{1} = 1$

⑤ $\frac{5-3}{1} = 2$



$X \approx SND \quad (\mu = 0, \sigma = 1)$



→ how many standard deviations 4.25 is away from the mean

$x_i = 4.25$

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

Eg: Dataset

Age (years)	weight (kg)	Height (cms)	Salary (INR)
24	70	175	40k
25	60	160	50k
25	90	170	35k
26	35	163	60k
20	70	175	45k
31	90	180	65k

- In features units will be different, we try to bring all the feature in a same unit

↓

we use Standardization

$$Z\text{-score} = \frac{x_i - \mu_{age}}{\sigma}, \frac{x_i - \mu_{weight}}{\sigma}$$

Uniform Distribution :-

① Continuous uniform Distribution

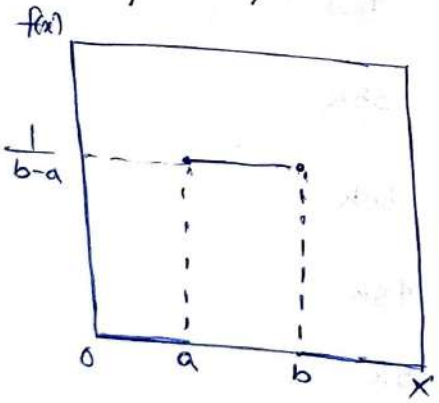
② Discrete uniform Distribution

① Continuous uniform Distribution (CRV)

In probability theory and ~~distribution~~ statistics, the Continuous uniform distributions or rectangular distributions are a family of symmetric probability distributions. Such a 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 the minimum and maximum values.

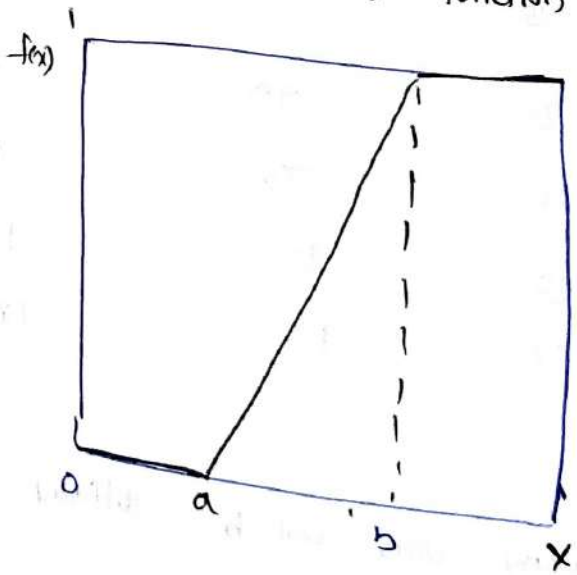
Continous uniform :-

Probability density function



using maximum convention

Cummulative distribution function



Notation :- $U(a,b)$

parameters :- $-\infty < a < b < \infty$

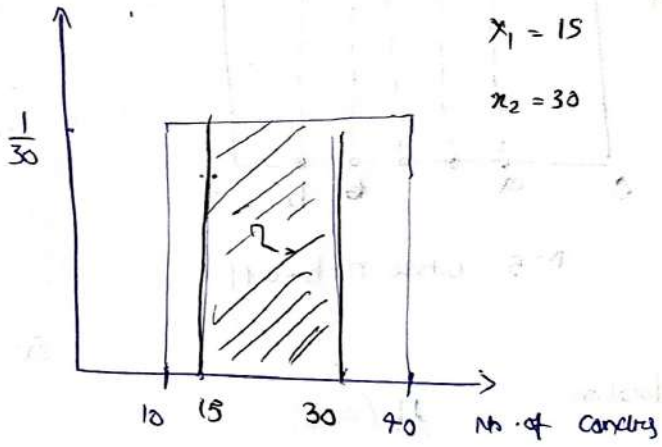
$$\text{Pdf} = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

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

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

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

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



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

(2) Probability of daily sales to fall between 15 and 30 ?

$$P(15 \le X \le 30) = (x_2 - x_1) \times \frac{1}{b-a} = (30-15) \times \frac{1}{30} = 0.5 //$$

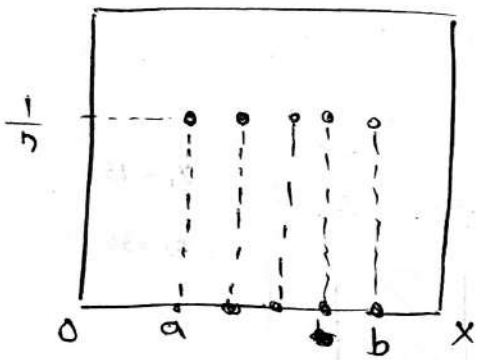
$$P(x \ge 20) = (40-20) \times \frac{1}{30} = 0.66 = 66\%$$

② Discrete uniform Distribution :-

In probability theory & statistics, the discrete uniform distribution is a symmetric probability distribution wherein a finite no of values are equally likely to be observed; every one of n values has equal probability equally likely to happen

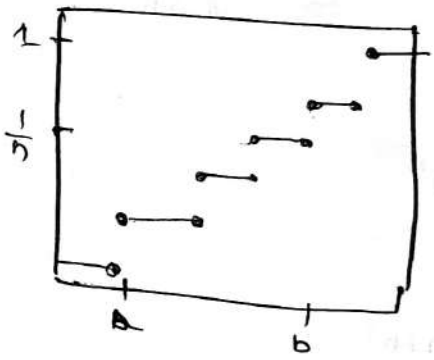
Discrete uniform

PMF



n=5 where n=b-a+1

Cummulative distribution function



Notation :- $\mathcal{U}(a,b)$

parameter : a, b where $b \geq a$

PMF :- $1/n$

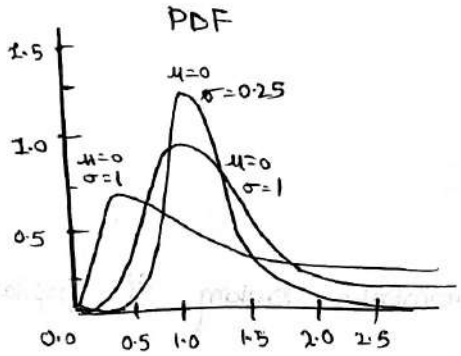
mean, median $\rightarrow \frac{a+b}{2}$

Ex:- Rolling a dice $\{1,2,3,4,5,6\}$ $P(1) = 1/6$ ⋮ $P(6) = 1/6$

Log Normal Distribution :-

In probability theory, a log-normal (or lognormal) distribution is a continuous probability distribution of a 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.

Log-normal D * It is a right skewed distribution



$x \approx$ log Normal distribution (μ, σ)

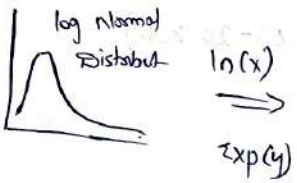
$y \approx \ln(x)$ → Normal distribution

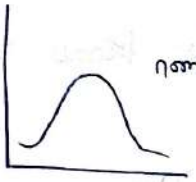
↓

Natural log

[$\log_e$]

$x \approx \exp(y)$ => log normal distribution



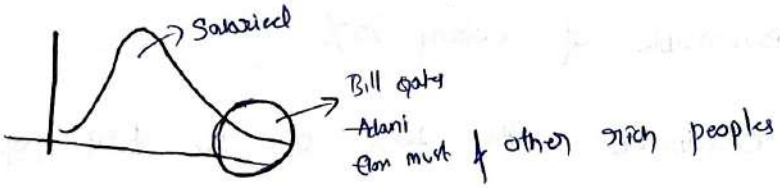


checked using Q-Q plot

↑

normal distribution

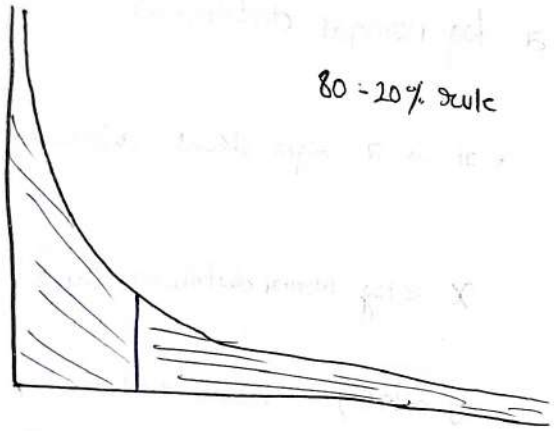
Ex:- ① Wealth distribution of the world



② Discussion forum → length of the comments

Power Law Distribution:

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



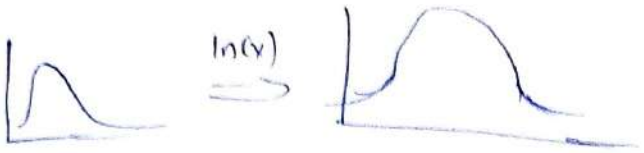
An example power-law graph that demonstrates ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as 80-20 rule)

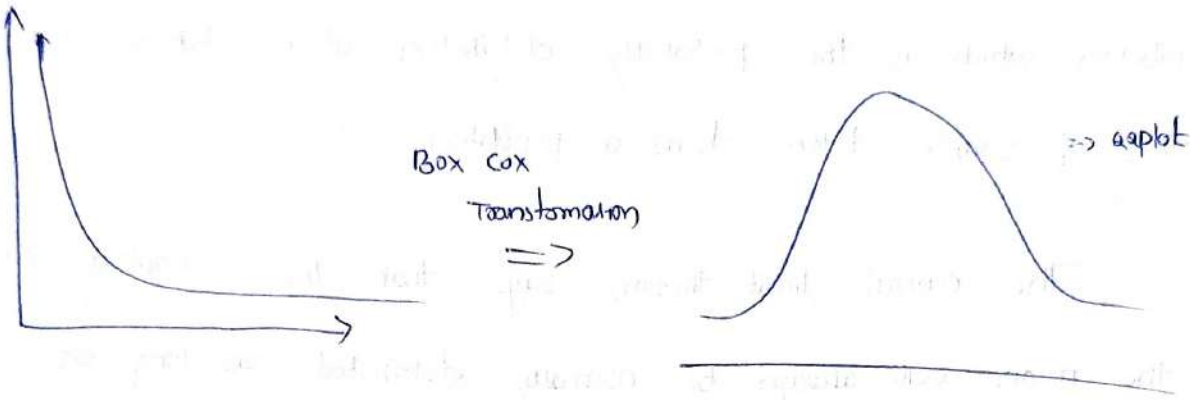
Eg:- IPL

① 20% of Team is responsible of wining 80%

② 80% of wealth are distributed with 20% of the total population

③ 80% of the total oil is with 20% of the nation





The data which follows power law distribution we call it as pareto distribution.

Pareto Distribution :- Eg: 80% of the entire project is done by 20% of the team

* In statistics, a power law is a

The pareto distribution, named after the Italian civil engineer, economist & Sociologist Vilfredo, pareto is a power-law PD that is used in description of social, quality, control, scientific, geophysical, actuarial & many other types

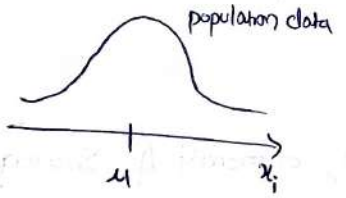
The principle originally applied to describing the distribution of wealth in a society, fitness the trend that a large portion of wealth is held by a small fraction of the population.

① Central Limit Theorem :-

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

The central limit theorem says that the sampling distribution of the mean will always be normally distributed as long as the sample size is large enough. Regardless of whether the population has a normal, poisson, binomial or any other distribution, the sampling distribution of mean will be normal

① $X \approx N(\mu, \sigma)$ n = sample size => any value



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

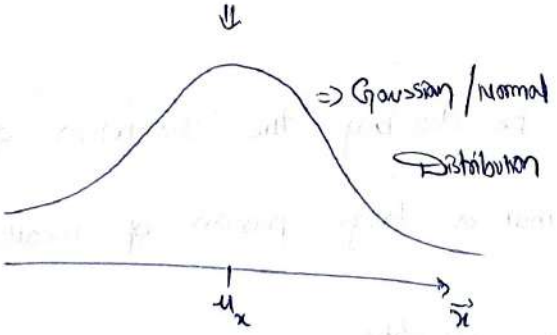
$S_2 = \{x_2, x_3 \ldots x_n\} = \bar{x}_2$

$\vdots$

$S_n = \bar{x}_n$

} Sample means

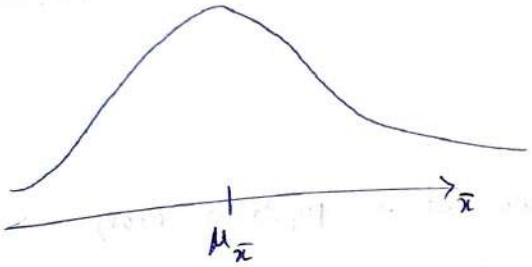
Sample distribution of means



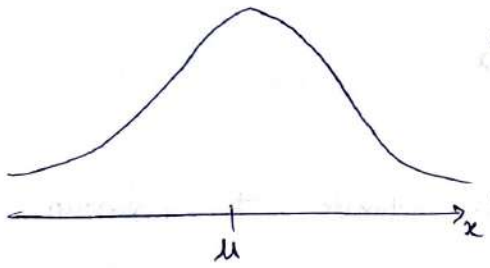
2. $X \not\approx N(\mu, \sigma)$

$S_1 \rightarrow \bar{x}_1$
$S_2 \quad \bar{x}_2$
$\vdots \quad \vdots$
$S_n \quad \bar{x}_n$

$n \geq 30$ sample size



① Normal Distribution



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

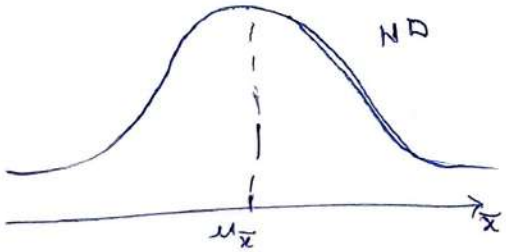
$\Downarrow$

σ = population std

μ = population mean

n = Sample size

Sampling Distribution of mean



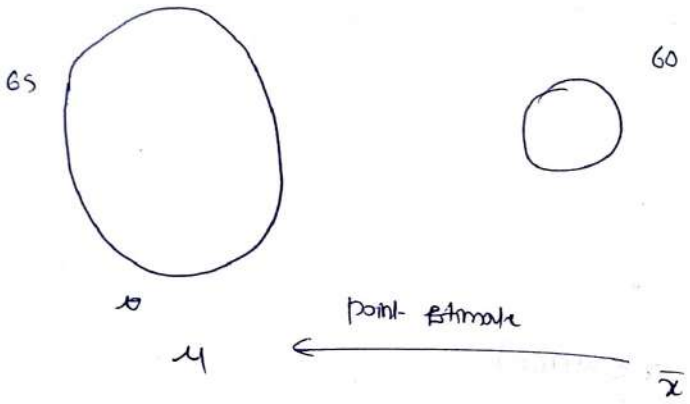
$\Rightarrow$ $x = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Estimate :- It is a Specified observed numerical value used to estimate an unknown population parameter

Types:-

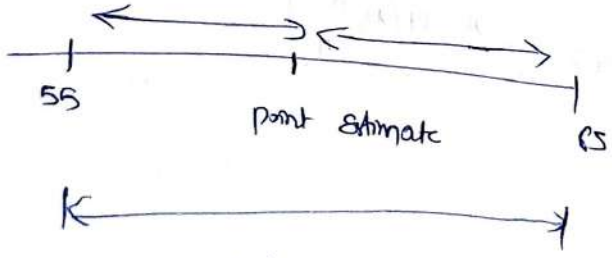
① point Estimate :- Single numerical value used to estimate an unknown population parameter

Eg :- Sample mean is a point estimate of a population mean



② Interval Estimate :- Range of values used to estimate the unknown population parameter

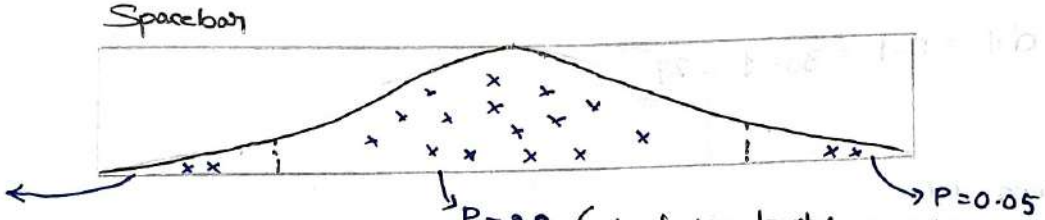
55-65 => Sample mean



confidence interval.

P-Value : The P-value is a number, calculated from a statistical test that describes how likely you are to have found a particular set of observations if null hypothesis were true. P values are used in hypothesis testing to help decides wheather to reject the null hypothesis.

Example 1 :- P-value giving the probability of spacebar touch at particular region



P=0.05 (out of 100 times, a person may touches 5 times at this region

P=0.9 (out of 100 touches, a person may touches 90 times at this region)

Example 2 :- Coin is fair or not

① Null hypothesis : H_0 : coin is fair

② Alternate hypothesis : H_1 : coin is not fair

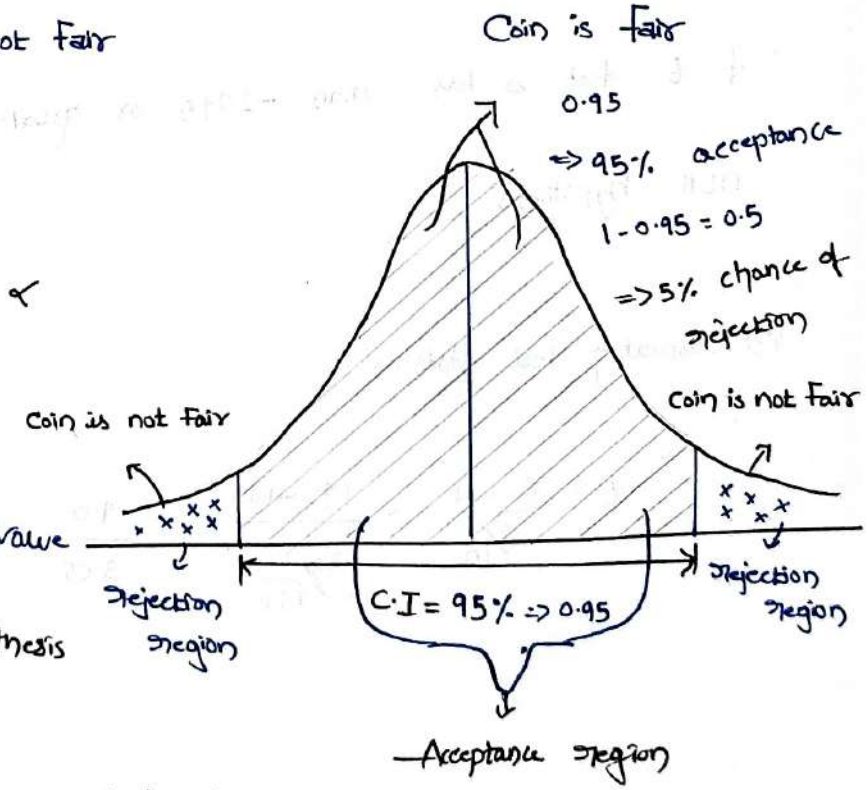
③ Experiment :- 100 times

④ Significance value :- $\alpha = 0.05$

Confidence interval = $1 - \alpha$

$C.I = 1 - 0.05$

$= 0.95$



⑤ Conclusion :- If $p < \alpha$ Significance value

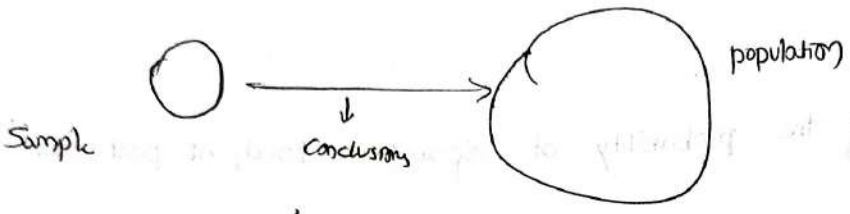
Reject the null hypothesis

else

Fail to reject the null hypothesis

A hypothesis and hypothesis Testing Mechanism :-

Inferential Stats :- Conclusion or Inference



Hypothesis testing

person → crime → court

Hypothesis Testing Mechanism :-

① Null hypothesis (H_0) - person is not guilty

- The assumption you are begining with

② Alternative hypothesis (H_1) : The person is guilty

- opposite of Null hypothesis

③ Experiments → Statistical Analysis {P value, Significance value}

→ Collect proof (DNA, finger test)

④ Accept the Null hypothesis or Reject the NULL hypothesis

Ex:- Colleges at Distaics A states its average passed percentage of students are 85%. A new college opened in the distaict and it was found that a sample of student 100 have a pass percentage of 90% with a standard deviation of 4%. Does college have a different passed percentage ?

Ans Null hypothesis (H_0) = $\mu = 85\%$

Alternate hypothesis (H_1) = $\mu \neq 85\%$

Hypothesis And Statistical Analysis :-

① Z test
② t test
} ⇒ data deling with average Ex :- Sample average

Z-table
t-table

③ CHI SQUARE ⇒ Statistical analysis of categorical data

④ ANNOVA ⇒ Variance of data

Z-test :-

⊛ When to use Z-test ?

If ⅰ) population std is know & ⅱ) $n \geq 30$

② The average heights of all residents in a city is 168 cm with a $\sigma = 3.9$ A doctor believes the mean to be different. He measured the height of 36 individuals & found the average height to be 169.5 cm

(a) State null & Alternate hypothesis

(b) At a 95% confidence level, is there enough evidence to reject the null hypothesis

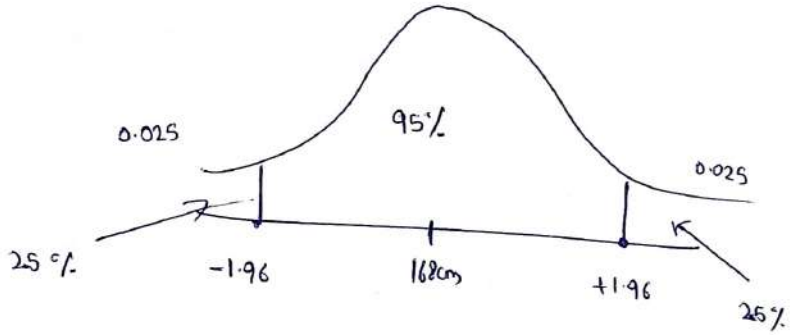
Ans $\mu = 168\text{ cm}$ $\sigma = 3.9$ $n = 30$ $\bar{x} = 169.5$

$CI = 0.95$ $\alpha = 1 - 0.95 = 0.05$

① Null hypothesis (H_0) = $\mu = 168\text{ cm}$

② $H_1 = \mu \neq 168\text{cm}$

③ Based on CI we will draw the Decision boundary



$1 - 0.025 = 0.9750$

Z-score = 0.9750

⇓

Area $\Rightarrow +1.96$

if Z is less than -1.96 or greater than $+1.96$, reject the Null hypothesis

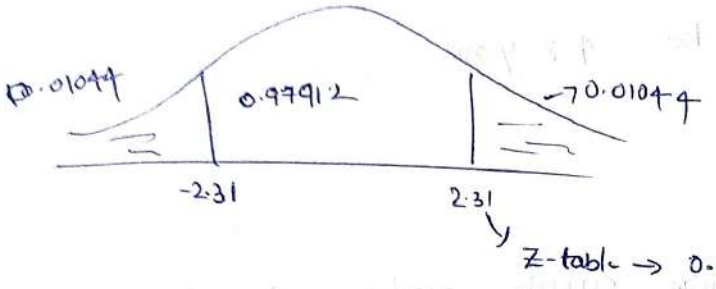
Z-Test $Z_x = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{4}} = \dfrac{169.5 - 168}{39/\sqrt{36}}$ Z-Score = $\dfrac{\bar{x} - \mu}{\sigma}$

$Z_d = \dfrac{1.5}{0.65} = \boxed{2.31}$

Conclusion:

2.31 > 1.96 Reject the null hypothesis

$P < 0.05$



Z-table → 0.98956

1 - 98956 = 0.01044

⇒ 0.01044 x 2

1 - 0.02088

P value = 0.01044 + 0.01044

= 0.02088

$P < 0.05$

0.02088 < 0.05 ⇒ Reject the null hypothesis

Final conclusion

the average height is not equal to 168 cm

The average height seems to be increasing based on sample data

(2) A factory manufactures bulbs with an average warranty of 5 years with standard deviation of 0.50. A worker believes that the bulb will manifaction is less than 5 years. He tests a sample of 40 bulbs and find the average time to be 4.8 years

(a) State null & alternate hypothesis

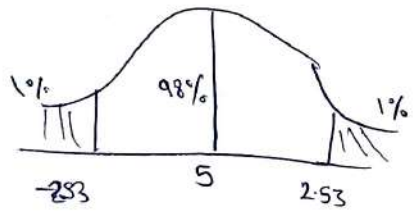
(b) At a 2% Significance level, is there enough evidence to support the idea that the warranty should be revised?

Ans $\mu = 5$, $\sigma = 0.50$, $n = 40$ & $\bar{x} = 4.8$

① $H_0 : \mu = 5$

② $H_1 : \mu < 5$

③ Decision boundary



$\alpha = 0.02$ ⇒

z-score = $1 - 0.02$

$= 0.98$

⇒ 2.53

$$2\text{-test} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.5/\sqrt{48}}$$

$= -2.53 = 0.0570$

Area under curve with z-score -2.53 = 0.0570

P-value = 0.0570

Compare p-value with α

$0.0570 < 0.02$ ⇒ false

We accept the Null hypothesis

We fail to reject null hypothesis

Student t distribution :-

In z stats when we perform any analysis using z-score we require σ (population standard deviation) → is already known

→ How do we perform any analysis when we don't know the population standard deviation ?

Student's t distribution :-

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

z-test

~~z-table~~

s = sample standard deviation

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

t-table ⇒ t-test

Degree of freedom

$$df = n - 1 = 3 - 1 = 2$$

3 people ☒ ☒ ☐

T-Stats :- T-test → one sample test

① In the population the average IQ is 100. A team of researches want to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a Standard deviation of 20. Did the medication affect intelligence

Ans

$\mu = 100$ $n = 30$ $\bar{x} = 140$ $S = 20$ CI = 95% $\alpha = 0.05$

(i) Null hypothesis H_0 : $\mu = 100$

(ii) Alternate hypothesis H_1 : $H_1 : \mu \neq 100$ {2 tail Test}

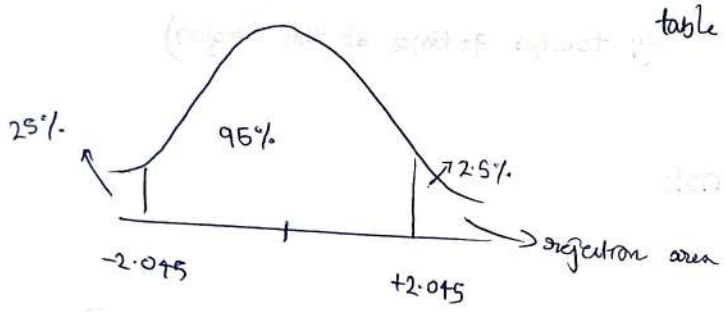
② $\alpha = 0.05$

③ Degree of freedom :-

$$dof = n - 1 = 30 - 1 = 29$$

④ Decision Rule

table → t
v & 95%



if t test is less than -2.045 or greater than 2.045, reject the null hypothesis

⑤ Calculating test statistics

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{140 - 100}{20/\sqrt{30}} = \frac{40}{3.65} = 10.96$$

$t = 10.96$

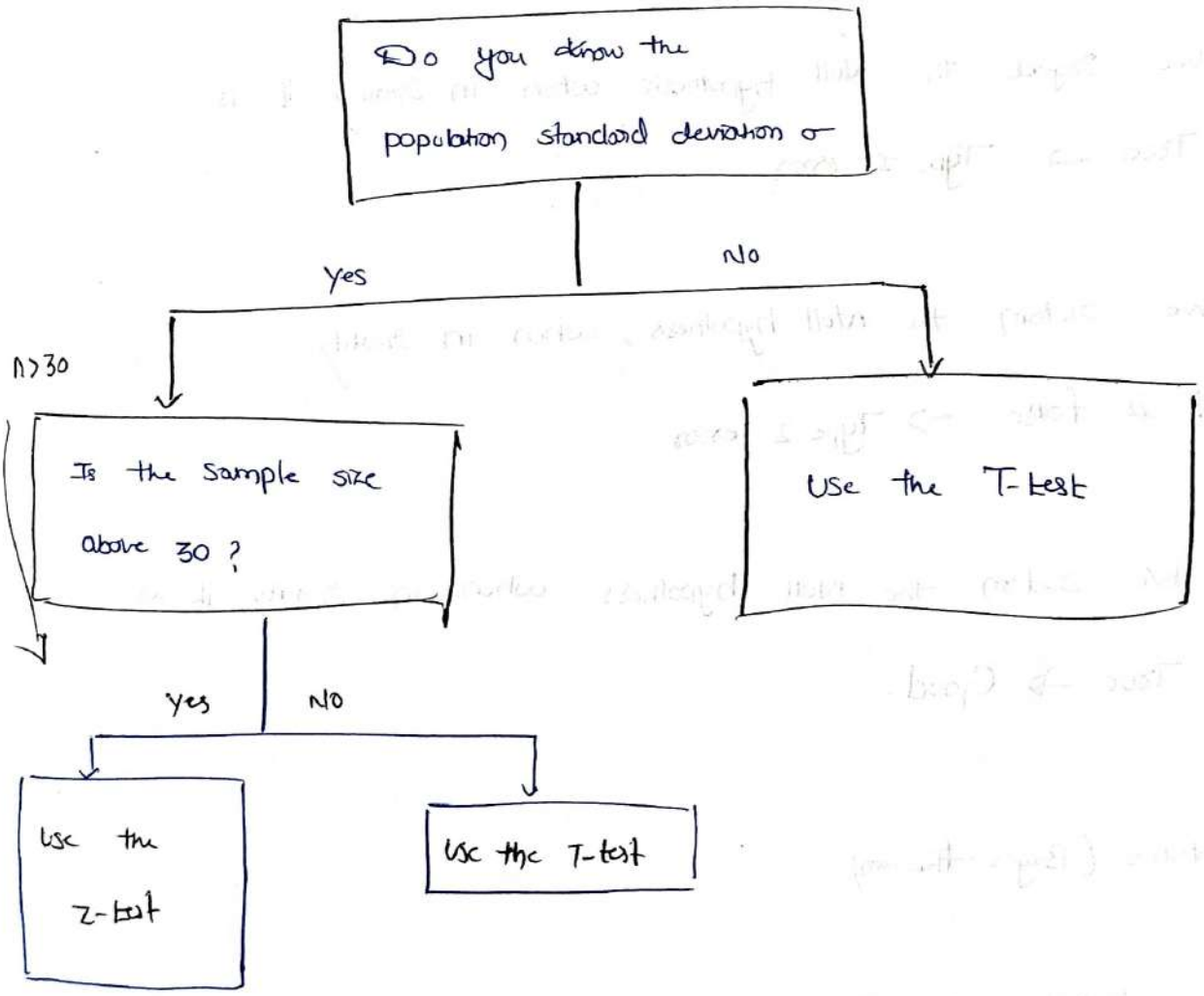
Since

$t = 10.96 > 2.095$ ∴ reject the null hypothesis

Conclusion: Medication used has affected the intelligence

Medication has increased

When To use T-tests VS Z-test



Type 1 and Type 2 Errors

Reality :- Null hypothesis is True or Null hypothesis is False

Decision :- Null hypothesis is True or Null hypothesis is False

OUTcome 1 :- We reject the Null hypothesis when in reality it is false → Good

OUTcome 2 :- We reject the Null hypothesis when in reality it is True → Type 1 error

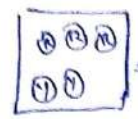
OUTcome 3 :- We retain the Null hypothesis, when in reality it is false → Type 2 error

OUTcome 4 :- We retain the Null hypothesis when in reality it is True → Good.

Bayes Statistics (Bayes Theorem)

Bayesian statistic is an approach to data analysis & parameter estimation based on Bayes theorem

probability — [Independent Events → Rolling a dice
Dependent Events



$P(R \text{ and } Y) = P(R) * P(Y/R)$ → Conditional probability

$= \frac{3}{5} * \frac{2}{4}$

$= \frac{6}{20}$

$P(A/B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A/B) * P(B)$ — (1)

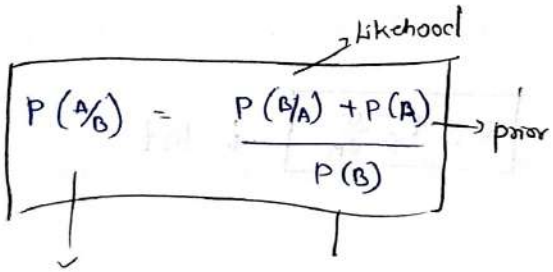
$P(B/A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(B/A) * P(A)$ — (2)

(1) & (2), we get

$P(A/B) * P(B) = P(B/A) * P(B)$

$P(A/B)$ = probability of A given B is true

$P(B/A)$ = probability of B a given A is true



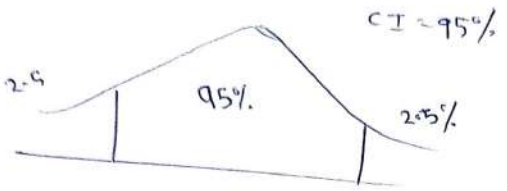
Dataset

Size of Mase	No of Room	Location	price

$P(y/x_1 x_2 x_3) = \frac{P(y) * P(x_1 x_2 x_3 / y)}{P(x_1, x_2, x_3)}$

Bayes's theorem

Confidence Intervals and margin of error



Point Estimate

$\bar{x}$ can less or greater than 3, this is a problem of point estimation we should define a range of confidence Interval

↓

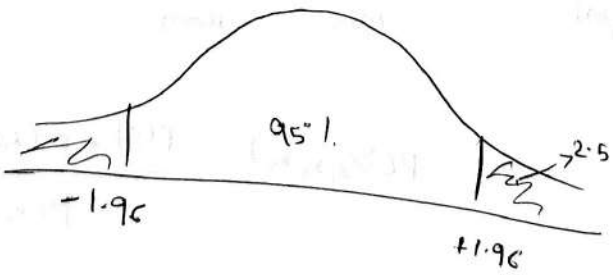
like PE lyes in 2 – 4

Confidence Interval

point estimate ± Margin of error

$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ → Z-test

$\bar{x} \pm t_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ t-test

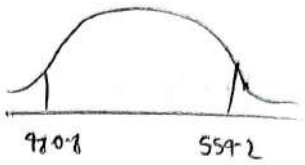


1 - 0.025 = 0.9750

±1.96 ⇒ 5 table

Lower CI $= 520 - (1.96)\left(\frac{100}{\sqrt{25}}\right) = 480.8$

Higher CI $= 520 + (1.96)\frac{100}{\sqrt{25}} = 559.2$



Conclusion: I am 95% confident that about the mean CAT score is between 480.8 to 554.2

CHI- SQUARE (χ^2) Test:-

When a coin is tossed 100 times the theoretical considerations leads us the Expect 50 heads & 50 tails. But in practice, these results are rarely achieved

i.e The result obtained in a experiment do not agree exactly with the theoretical results.

-> The magnitude of discrepancy between the theory and observation is given by the quantity χ^2 (a greek letter, pronounced as 'chi-square').

-> If $\chi=0$, the observed & Expected frequencies completely coincide.

-> As the value χ^2 increases the ~~affords as a measure of the~~ corresponding discrepancy between the observed & theoritical frequencies increases.

-> Thus, χ^2 affords a measure of the correspondence between theory & observation

Definition: If a set of events $A_1, A_2, A_3 \ldots A_n$ are observed to occur with frequencies $O_1, O_2, O_3 \ldots O_n$ respectively and according to probability rules $A_1, A_2, \ldots A_n$ are expected to occur with frequencies $E_1, E_2, \ldots E_n$ respectively with $O_1, O_2, \ldots O_n$ are called observed frequencies and $E_1, E_2, \ldots E_n$ respectively with $O_1, O_2 \ldots O_n$ are called observed frequencies & $E_1, E_2 \ldots E_n$ are called Expected frequencies

Def If O_i $(i=1,2\ldots n)$ is set of observed (experimental) frequencies & E_i $(i=1,2,\ldots,n)$ is a (Corresponding) set of ~~observed~~ (~~Experimental~~ theoretical / Expected) frequencies, then χ^2 is defined as $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ with $(n-1)$ degrees of freedom

It is used to test whether the differences b/w of & Ef are significant

Note: If the data is given in a series of 'n' numbers the degrees of freedom $= n-1$

In case of Binomial distribution, d.f $= n-1$

Poisson distribution, d.f $= n-2$

Normal distribution, d.f $= n-3$

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots \frac{(O_n - E_n)^2}{E_n}$$

CHI Square for Goodness of Freedom

In 2010 Census of the city, the weight of the individuals in a small city were found to be the following.

<50kg	50-75	>75
20%	30%	50%

In 2020, weight of $n = 500$ individual were sampled. Below are the result

<50	50-75	>75
140	160	200

using $\alpha = 0.05$, would you conclude the population differences of weights has change in the last 10 years?

Ans

2010

Expected

<50kg	50-75	>75
20%	30%	50%

2020

observed

$n = 500$

<50	50-75	>75
140	160	200

Expected

<50	80-75	>75
0.2 x 500 = 100	0.3 x 500 = 150	0.5 x 500 = 250

① Null hypothesis H_0 : The data meets the Expectation

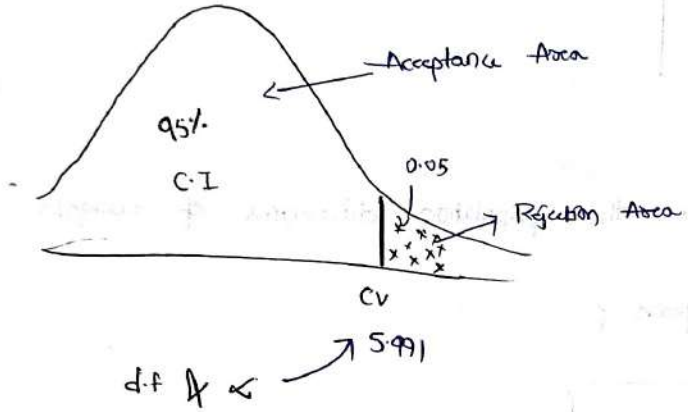
② Alternate hypothesis H_1 : The data does not meet the Expectation

③ $\alpha = 0.05$ C.I = 95%

④ Degree of freedom

$d.f = k-1 = 3-1 = 2$

⑦ Decision boundary



If x^2 is greater than 5.99, reject no else

we fail to reject the Null hypothesis

⑤ Calculation:-

$$x^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$= 16 + 0.66 + 10$ $\quad 26.66 > 5.99$, reject H_0

$x^2 = 26.66$ The weights of 2020 populations are different than those experiment in 2010. population

Analysis of variance (ANNOVA) super imp Statistical method

Definition:- ANOVA is a statistical method used to compare the means of 2 or more groups

ANNOVA

1) Factors (variable)
2) levels

Eg:- Medicine (factor

[Dosage] 5mg 10mg 15mg → level

Eg:- Mode of payment (factors)

GPay Phonepay Supermoney (levels)

Assumptions of ANONA (imp of interview)

→ centre limit theorm

1) Normality of Sampling Distribution of mean

The distribution of Sample mean is normaly distributed

2) Absence of outliers

outlying score need to be removed from the dataset

3) Homogenity of variance $[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$

Population variance in different levels of each independent variables are equal

4) Samples are independent & randomly selected

Types of ANNOVA (3 Types)

① One Way ANNOVA :- one factor with atleast 2 levels, these levels are Independent

Eg:- Doctor wants to test a new medication to decrease headache they split the participants in 3 conditions [10 mg, 20mg, 30 mg] Doctor ask the participants to state the headache [1-10]

Medication → Factor

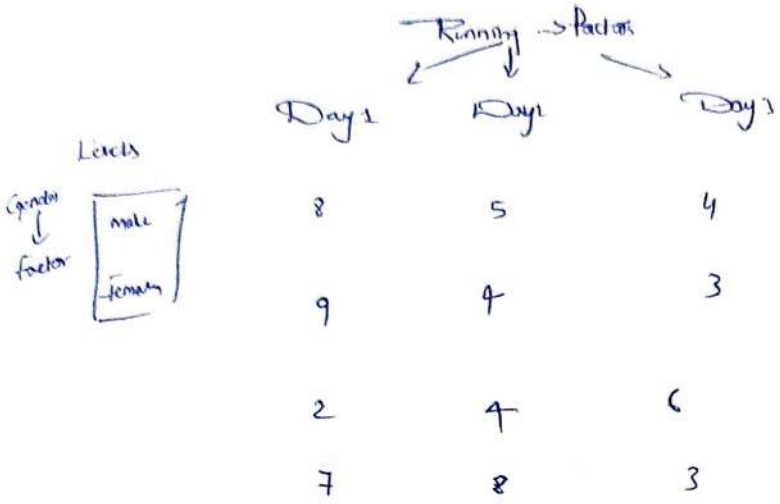
10mg	20mg	30mg
5	7	2
3	9	1
—	—	—
—	—	

② Repeated Measures Annova :- one factor with atleast 2 levels, levels are dependent

Running → Factor

levels	Day 1	Day 2	Day 3
	8	5	4
	7	4	9
	—	—	—
	—	—	—

③ Factorial ANOVA: Two or more factors (each of which with atleast) 2 levels, levels can be independent and dependent



Hypothesis Testing In ANOVA (Partitioning of variance in the Anova)

Null Hypothesis : H_0 $\mu_1 = \mu_2 = \mu_3 \cdots \mu_n$

Alternate hypothesis H_1 Atleast one of the Sample mean is now equal

XX $\mu_1 \neq \mu_2 \neq \mu_3 \neq \cdots \mu_n$ → wrong

Test Statistics

$$F = \frac{\text{Variance between Sample}}{\text{Variance within Sample}}$$

Variance between sample

	x_1	x_2	x_3
Variance within Samples	1	7	2
	5	6	3
	3	7	7
	9	8	8
	6	9	6

$\bar{x}_1 = 3 \quad \bar{x}_2 = 19/5 \quad \bar{x}_3 = 4$

Ex:- One way ANOVA

One factor with atleast 2 levels, levels are independent

① Doctors want to test a medication which reduces headache. They splits the particular into 3 condition [15mg, 30mg, 45mg]. Later on the doctor ask the patient to rate the headache between [1-10]. are there any difference bifferance between the 3 condition using $\alpha = 0.05$?

Sol

15mg	30mg	45mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

① Define Null & Alternate hypothesis?

$H_0 : \mu_{15} = \mu_{30} = \mu_{45}$

H_1 : ~~Not~~ Atleast one mean is not equal

② Significance

$\alpha = 0.05$, $C.I = 1 - 0.05 = 0.95$

③ Calculate Degree of freedom

$N = 21$ $a = 3$ $n = 7$

↓
Categories

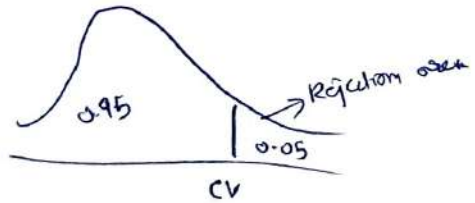
$df_{between} = a - 1 = 3 - 1 = 2$

$df_{within} = N - a = 21 - 3 = 18$

$df_{Total} = N - 1 = 20$

df_1 df_2
$(2, 18)$
↓ with $\alpha = 0.05$
F-table
↓ find's
Critical value

④ Decision Boundary



Decision Rule

if f is greater than 3.5546, reject the NULL Hypothesis

⑤ Calculate F Test statistic

~~Between~~	Sum of Square	df	mean Square	F
Between	98.67	2	49.34	
within	10.29	18	0.54	
Total	108.96	20		

① SS between $= \frac{\sum (\sum_{ai})^2}{n} - \frac{T^2}{N}$

15mg : 4+8+7+8+4+5+9+8 = 57

30mg = 7+6+7+8+7+4 = 47

45mg = 4+3+2+3+4+2+2 = 21

$= \frac{57^2+47^2+21^2}{7} - \frac{[57^2+47^2+21^2]}{21}$

$= 98.67$

② SS within $= \sum Y^2 - \sum (\sum_{ai})^2$

$\sum Y^2 = 9^2+8^2+7^2+8^2+4^2 \ldots\ldots$

$= 853$

$= 853 - \left(\frac{57^2+47^2+21^2}{7}\right)$

$= 10.29$

F. test = $\frac{MS_{Between}}{MS_{within}}$

F-test$_{(i.)}$ = $\frac{49.34}{0.54}$

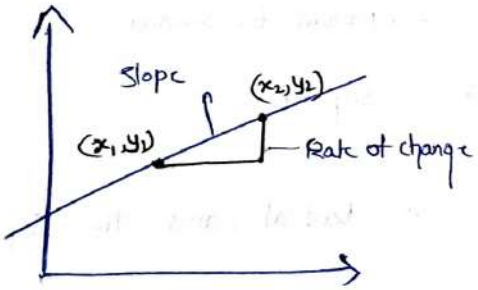
= 86.56

If f is greater than 3.554, Reject the H_0

86.56 > 3.554 Reject the H_0

Slope → Derivate as a concept

The Slope of a line is a measure of how steep the line is, and it represents the rate of change of one variable with respect to another. In the context of a two-dimensional cartesian co-ordinate System, the Slope indicates the ratio of the vertical change (rise) to the horizontal change (run) between two points on a line



Slope = $\frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x}$

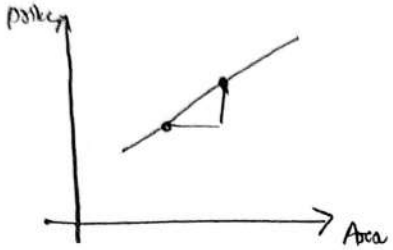
Where $y_2 - y_1$ is the vertical (rise)

$x_2 - x_1$ is the horizontal (run)

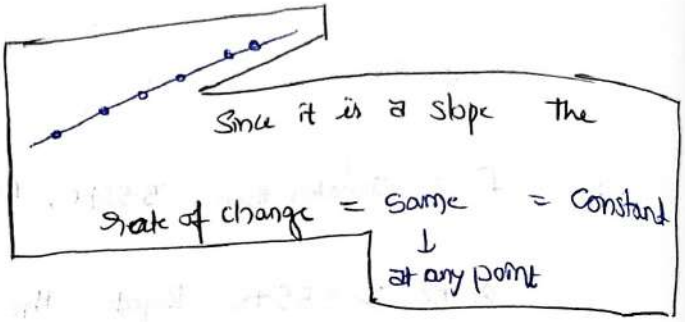
Example Data Set:

↓

2 features : Area & price



as area increase price decreases

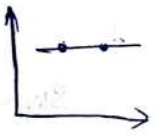


Interpretation of slope:

* Positive slope: If slope >0, the line rises as it moves left to right, The longer the slope, the steeper the line. 

* Negative slope: If slope <0, the line falls as it moves from left to right, the more negative the slope, the steeper the line in downwards direction 

* Zero slope: If slope =0 the line is horizontal, meaning there is no vertical change as the line move from left to right

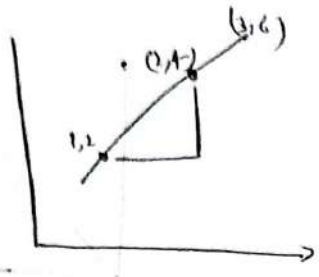


⇒ parallel to x-axis

slope =0

* Undefined slope: If $x_2 = x_1$, the line is vertical, and the slope is undefined because you cannot divide by 0

Example 1:- Positive Slope:



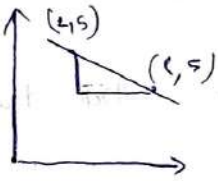
$$\text{Slope} = \frac{6-2}{3-1} = \frac{4}{2} = 2$$

This means the for every 1 unit you move horizontally from left to right the line moves 2 units vertically up

Example 2:- Negative slope:-

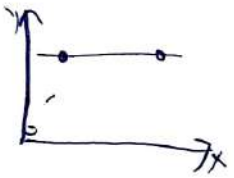
(2,5) (6,3)

$$\text{Slope} = \frac{3-5}{6-2} = \frac{-2}{4} = -\frac{1}{2}$$



This mean for every 2 units you move horizontally to the right, the line moves 1 unit vertically down

Example 3:- Zero slope Consider the point (4,4), (5,4)



$$\text{Slope} = \frac{4-4}{5-1} = \frac{0}{5-1} = 0 \Rightarrow \frac{0}{4} = 0$$

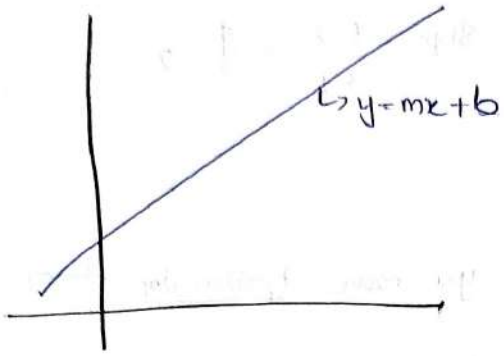
The means the line is horizontal

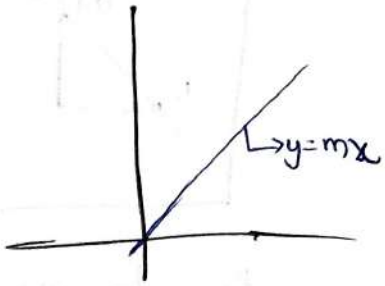
Example 4:- Undefined slope

(3,2) (3,7)

$$\text{Slope} = \frac{7-2}{3-3} = \frac{5}{0}$$

Slope in a equation of a line





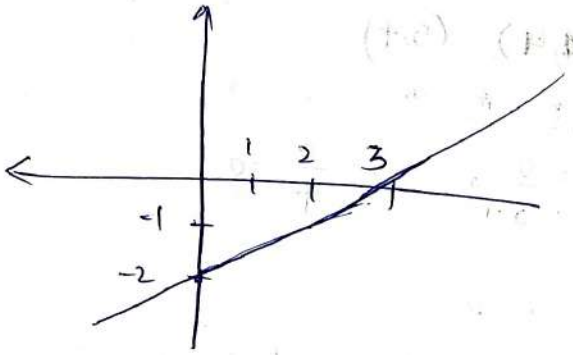
$y = mx + b$

m is the slope of the line

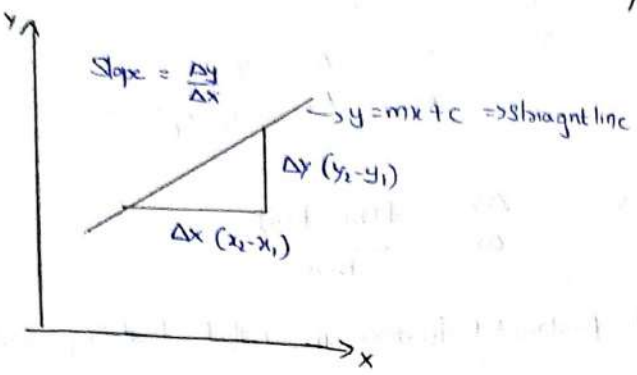
b is a the y intercept where the line arocces y axis

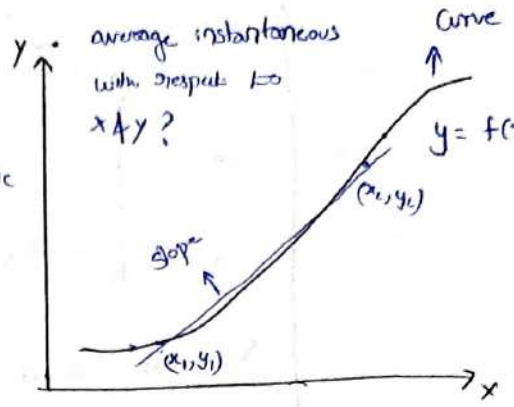
$m = 3$ $b = -2$

$\Rightarrow y = 3x - 2$



Derivate:

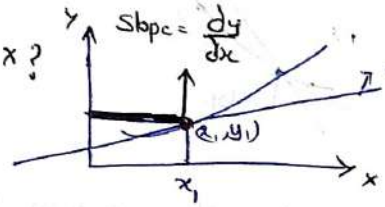




In this figure the slope is change as it is a curve

=> To get average instantaneous we draw a Secant line

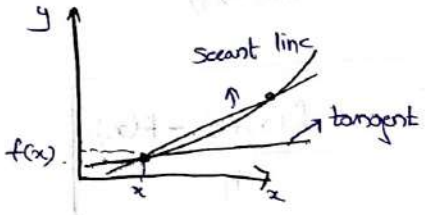
with respect to x ?

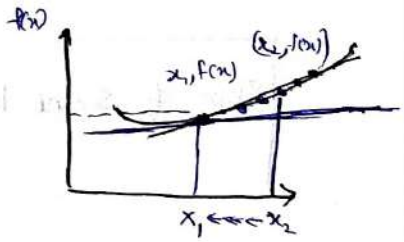


Rate of change of x_1 with respect to y ? cannot draw Secant so, we calculate by drawing tanget

Slope of tanget = Instantaneous rate of change of x with respect to y

=> slope = $\frac{dy}{dx}$

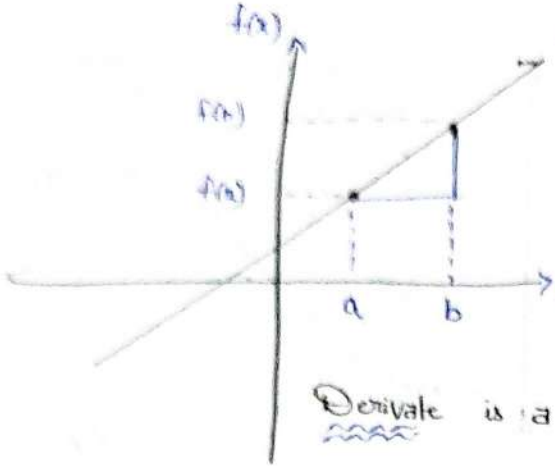




Secant to Tangent then slope

Mathematics Notation of Derivative with limits :

The derivate is a fundamental concept in calculus that represents the rate of which a function is changing at any given point. It is essentially the slope of the tangent line to the function's graph at that point. The derivatives is used to understand how a function behaves as its input changes & it is a key tool for analyzing the dynamics of systems in mathematics, physics, economics, engineering & many other fields

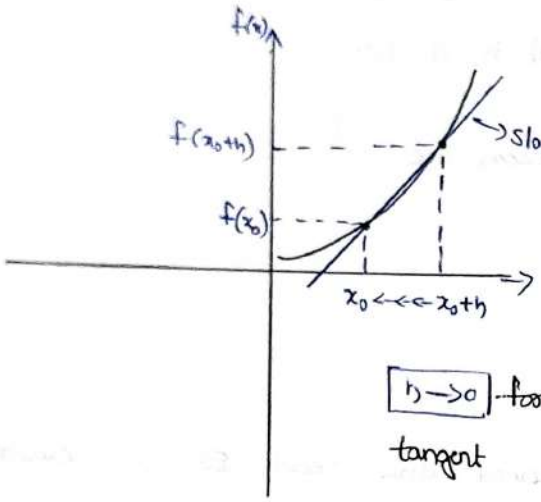


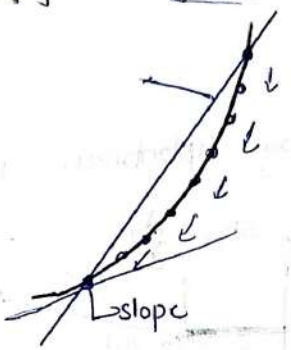
$f(x) = mx + c$ → intercept

slope = $\frac{\text{rise}}{\text{run}} = \frac{\Delta y}{\Delta x} = \frac{y_1 - y_1}{x_1 - x_1}$

$$\frac{\Delta y}{\Delta x} = \frac{f(b) - f(a)}{b - a}$$

Derivate is a fundamental function in calculs that represents the rate of which a function is changing at a given point.





Secant to tangent

$h \to 0$ for calculating tangent

- Secant line slope is going to give the average rate of change
- To get exact slope at a point you need to draw a tangent line & then find a slope

$$\text{Slope} = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{f(x_0+h) - f(x_0)}{(x_0+h) - x_0} = \frac{\Delta y}{\Delta x}$$

$$\boxed{\text{Slope of Secant line} = \frac{f(x_0+h) - f(x_0)}{h}}$$

To Calculate rate of change at specific point we need to convert Secant line to tangent by moving from $x_0 + h$ to x_0 [we use limit to move]

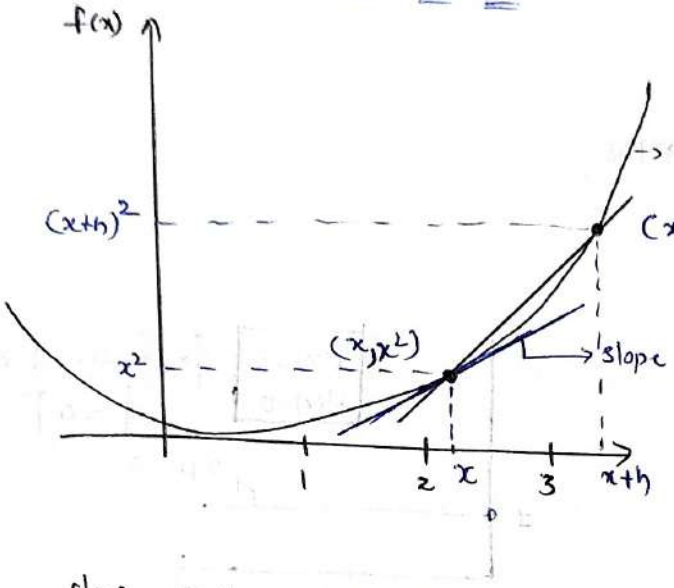
$$f'(x) = \boxed{\lim_{h \to 0} \frac{f(x_0+h) - f(x_0)}{h}} \leftarrow \text{derivative of } f(x)$$

$$\boxed{f'(x) = \frac{dy}{dx} = \frac{d(f(x))}{dx}}$$

Finding a Derivative At a point :

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

* proofing derivative of x^2 is $2x$



$f(1) = 1$
$f(2) = 4$
$= 9$
$\vdots$
n^2

limit $h \to 0$ → for calculating slope at tanget

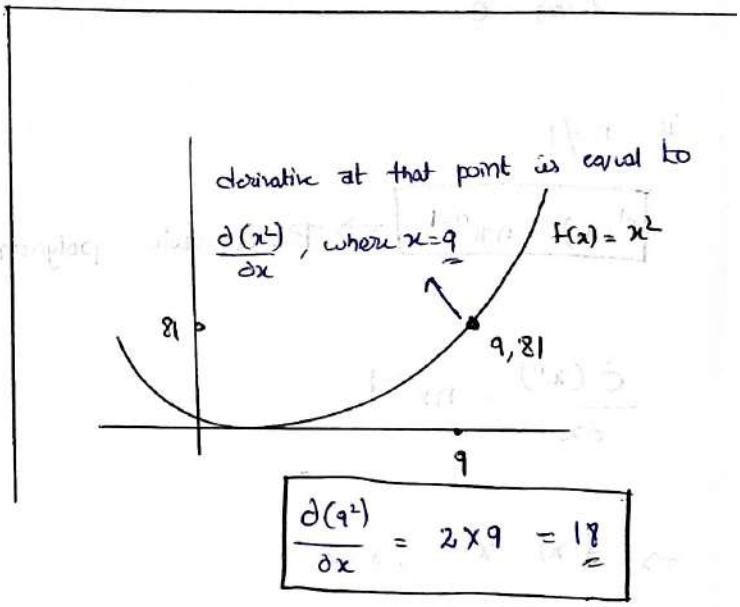
$$f'(x) = \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h}$$

$$f'(x) = \lim_{h \to 0} \frac{x^2 + 2xh + h^2 - x^2}{h}$$

$$= \lim_{h \to 0} \frac{h(2x+h)}{h}$$

$$= 2x + 0$$

$$f'(x) = 2x$$ ⟶ find a derivative at a point **X**



$$\frac{d(9^2)}{dx} = 2 \times 9 = 18$$

where $f(x) = x^2$

Hence proved that $\frac{d(f(x))}{dx} \Rightarrow \boxed{\frac{d(x^2)}{dx} = 2x}$ $\quad f'(x^2) = 2x$

$$\frac{d(x^n)}{dx} = nx^{n-1} \Rightarrow \frac{d(x^2)}{dx} = 2x^{2-1} = 2x$$

Power Rule In Derivatives

→ It is applied to polynomial

$f(x) = x^2$

$f'(x) = 2x$

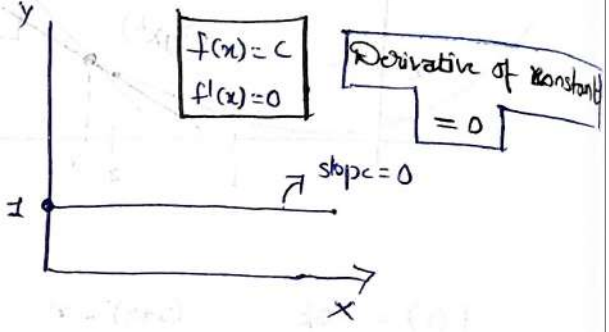
$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

different polynomial equations have different graphs

↓

x^2+3, x^3, x^2+2x+1 ... etc

$f(x) = x^n$, $n \neq 0$



if $n = 0$ $\quad f(x) = x^0 = 1$ => constant value

$f'(x) = 0$

if $n \neq 1$

$\boxed{f'(x) = n x^{n-1}}$ => Power rule polynomial Expression.

$$\frac{\partial (x^n)}{\partial x} = n x^{n-1}$$

=> $f(x) = x^3$, $x = 2$

$$\frac{\partial (x^3)}{\partial x} = 3x^2 = 3 \times 2^2 = 12$$

=> $$\frac{\partial (3x^2)}{\partial x} = 3 \frac{\partial x^2}{\partial x} = 3 \cdot 2x = 6x$$

=> $$\frac{\partial (1/x)}{\partial x} = \frac{\partial (x^{-1})}{\partial x} = -x^{-2} = \frac{-1}{x^2}$$

Assignment :-

① $f(x) = x^8$ Ans $f'(x) = \frac{\partial x^8}{\partial x} = 8x^{8-1} = 8x^7$

② $f(x) = x^{-1}$ $\Rightarrow f'(x)$ at $x = -1$

Ans $\frac{\partial x^{-1}}{\partial x} = -\frac{1}{x^2} = \frac{-1}{(-1)^2} = -1$

* Derivative Rules :- Constant, sum, difference and Constant Multiple

$\frac{\partial (x^n)}{\partial x} = nx^{n-1}$, $n \neq 0$ {power rule} => polynomial

$\frac{\partial (x^0)}{\partial x} = \frac{\partial (1)}{\partial x} = 0$ => Derivation of a constant is 0

$\frac{\partial (c)}{\partial x} = 0$ ← constant

$\Rightarrow \frac{\partial (c\, f(x))}{\partial x} = c\frac{\partial (f(x))}{\partial x} = c f'(x)$

$\frac{\partial (3x^4)}{\partial x} = 3\frac{\partial (x^4)}{\partial x} = 3 \times 4x^{4-1} = 12x^3$

at $x = 2$ $\Rightarrow 12 \times 2^3 = 12 \times 8 = 96$

Assignment $\frac{\partial [A f(x)]}{\partial x} = \frac{\partial [2x^5]}{\partial x} = 2\frac{\partial (x^5)}{\partial x} = 2 \times 5x^{5-1} = 10x^4$

* Sum of 2 functions $f(x)$, $g(x)$

$$\frac{\partial [f(x) + g(x)]}{\partial x} = \frac{\partial (f(x))}{\partial x} + \frac{\partial (g(x))}{\partial x}$$

$$\frac{\partial [x^4 + x^{-2}]}{\partial x} = \frac{\partial (x^4)}{\partial x} + \frac{\partial (x^{-2})}{\partial x}$$

x^{-2}

$-2x^{-2-1}$

$$= 4x^3 - \frac{2}{x^3}$$

~~Assignment~~ :-

(1) $$\frac{\partial (x^2+2)}{\partial x} = \frac{\partial (x^2)}{\partial x} + \frac{\partial (2)}{\partial x}$$

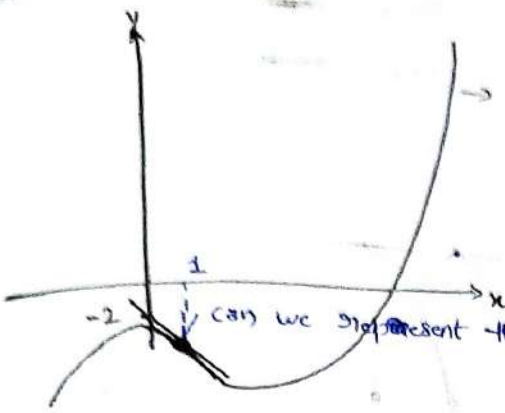
$$= 2x =$$

(2) $$\frac{\partial [4x^3 - 6x^2 + 2x + 100]}{\partial x} = \frac{\partial [4x^3]}{\partial x} - \frac{\partial [6x^2]}{\partial x} + 2\frac{\partial (x)}{\partial x} + \frac{\partial (100)}{\partial x}$$

$$= 12x^2 - 12x + 2 + 0$$

$$= 12x^2 - 12x + 2$$

Tangent of polynomial



$\rightarrow f(x) = x^3 - 6x^2 + x - 7$

can we represent this line using some equation ?

$$f'(x) = \frac{d(x^3)}{dx} - \frac{d(6x^2)}{dx} + \frac{d(x)}{dx} - 7$$

$$= 3x^2 - 12x + 1 - 0$$

$$= 3x^2 - 12x + 1$$

at $x = 1$

$y = f(1) = 1^3 - 6(1)^2 + (1) - 7$

$= 1 - 6 + 1 - 7 = -11$ //

$f'(1) = 3 - 12 + 1 = -9$ => slope

$y = mx + c$

$y = -9x + c$

$-11 = -9x + c$

$c = +9(1) - 11$

$\boxed{c = -2}$

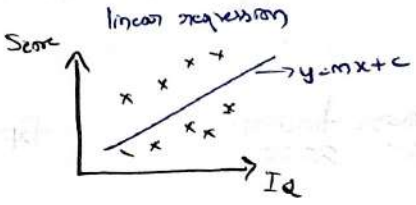
=> $y = mx + c$

$\boxed{y = -9x - 2}$ → equation that particular target at that particular point at $x=1$

↓ slope ↓ intercept

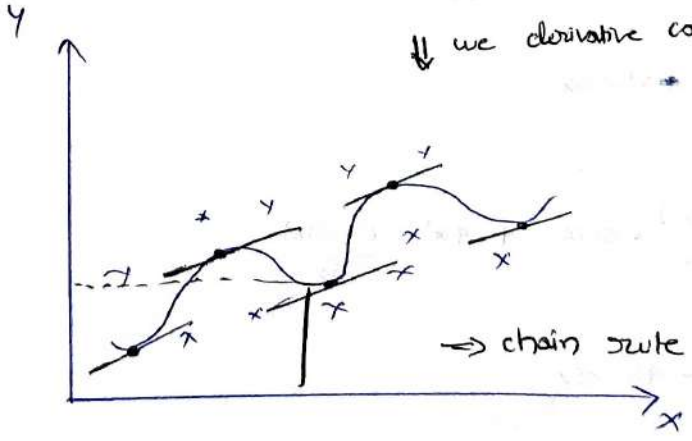
Ex: IQ Dataset

IQ	Score
100	91
100	89
90	90
85	88



IQ → model → Score

we by to find right slope & right intercept we will satisfied this line such that for any data point model should predict the output

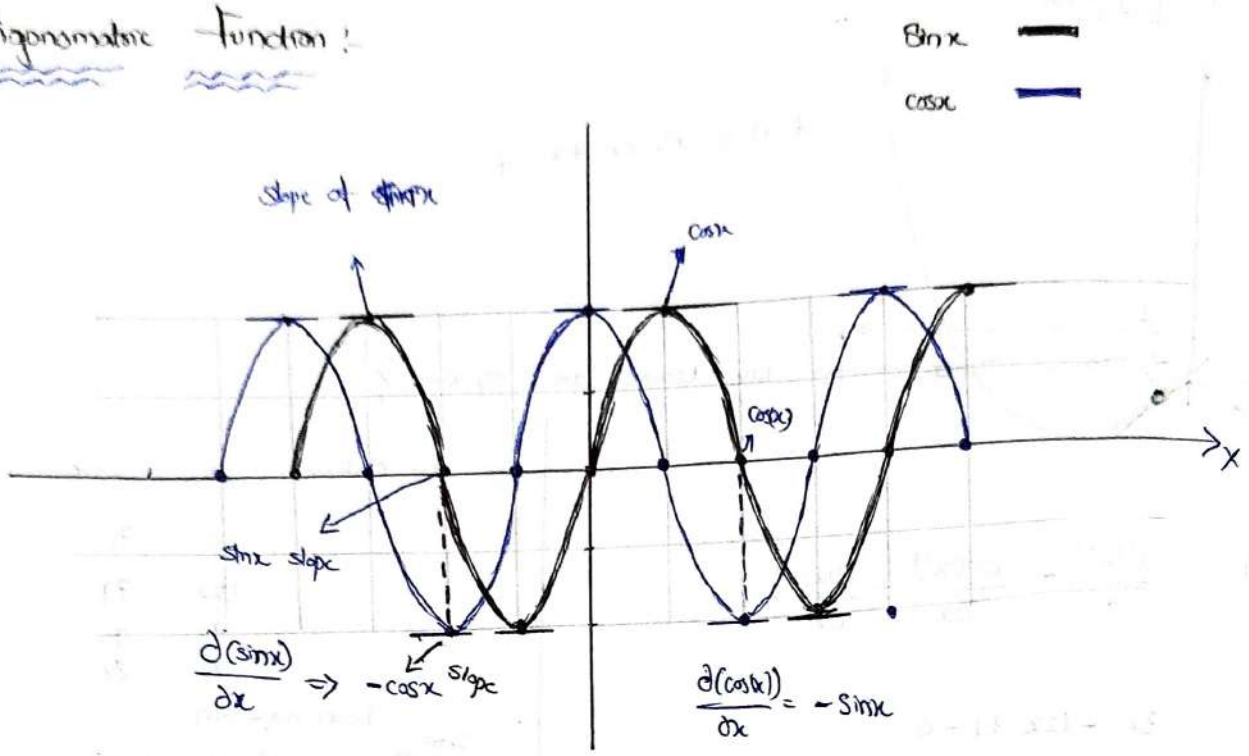


we derivative concepts to get slope.

=> with respect any value predict the output

=> chain rule

Trigonometric function :



Logarithmic function:

$f(x) = \ln(x)$ then

$f'(x) = \frac{1}{x}$

Exponential function

$f(x) = e^x$

$f'(x) = e^x$

constant

if $f(x) = c$

then $f'(x) = 0$

power rule

if $f(x) = x^2$

$f'(x) = nx^{n-1}$

Production rule in Derivatives:

$$\frac{\partial [h(x) \cdot f(x)]}{\partial x} = h'(x) \cdot f(x) + h(x) \cdot f'(x)$$

① $\frac{d(x^2 \cos x)}{dx} = \frac{d(x^2)}{dx} \cdot \cos x + x^2 \cdot \frac{d(\cos x)}{dx}$

$= 2x \cdot \cos x - x^2 \cdot \sin x$

Assignment

② $\frac{d(4x^2 \sin x)}{dx} = \frac{d(4x^2)}{dx} \cdot \sin x + 4x^2 \cdot \frac{d(\sin x)}{dx}$

$= 8x \sin x - 4x^2 \cos x$

Chain Rule :- [of Derivation]

The Chain Rule is a fundamental theorem in calculus that is used to find the derivates of a Composite function. when a function is Composed of other function, the chain rule allows us to differentiate it with respect to the innermost variable.

formal defination :-

If $y = f(g(x))$ where $y = f(u)$ & $u = g(x)$, then the derivative of y with respect to x

$$\boxed{\frac{dy}{dx} = \frac{df}{du} \cdot \frac{du}{dx}}$$

In simpler term, this can be expressed as

$$\boxed{\frac{dy}{dx} = f'(g(x)) \cdot g'(x)}$$

This means that to differentiate a composite function, you first differentiate the outer function with respect to its inner function and then multiply by the derivatives of the inner function.

Example 1:-

$$y = \underbrace{(3x^2+2x+1)}_{u}^{5}$$

Step 1: Identify the outer and inner function

Outer function $f(u) = u^5$

Inner function $u = g(x) = 3x^2+2x+1$

Step 2: Differentiate the outer function

$$f'(u) = \frac{df}{du} = \frac{du^5}{du} = 5u^4$$

Step 3: Differentiate the inner function

$$g'(x) = \frac{du}{dx} = \frac{d(g(x))}{dx} = \frac{d(3x^2+2x+1)}{dx} = 6x+2$$

Step 4:- $\frac{du}{dx} = f'(g(x)) \cdot g'(x)$

$$= 5u^4 \cdot (6x+2)$$

$$\frac{du}{dx} = 5(3x^2+2x+1)^4 \cdot (6x+2)$$

② Trigonometric

$y = \sin(4x^3+x)$ $\quad u = g(x) = 4x^3+x$

$= \sin(u)$

$\Rightarrow y = f(g(x))$

Step1: Identify the outer and inner function.

Outer function: $f(u) = \sin(u)$

inner function: $u = g(x) = 4x^3+x$

Step2: $f'(u) = \frac{df}{du} = -\cos(4)$

Step3: $g'(x) = \frac{du}{dx} = \frac{d(4x^3+x)}{dx} = 12x^2+1$

Step4: Chain Rule

$$\frac{du}{dx} = f'(g(x)) \cdot g'(x) = -\cos(4x^3+x) \cdot (12x^2+1)$$

$$\boxed{\frac{du}{dx} = -\cos(4x^3+x) \cdot (12x^3+1)}$$

Example of ~~three~~ Composition of three function

$y = \sqrt{\sin(3x)}$ {3 composite function}

Step1: Identify the function

- outer function: $f(u) = \sqrt{u} = u^{1/2}$ — $u = \sin(3x)$
- Middle function: $g(v) = \sin(v)$ — $v = 3x$
- Inner function: $v = h(x) = 3x$

Step 2: $f'(u) = \frac{1}{2}u^{-1/2} = \frac{1}{2\sqrt{u}}$

$g'(v) = -\cos v$

$h'(x) = 3$

Step 3: Apply chain rule

$$\frac{du}{dx} = \frac{df}{du} \cdot \frac{du}{dv} \cdot \frac{dv}{dx} \qquad u = \sin(3x)$$

$$= f'(g(h(x)) \cdot g'(h(x)) \cdot h'(x)$$

$$= \frac{1}{2\sqrt{\sin(3x)}} \cdot \cos(3x) \cdot 3$$

$$\boxed{\frac{du}{dx} = \frac{3\cos(3x)}{2\sqrt{\sin(3x)}}}$$

The chain rule is a powerful tool in calculus for differentiating composite function. By breaking down the differentiation process in to manageable steps; it allows us to compute derivatives of complex expressions efficiently. Understanding and applying the chain rule is essential for solving a wide range of problems in mathematics, physics, engineering, economics, and datascience

Applications of chain Rule In Data Science:

1] Backpropagation: Training Deep learning models

Chain rule for calculating the gradients of Loss function with respect to weight that are initialized.

2] Gradient Descent optimization:

Linear Regression - the goal is to minimize a cost function

↳ update slopes — Gradient Descent optimization - chain rule

3] Chain rule in F.E

$z = \log(x^2+y^2)$

x z $\quad \boxed{\frac{\partial z}{\partial x} = \frac{1}{x^2+y^2} \cdot 2x}$ => How small change in x & y affect z, chain rule

y z $\quad \boxed{\frac{\partial z}{\partial x} = \frac{1}{(x^2+y^2)} \cdot 2y}$

4] Regularization Techniques:

Overfitting And underfitting.

Application of Linear Algebra, Stats and Differential Calculus

Main Aim:-

1] Linear Algebra
2] Statistics
3] Differential Calculus

} => Application in Data Science

① Simple linear Regression
② Multiple linear Regression
③ Dimensionality Reduction → PCA
④ Artificial Neural Network

} => Linear Algebra + Differential calculus + Stats

=> Model Trained

⇓

Deep learning

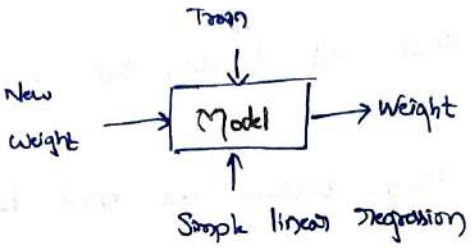
Simple linear Regression :- (training Single input/independent feature):

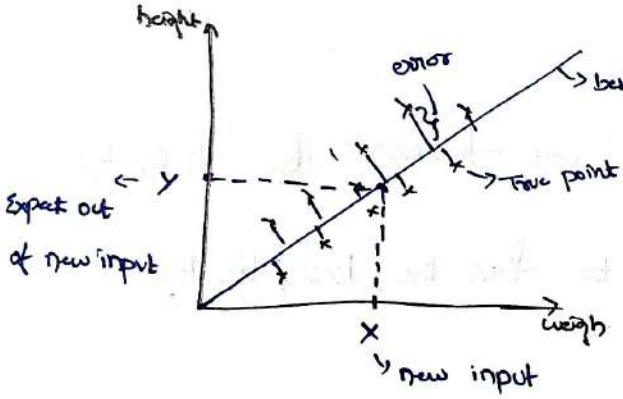
In Supervised ML the Regression problems are solved by using Simple linear regression

Ex: **Dataset:-**

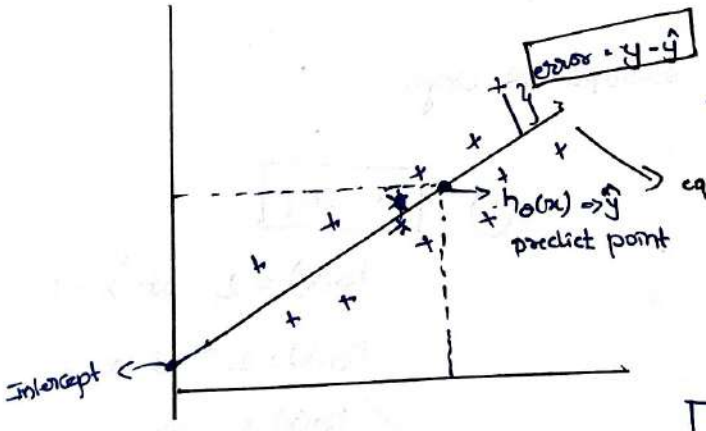
Weight (independent)	Height (dependent or o/p)
74	170 cm
80	180 cm
75	175.5 cm
—	—

⇒ True output





The best-fit line should be created in such a way that sum of all distance between the all the true point and expected point should be minimal



Dataset:-

equation of bestfit line

$y = mx + c$

$y = \beta_0 + \beta_1 x$

$h_\theta(x) = \theta_0 + \theta_1 x$ → Adrien ng

for many independent features:-

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \dots \theta_n x_n$

if $x = 0$

$h_\theta(x) = \theta_0$ intercept

θ_0 = Intercept

θ_1 = Slope / Co-efficient { a unit movement in x axis what is the movement with respect to y-axis }

Error = $y - \hat{y}$

Cost function :-

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x)^{(i)} - (y)^{(i)} \right)^2$$

- θ_1 : Slope
- θ_0 : intercept
- $h_\theta(x)^{(i)}$: predicted
- $(y)^{(i)}$: true o/p
- $\left(h_\theta(x)^{(i)} - (y)^{(i)} \right)$: Error
- $\Rightarrow$ Mean Square Error

we also have other cost function that are mean absolute error, root MSE ... etc

we are using MSE because we need to minizing in such a way that by changing the theta zero & theta one

Main Aim ? what we need to solve.

The main aim is to minimize the cost function $J(\theta_0, \theta_1)$

By continously change θ_0 & θ_1 we try to find the best fit line

① $h_\theta(x) = \theta_0 + \theta_1 x$

Let assume $\theta_0 = 0$ $\Rightarrow$ that implies intercept at Origin

for simple understanding

$$h_\theta(x) = \theta_1(x)$$

Dataset

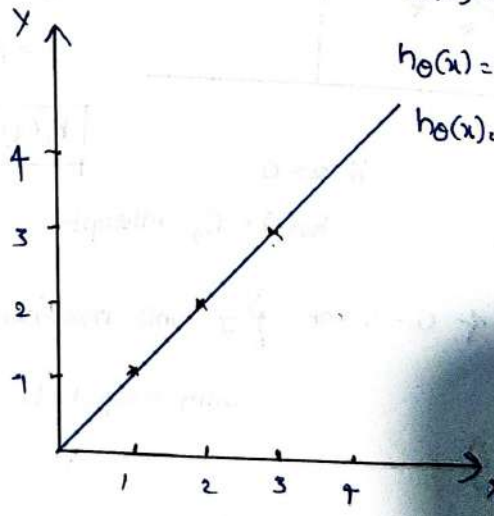
x	y
1	1
2	2
3	3

(i) Let $\theta = 1$

$h_\theta(x) = 1$ at $x = 1$

$h_\theta(x) = 2$ at $x = 2$

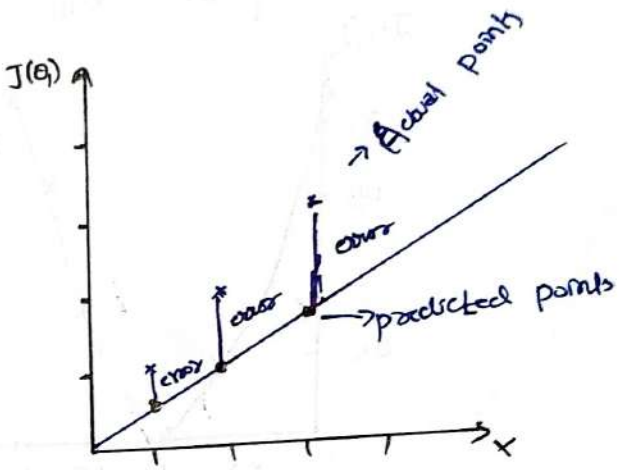
$h_\theta(x) = 3$ at $x = 3$



$$J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{n} \left(h_\theta(x)^i - y^{(i)} \right)^2$$

$$= \frac{1}{2(3)} \left[(1-1)^2 + (2-2)^2 + (3-3)^2 \right]$$

$$\boxed{J(\theta_1) = 0}$$



(ii) Let $\theta_1 = 0.5$

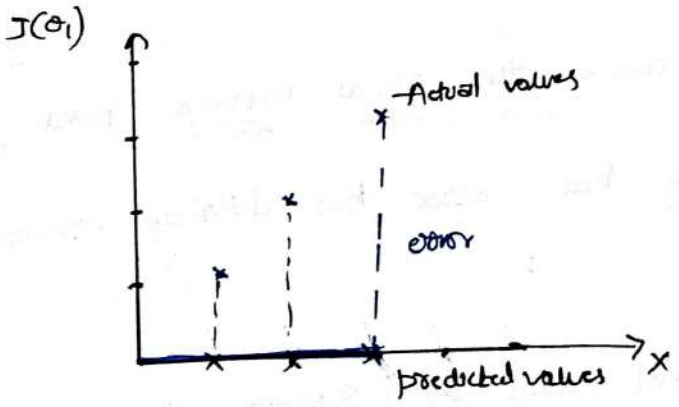
$x = 1$, $h_\theta(x) = (0.5) \times (1) = 0.5$

$x = 2$, $h_\theta(x) = 1$

$x = 3$, $h_\theta(x) = 1.5$

$$J(\theta_1) = \frac{1}{2 \times 3} \left((0.5-1)^2 + (1-2)^2 + (1.5-3)^2 \right)$$

$$\boxed{J(\theta_1) \approx 0.58}$$



(iii) Let $\theta_1 = 0$

$h_\theta(x) = 0 \quad x = 1$

$h_\theta(x) = 0 \quad x = 2$

$h_\theta(x) = 0 \quad x = 3$

$$J(\theta_1) = \frac{1}{2 \times 3} \left[(0-1)^2 + (0-2)^2 + (0-3)^2 \right]$$

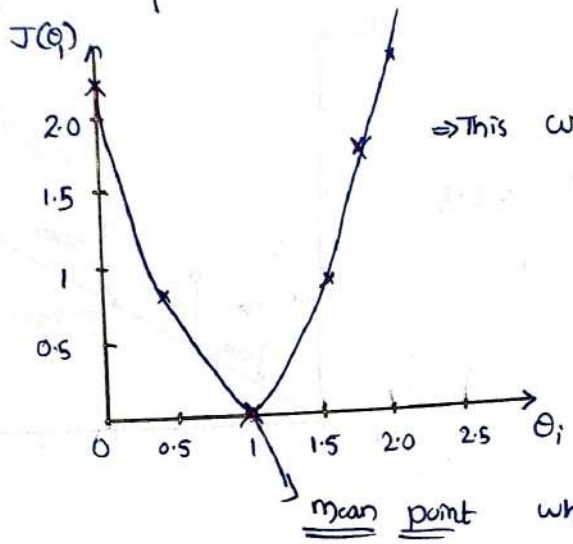
$$\approx 2.3$$

$\Rightarrow$ $\theta_1 = 0$ $\quad J(\theta_1) = 2.3$

$\theta_1 = 0.5$ $\quad J(\theta_1) = 0.58$

$\theta_1 = 1$ $\quad J(\theta_1) = 0$

After continuing calculating $J(\theta_1)$ value we get the following curve



$\Rightarrow$ This curve is called Gradient descent

mean point when we draw bestfit line using mean point error was very less. So, we can say this point is Global minima

$\Rightarrow$ we need achive the global minima point or atleast near to this point to say that error has definitely minimized.

Note:- We cannot keep on selecting different - different theta one (θ_1) values all the time. So we show apply convergence algorithm we should select one θ_1 value & try to find out mechanism of changing θ_1 values.

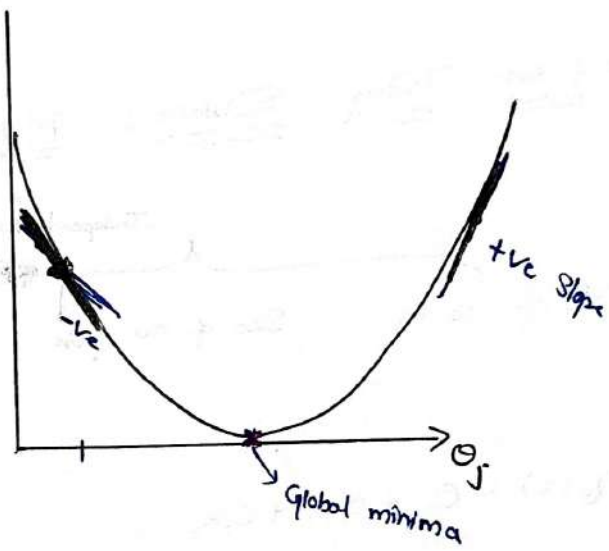
Convergence Algorithm:- The main aim is to optimize the changes of θ_j value

Repeat until Convergence {means repeat until global minima}

{

$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta_j)}{\partial \theta_j}$$ → slope

}



If Slope is -ve

$$\Rightarrow \frac{\partial J(\theta_j)}{\partial(\theta_j)} = -ve$$

$$\Rightarrow \theta_j = \theta_j - \alpha(-ve) = \theta_j + (+ve)$$ } → it add some value to θ_j

θ_j is increased until it get at or near Global minima

If Slope is +ve $\Rightarrow \theta_j = \theta_j - \alpha(+ve) = \theta_j = \theta_j - (+ve)$

It will decrease θ_j until it get at or near global minima

α - alpha is a Learning rate $\boxed{\alpha = 0.001}$ → very good practice

It controls the speed at which convergence should happen

If it is more small it will take more time to converge

If it is very-very big it will jump continuously and it may never converge

Select a small value but not very-very small

Linear Regression Example

Dataset

I/p	o/p
Weight	height
—	—
—	—

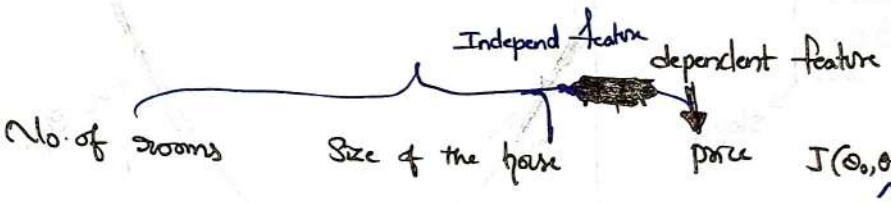
I/p or independent

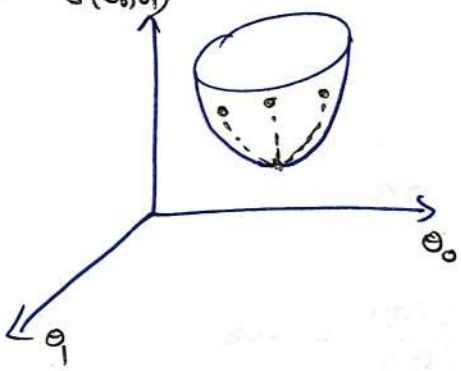
$h_\theta(x) = \theta_0 + \theta_1 x$

θ_0 = Intercept

θ_1 = slope

House Pricing Dataset :- Multiple Regression :-





$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

θ_0 = intercept

θ_1, θ_2 = slope

Multiple Regression formula:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots\cdots\cdots \theta_n x_n$$

Performance matrix used in Linear Regression:

It is used to determine whether the model is good or not for a specific problem statement

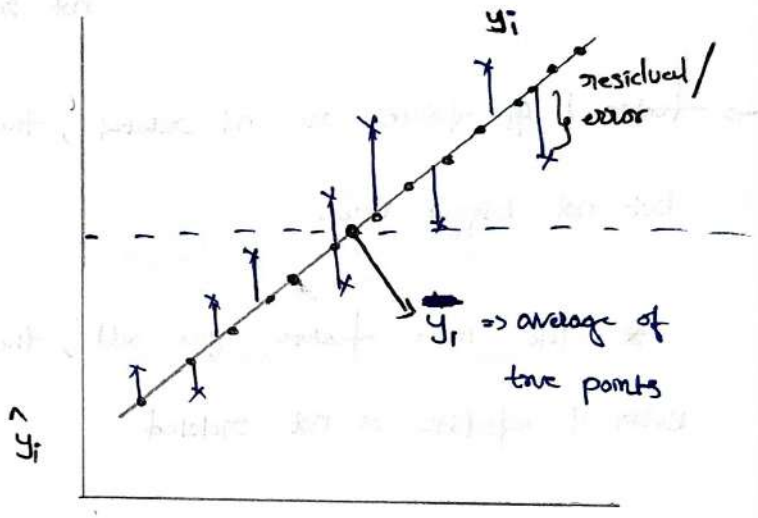
=> The matrics that basically used as

① R Squared

② Adjusted Squared

① R-Squared:-

$$R\text{-Squared} = 1 - \frac{SS_{Res}}{SS_{Total}}$$



$$= 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$$ => Smaller compare to denominator

$$= 1 - \frac{\text{Small number}}{\text{big number}}$$

$$= 1 - \text{Smaller number}$$

$$\approx 1$$

0.70 => 70% Accuracy

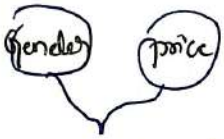
1 => 100% Accuracy

0.80 => 80%

0.94 => 94%

2] Adjust R squard :

Dataset

- Size of the house price ⇒ R - squared = 75%
- Size of the house Location price ⇒ R- Squared ⇒ 80%
- Size of the house Location 

⇒ R- Squared ⇒ 83% ⇒ 0.83

not related

⇒ feature & o/p feature are not related, then also R-squared value will increase but not bigger value

* The more feature you add, the R-Squared value is going to increase even if feature is not related

★ To overcome this we use Adjusted R-Squared

- Adjusted R-Squared = $1 - \frac{(1-R^2)(N-1)}{N-P-1}$

N = No of data points

P = No. of independent features

If the feature is related ⇒ value increase

If the feature is not related ⇒ value decrease

Ex: P=2 R^2 = 90% R^2-adjusted = 86%

one no-related feature is added

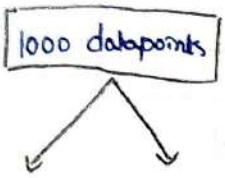
⇒ P=3 R^2 = 92% R^2-adjusted = 82%

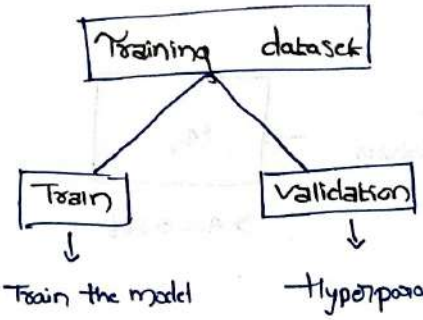
Overfitting And underfitting (Bias And Variance)

1. Training data
2. Test data
3. validation dataset

Dataset :-

Size of house	No of bedrooms	price
—	—	—
—	—	—
—	—	—
—	—	—





90%

Train | very Good Accuracy [Low bais]

Test | very Good Accuracy [Low variance]

85%

=> Generalized model

Train | very Good (90%)

Test | Bad Accuracy (50%)

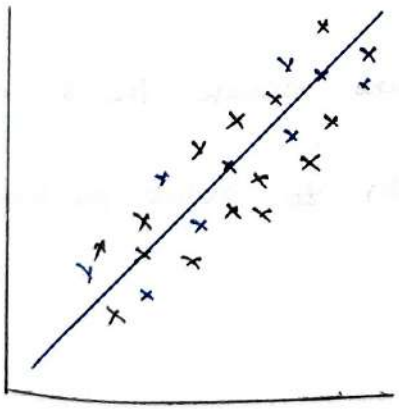
=> Model is overfitting

Train | Model Accuracy is low [High Bias]

Test | Model Accuracy is low [High variance]

⇓

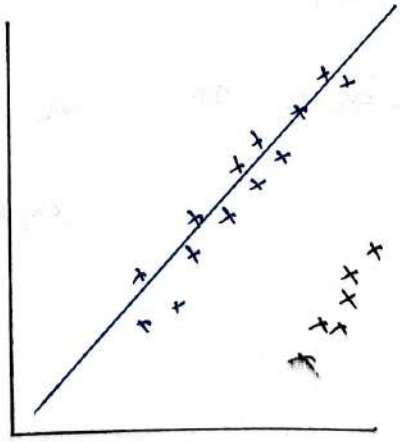
Model is underfitting



Generalized model

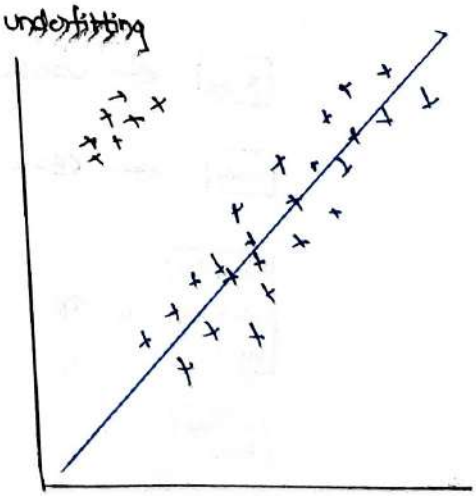
⇓

low Bias, Low variance



Overfitting

Low Bias, High variance



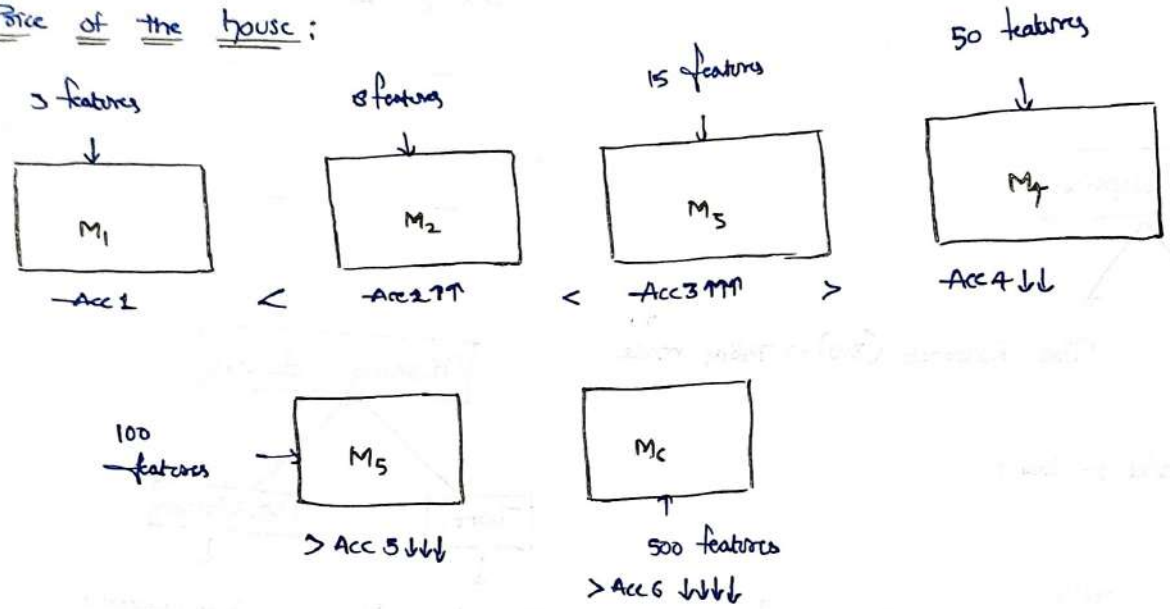
underfitting

Principal Component Analysis (PCA) [Dimensionality Reduction]

① Curse of Dimensionality :-

Dimensions {features}

=> Price of the house :



② Model performance degrade :-

A person asking the domain expert about the house price

f_1

Loc A ← 500k - 550k

3bhk ← 600k ↑↑

Beach ← 650k - 700k ↑↑

Near to celebrity house ← ↑↑

Grocery shop

As the features increase the Domain Expert confuses to predict the price

Two different ways to remove curse of Dimensionality

1 Feature Selection ⇓ The important feature are Selected

② Dimensionality Reduction (PCA) ⇓ Feature Extraction

original f_1 f_2 f_3 o/p ⇓ feature extraction D_1 D_2 o/p

Feature Selection vs Feature Extraction

↳ Dimensionality Reduction

② Why Dimensionality Reduction ?

* Prevent → Curse of Dimensionality
* Improve the performance of the model
* Visualize the data → understand the data

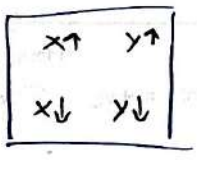
100 feature are difficult to visualize

=> 3-5 feature are easy to visualize

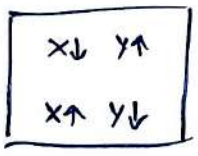
Feature Selection:

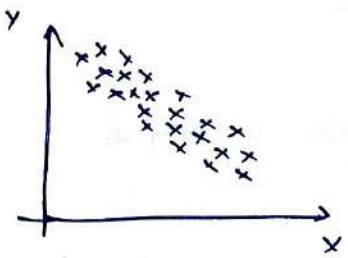
I/p	o/p
x	y
–	–
–	–
–	–

+ve

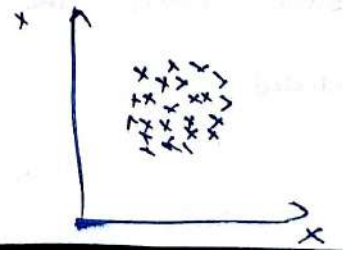


-ve





No relationship b/w x & y



$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) * (y_i - \bar{y})}{N-1}$$

It may be +ve / -ve / ≈ 0

Pearson correlation = $\frac{Cov(x,y)}{\sigma_x * \sigma_y}$ = **-1 to 1**

The more towards the value of +1 the more the +ve correlated x & y

Dataset Housing :-

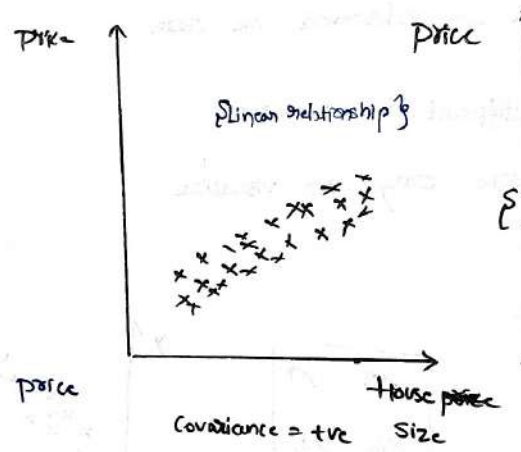
Example

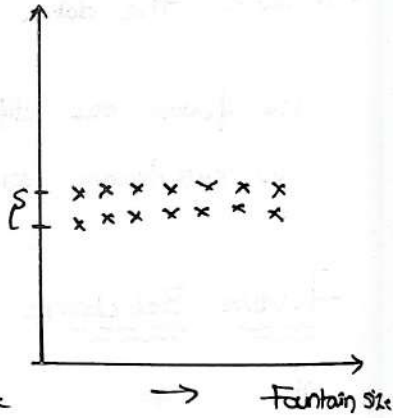
① Feature Selection :-

Independent features		
House Size	Fountain Size	price
—	—	—
—	—	—
—	—	—

Fountain size feature is droped

So, basically in feature selection the important features are selected





Covariance ≈ 0

even though fountain size increase the house price is stagnent in a specific region

=> covariance is very low or zero

If is not that important feature

Example :- Feature Extraction

Room Size	No. of Rooms	price

Dimensionality reduction 2f → 1f

We cannot use feature selection to perform dimensionality reduction because both the features are +vely correlated or -vely correlated

* We cannot drop one feature

⇒ So use feature extraction by taking the features & applying transformations to extract new features

⇒ Feature extraction :-

Room Size No. of Rooms

↓ Transformation

House price | price

Principle Component Analysis Geometric Intuitions

{ Dimensionality reduction }

=> Housing dataset

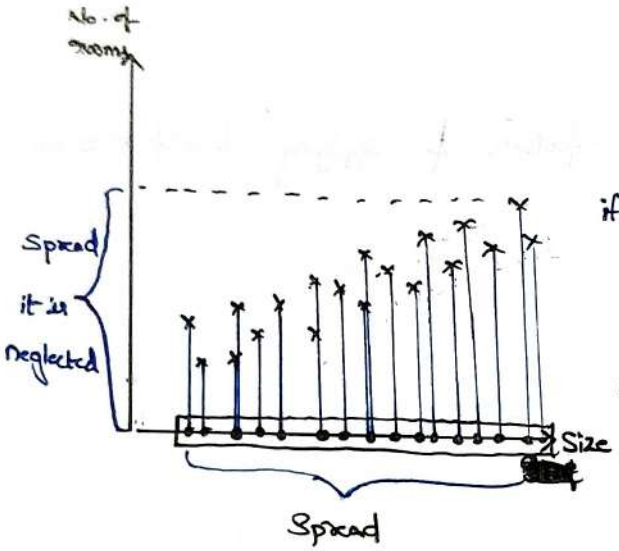
Size of house | No. of rooms | price

PCA

2D → 1D

↓

our aim



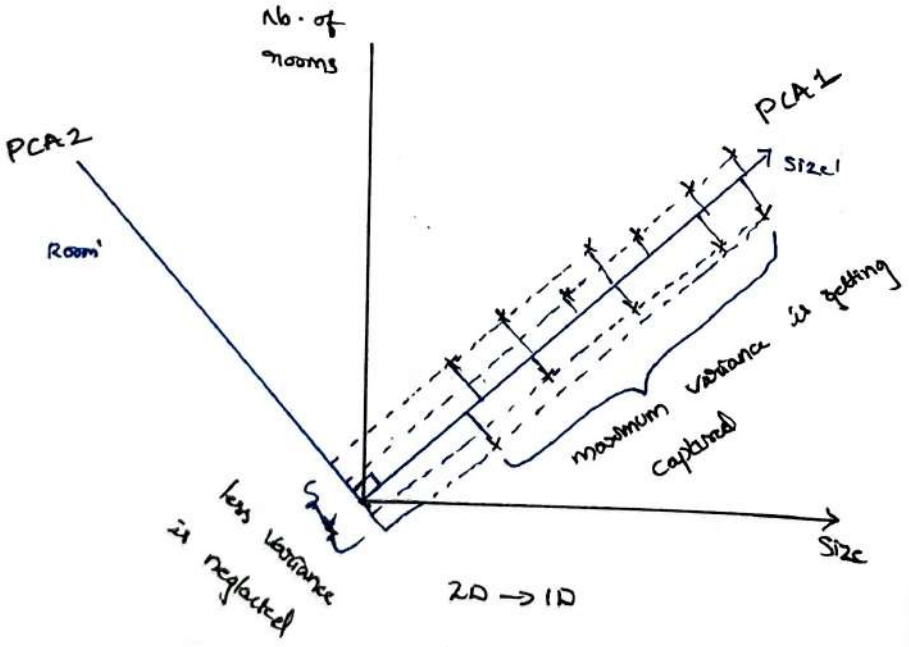
if project of x-axis a huge Spread is neglected

=> Loss of information (No of Rooms)

=> In this feature Extraction lots of information is lost

Spread increases → Variance increase

★★★ To prevent huge Loss of information PCA is used.



less information is lost

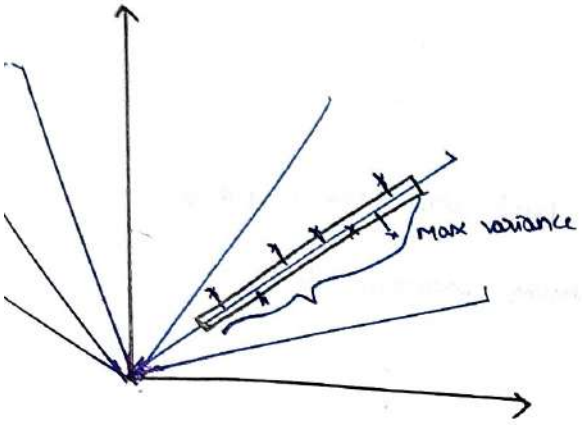
We will apply Eigen decomposition on matrix ◀ It is transformation To get new axis

3 Dimension

=> PC1, PC2, PCA

Variance(PC1) > Var(PC2) > Var(PC3)

$\Rightarrow \sigma^2(PC1) > \sigma^2(PC2) > \sigma^2(PC3)$



2D - 1D

We Select 2 Best PCA line (i.e the lines capturing high variances)

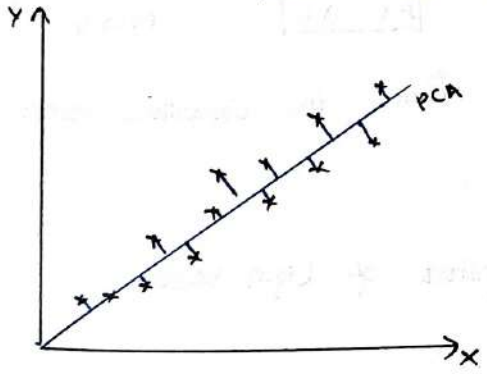
3Dimensions: To get the best principle component which captures maximum variance

$$Var(PC1) > Var(PC2) > Var(PC3)$$

3D => PC1, PC2, PC3

3D => 1D 3D - 2D

Maths Intuition behind PCA Algorithm:

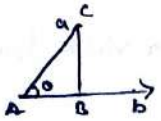


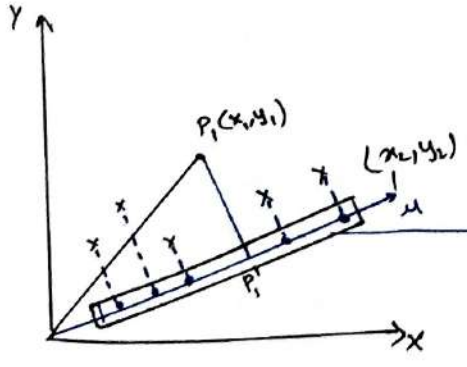
Aim

2D -> 1D

① projection

② cost function ← variance





$$Proj_{P_1} U = \frac{P_1 \cdot U}{\|U\|} \Rightarrow \boxed{proj_{P_1} U = P_1 \cdot U}$$

-> we will take all these points & calculate variance

$P_0', P_1', P_2', P_3', P_4' \ldots P_n'$ scalar values

$p_0', p_1', p_2', \ldots p_n'$

$x_0', x_1', x_2', x_3' \ldots x_n'$

$$\text{Variance} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$

{ goal :- find the best unit vector which captures maximum variance }

⇓

Cost function.

Obviously we ~~does~~ cannot selecting different-different unit vectors

The technique eigen decomposition which we specifically say as Eigen vector & Eigen values

Eigen vector & Eigen values

To find the best unit vector following steps are ~~acquired~~.

① Covariance matrix between features

② Eigen vectors & Eigen values will found out from the covariance matrix

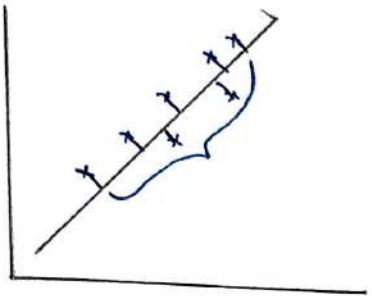
$AV = \lambda V$ linear transformation of Matrix

③ Largest Eigen vector ⇒

~~Eigen~~ Eigen vector having Eigen value high → magnitude of Eigen vector

It will capture the maximum variance

Eigen decomposition of covariance matrix :-



↳ we will find Eigen vectors & Eigen values for the matrix.

$$\begin{bmatrix}\text{covariance} \\ \text{Matrix } A\end{bmatrix} \times [v] = \lambda \times v \qquad \boxed{A \cdot v = \lambda \cdot v}$$

(Eigen value)

Linear transformation on A

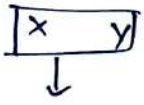
* Eigen vector having maximum magnitude that will be use as principal component and it will calculate maximum variance

$$\boxed{A \cdot v = \lambda \cdot v}$$

Eigen vector ⇒ Max magnitude ⇒ Max Eigen value ⇒ Best PC → PC1 (capture max variance)

Steps to calculate Eigen value and vectors :-

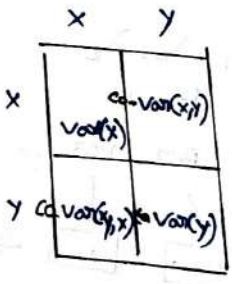
① Covariance of feature



z

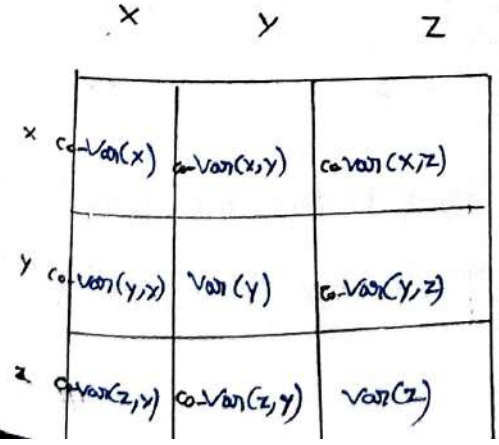
x'

$$cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$



⇒ covariance matrix = A

If it is a 2×2 matrix



$$A \cdot v = \lambda \cdot v$$

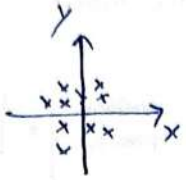
Since we are having 2 features x,y we will get 2 λ values λ_1 & λ_2 ⇒ Eigen values.

Magnitude

↓ PC1 ↓ PC2

2D → 1D

① Standardize the data



② Covariance matrix of x & y

A =

Var(x)	Covar(x,y)
Covar(y,x)	Var(y)

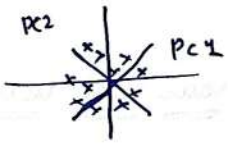
③ Find out Eigen vectors & Eigen values

$$A \cdot V = \lambda V$$

λ_1, λ_2 → Eigen values

↓ ↓

PC1 PC2



3D → 1D

$\lambda_1, \lambda_2, \lambda_3$

↓ ↓ ↓

PC1 PC2 PC3

1D 1D → 2D

3D → 1D

$\lambda_1, \lambda_2, \lambda_3$

PC1 PC2 PC3

[1D]

2D - 1D

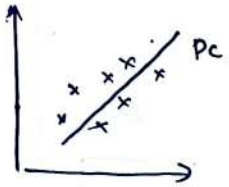
λ_1, λ_2

↓ ↓

PC1 PC2

→ projection

1D



We are trying to find best principle component that can fit for any data capturing maximum variance.

③ Perceptron :- It is a single layer neural network & it is a type of artificial neuron or the simplest form of a neural network

=> It performs binary classification that maps input features to an output features / decision, usually classifying data into one of two categories, such as 0 or 1

① Input layer
② hidden layer
③ Weights
④ Activation function

Dataset :

IQ	No. of study hours	Pass/Fail
95	3	0
110	4	1
100	5	1

Activation function :-

the aim is to transform the output b/w the value 0 to 1 {use step function}
-1 to 1

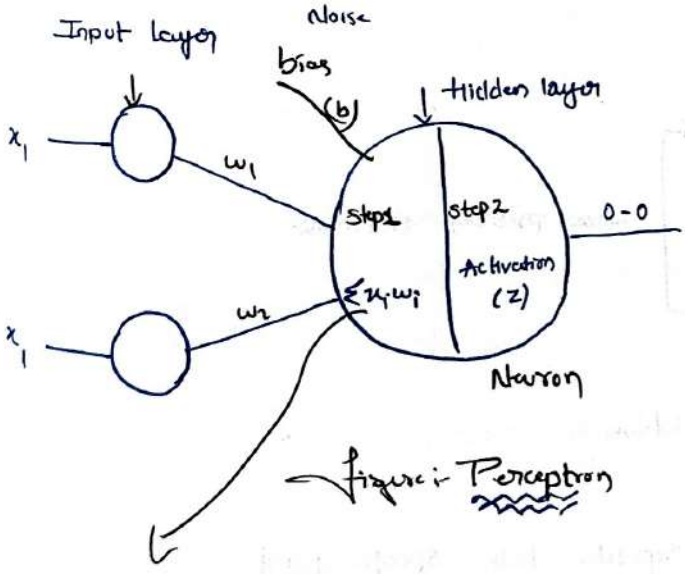


figure : Perceptron

$$Z = x_1 w_1 + x_2 w_2 + b$$

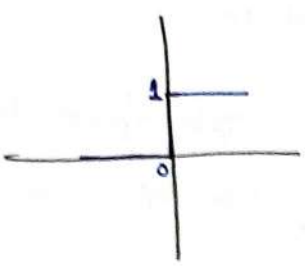
$$z = \sum_{i=1}^{n} w_i x_i + b$$

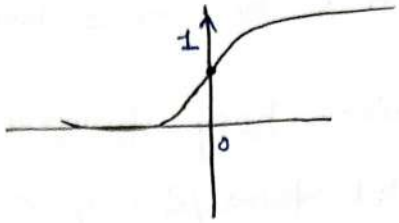
If weights are initialize zero then no processing happen. So this bias should add so that entire processing will not become zero

Noise : A data that may completely different & our model is able to handle it

The output & real output are same error is zero
if not same change weight

Step function





$$\begin{cases} 0 & z \le 0 \\ 1 & z > 0 \end{cases} \qquad \begin{cases} 1 & z > 0.5 \\ 0 & z \le 0.5 \end{cases}$$

Threshold value is change

0 → 0.5

=> If output & real output are not same then update change weights inorder to get correct output

Step 1

$$z = \sum_{i=1}^{n} w_i x_i + b$$

$$z = b + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

$$y = mx + c$$

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Linear problem statement

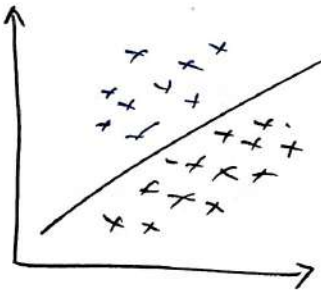
If i have the following classification problem

we will able to seperate both specific point using perseptron

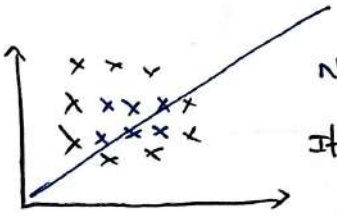
Perceptron models

Single Layer perceptron model

- It is a feed forward Neural N/w
- The real output is 0 . example we get 1 as output than again we go back & update weights & perform feed forward Neural N/w until it matches
- This not a very efficent technique Except linear Seperable problems
- problem



Binary classifier
⇓
linearly Sp Separable



Non linear Separable

It requires Multi layered NN/ perceptron model

Multi Layered perceptron model

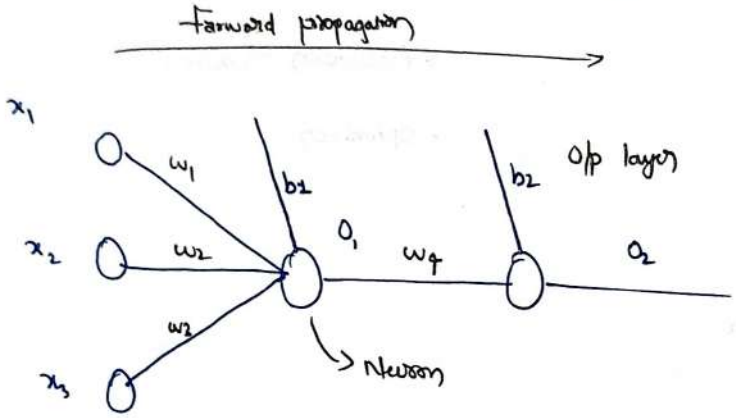
* Forward propagation
* Backward propagation
* Loss functions
* Activation functions
* optimizer

Artificial Neural network

1. Forward propagation
2. Backward propagation
3. Logs function
4. Activation function
5. Optimizers

Dataset:-

	x_1	x_2	x_3	
	IQ	study hours	play hours	pass / fail
	95	4	4	1
	100	5	2	1
	95	2	7	0



The hidden layer can have more than one Neuron

① Foward propagation:-

Hidden layer 1

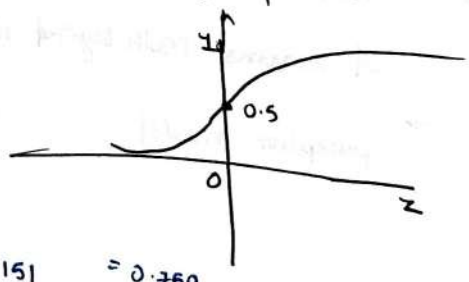
Step 1 :- $Z = \sum_{i=1}^{n} w_i^T x_i + b$

$= 95 \times 0.01 + 4 \times 0.02 + 4 \times 0.03 + 1 \times 0.01$

$= 1.151$

Step 2 :- Activation (z)

Sigmoid



$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-1.151}} = 0.759$

$O_1 = 0.759$

Hidden layer 2 :

Step 1 :- $z = o_1 * w_4 + b_2$

$= 0.759 * 0.02 + 0.03$

$= 0.04518$

Step 2 :- Activation (z)

$$z = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(0.04518)}} = 0.51129$$

$O_2 = 0.51129 = \hat{y}$ (predicted)

Loss function = $(1 \leftarrow \text{real output} - 0.51129)$

≈ 0.49 (error)

error is high our main aim is to reduce the error

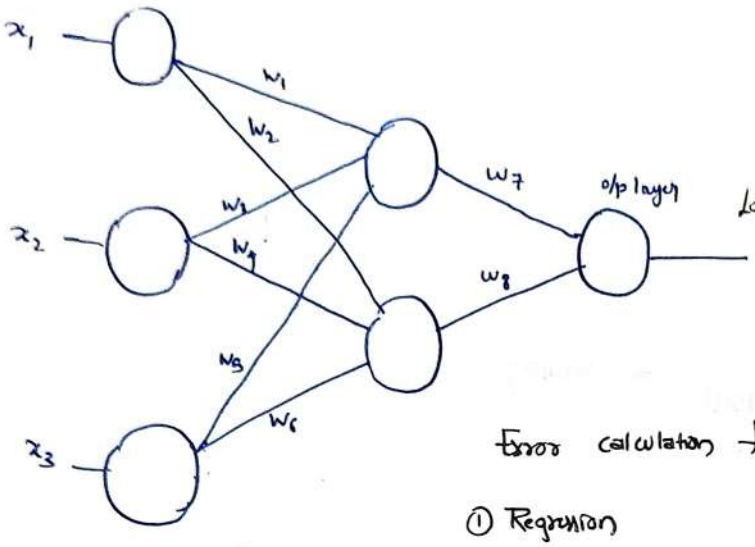
** To minimize the error the weights should be updated.

updated process will going from backward propogation & update the weight again all the steps repeated and loss function is used if error the again repeat .

Loss function vs cost function

→ $(y-\hat{y})^2$ $\sum_{i=1}^{\hat{}} (y-\hat{y})^2$

Backward propagation And Weight updation formula



Loss function $(y-\hat{y})^2$

x_1 IQ	x_2 study hours	x_3 play hours	pass/fail o/p
95	4	4	1
100	5	2	1
95	2	7	0

Error calculation functions

① Regression
- → MSE
- → MAE
- → huber loss

② classification
- → Binary class entropy
- → categorical class entropy

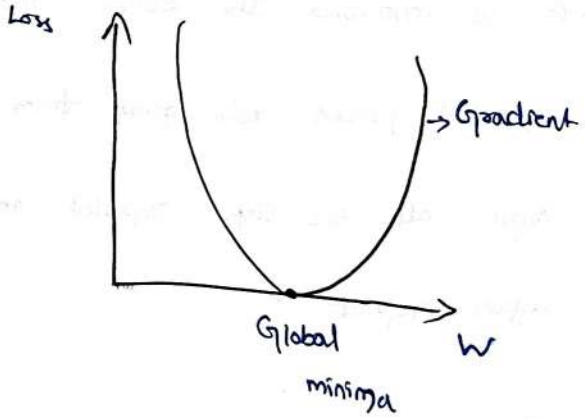
*if we get very big loss function we need to update the weight it will happen in backward propagation

Weight updation formula:

$$W_{new} = W_{old} - \eta \boxed{\frac{dL}{dW_{old}}}$$

slope

η => It is learning rate



If it is more small it will take more time to converge

If it is very big value will jump continuous & may never converge

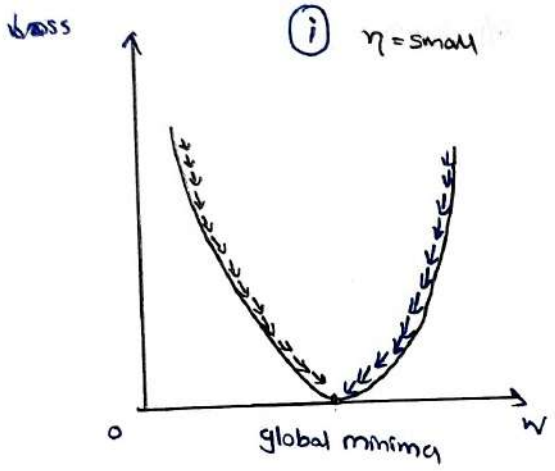
* Select a small value but not very - very small

$$W_{7_{new}} = W_{7_{old}} - \eta \frac{dL}{dW_{7dd}}$$

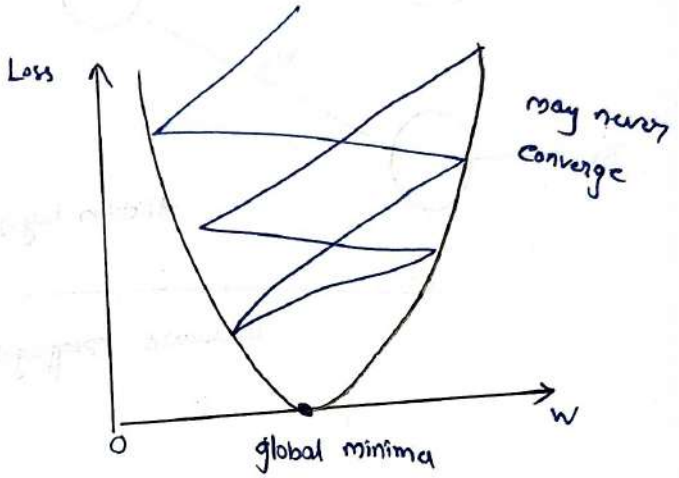
optimizers :- To reduce the loss function

Gradient Descent optimizer :

(i) η = small



(ii) η = very large



$$W_{new} = W_{old} - \eta(-ve)$$

$$= W_{old} + \eta(+ve)$$

↳ learning rate

η = 0.001 small value

$$\boxed{W_{new} >> W_{old}}$$

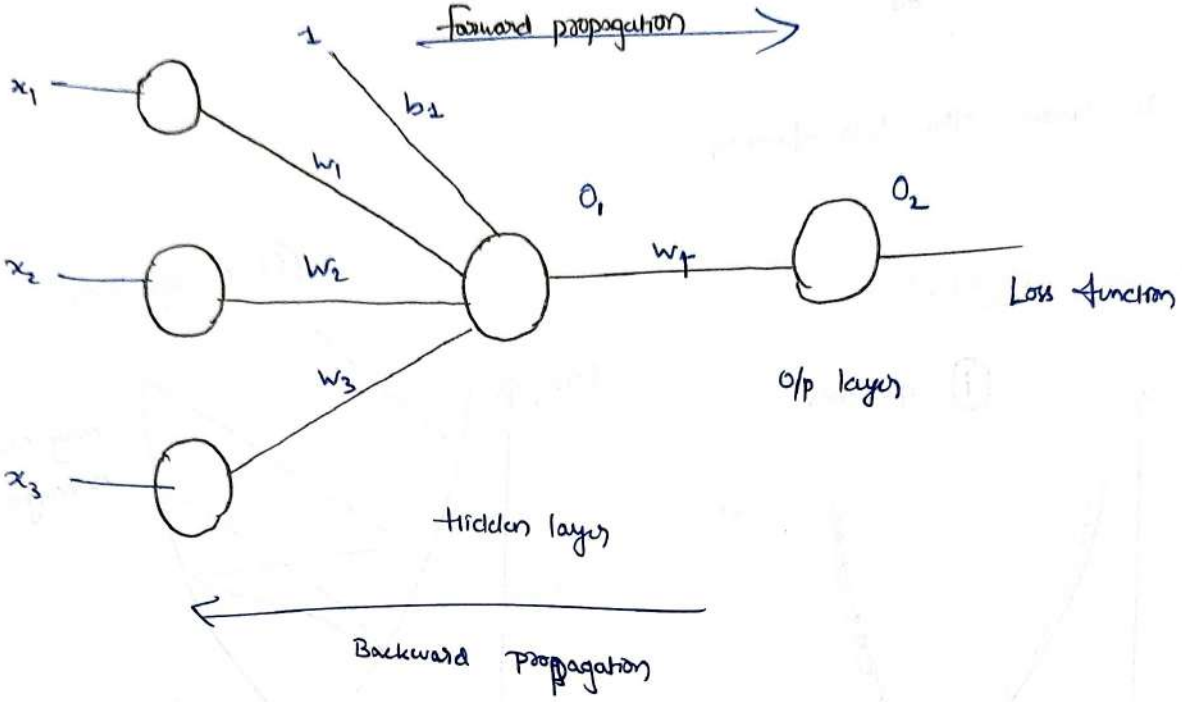
$$W_{new} = W_{old} - \eta(+ve)$$

$$= W_{old} - \eta(+ve)$$

$$\boxed{W_{new} << W_{old}}$$

It will stop when W reaches global Minima $\boxed{W_{new} = W_{old}}$

Chain Rule of Derivation :-



$$W_{4new} = W_{4old} - \eta \boxed{\frac{dL}{dW_{4old}}}$$

$$\frac{dL}{dw_{4old}} = \frac{dL}{dO_2} * \frac{dO_2}{dw_4}$$

This splitting of the derivate is basically called as chain rule of derivative

(1) $$W_{1new} = W_{1old} - \eta \frac{dL}{dW_{1old}}$$

$$\boxed{\frac{dL}{dW_{1old}} = \frac{dL}{dO_2} * \frac{dO_2}{dO_1} * \frac{dO_1}{dw_1}}$$

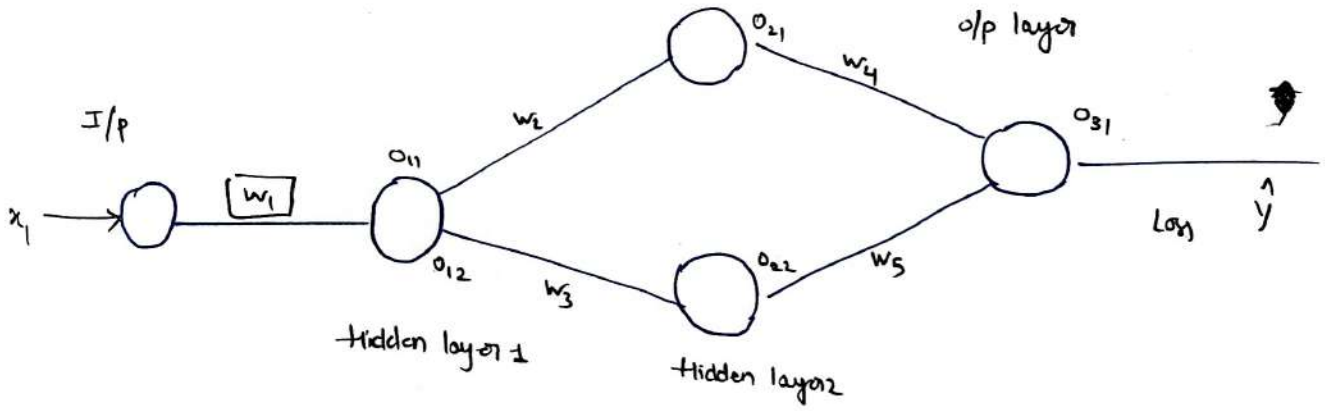
② $w_{2new} = w_{2old} - \eta \frac{dL}{dw_{2old}}$

$$\frac{dL}{dw_{2old}} = \frac{dL}{do_2} * \frac{do_2}{do_1} * \frac{do_1}{dw_2}$$

③ $w_{3new} = w_{3old} - \eta \frac{dL}{dw_{2old}}$

$$\frac{dL}{dw_{3old}} = \frac{dL}{do_2} * \frac{do_2}{do_1} * \frac{do_1}{dw_3}$$

Assignment :- How do you update w_{1new} ?



$$w_{1new} = w_{1old} - \eta \frac{dL}{dw_{1old}}$$

$$\frac{dL}{dw_{1old}} = \left[\frac{dL}{do_{31}} * \frac{do_{31}}{do_{21}} * \frac{do_{21}}{do_{11}} * \frac{do_{11}}{dw_{1old}} \right] + \left[\frac{dL}{do_{31}} * \frac{d31}{d22} * \frac{d22}{d12} * \frac{d12}{dw_{1old}} \right]$$