

Index :

ch : 1 Linear Algebra

ch : 2 Statistics

ch : 3 Probability

ch : 4 Calculus

Index = ~~dimensions of matrix~~ ~~number of rows and columns~~

## 1. Basics of Linear Algebra

1.1 scalars, vectors, Matrices, Tensors

1.2 Indexing and Slicing

## 2. Vector Operations

2.1 vector addition and subtraction

2.2 scalar multiplication

2.3 Dot product (Inner product)

2.4 cross product

2.5 Norms (L1, L2 Norms / magnitude / length)

2.6 unit vectors and direction

## 3. Matrix operations

3.1 Matrix addition and subtraction

3.2 scalar multiplication

3.3 Matrix multiplication (Dot product)

3.4 Transpose of matrix

3.5 Element wise operations

## 4. Special matrices and properties.

4.1 Identity matrix, zero matrix

4.2 Diagonal matrix, orthogonal matrix

4.3 Symmetric matrix, Inverse matrix

4.4 Rank of matrix, Trace of matrix

## 5. System of Linear Equations.

5.1 Representing Linear system as Matrix equations

5.2 Row Echelon form & Gaussian Elimination

5.3 Matrix Inverse method.

5.4 Cramer's Rule.

## 6. Matrix Inverse and Determinant

6.1 Determinant of  $2 \times 2$  and  $N \times N$  matrices

6.2 Properties of determinants.

6.3 Invertibility (condition of matrix to be invertible)

## 7. Linear Transformations

### 7.1 Function and Vector Transformation.

### 7.2 Linear Transformation

### 7.3 Linear Transformation in Data Science.

### 7.4 Projections.

## 8. Inverse Functions or Transformation.

### 8.1 Inverse of function

### 8.2 Application of Inverse Function

## 9. Eigen value and Eigen vector

## 10. Equation of Line, plane and Hyperplane

(1) Basics of Linear Algebra

## (1.1) scalars, Vectors, Matrices, Tensors

→ scalars : scalar algebra = P.E.

scalar is just a single number.

(single value with no direction, no dimension)

scalar not × scalar multiplication

eg.: Age = 21 hrs

Temperature =  $347^{\circ}\text{C}$

Price = 69 £

any number like  $\frac{2}{7}$ , 6.1 etc.

→ Vectors :

vector is a list of numbers (1D Array).

(A line of numbers - 1D (One direction)). Row or column

eg.: Juhil's marks in three subjects = [80, 85, 82]

position in 2D space [ $x=5, y=7$ ]

Daily temperature (For week): [30, 29, 41, 20

35, 32, 31]

→ Matrix :

Matrix is combination of rows and columns (2D)

(2D grid like excel sheet)

eg: Marks of 3 students in 4 subjects

[ [80, 85, 90, 75],

[70, 72, 92, 81],

[60, 78, 85, 80] ]  $(3 \times 4)$

→ Tensor:

Tensor is a multi-dimensional array

(more than 2D).

(can be 3D, 4D -- nD).

e.g.: - colored image have 3D structure

(Height x width x 3 colors - RGB)

eg. shape : [100, 100, 3]

It's mean  $100 \times 100$  pixels

and 3 color channels

- video data : [Frames, height, width, channels].

e.g. with deep explain:

one Box, its length = 5

width = 4

depth = 2

than it can be simply write like

$[5, 4, 2]$  but

in tensor it represent like

(2 depth, each with 5 rows & 4 cols)

tensor = [

  [ [ 1, 2, 3, 4 ],   :   :   : ]  
  [ 5, 6, 7, 8 ],   :   :   : ] } depth 1

  [ 9, 10, 11, 12 ],   :   :   : ]  
  [ 13, 14, 15, 16 ],   :   :   : ] } depth 1

  [ 17, 18, 19, 20 ],   :   :   : ] } depth 1

  [ 21, 22, 23, 24 ],   :   :   : ]  
  [ 25, 26, 27, 28 ],   :   :   : ] } depth 2

  [ 29, 30, 31, 32 ],   :   :   : ]  
  [ 33, 34, 35, 36 ],   :   :   : ] } depth 2

  [ 37, 38, 39, 40 ],   :   :   : ] } depth 2

  [ 41, 42, 43, 44 ],   :   :   : ] } depth 2

  [ 45, 46, 47, 48 ],   :   :   : ] } depth 2

  [ 51, 52, 53, 54 ],   :   :   : ] } depth 2

## (1.2) Indexing and Slicing

→ scalar :

scalar is a single value - so no indexing or slicing needed.

e.g.:  $x = 10$

→ Vector (1D) :

Indexing:

$v = [10, 20, 30, 40]$

0 1 2 | 3

Print( $v[0]$ ) o/p : 10

Print( $v[2]$ ) o/p : 30

Slicing : Extract(a part(range)) From vector

Print( $v[1:3]$ ) o/p : [20, 30]

Print( $v[:2]$ ) o/p : [10, 20]

Print( $v[2:]$ ) o/p : [30, 40]

→ Matrix (2D) :

Matrix is a list of lists

Indexing:

$m = [[1, 2, 3],$

[4, 5, 6],

[7, 8, 9]

]<sub>(3x3)</sub>

Print( $m[0]$ ) o/p : [1, 2, 3]

Print( $m[1][2]$ ) o/p : 6

↑      ↑  
row column

Slicing:

Print( $m[:2]$ ) o/p : First two rows

Print( $m[0][:2]$ ) o/p : [1, 2]

→ Tensor (3D or more) : what is it?

It's like a cube (list of matrices)

tensor = [

  [ # depth 0 ]

    [ 1, 2 ], similarly

    [ 3, 4 ]

  ],

[ # depth 1 ]

  [ 5, 6 ]

    [ 7, 8 ] similarly

  ],

or [ 2 ] [ 0, 1, 2 ] (row) + (col)

Indexing : [ 0, 1, 2 ] (row) + (col) + (row)

matrix print (t[0]) o/p : First full depth matrix

First row : [ 1, 2 ]

First col : [ 3, 4 ]

print (t[1][0][1])

                ↑   ↑   ↑   o/p : 6  
Depth row col

Slicing : [ 0 : 1 ] [ 0 : 1 ] [ 0 : 1 ]

print (t[1][1][0:1]) o/p : 7

Depth row col

[ 0 : 1 ] [ 0 : 1 ] [ 0 : 1 ]

print (t[1][1][1:2]) o/p : 8

Depth row col

[ 0 : 1 ] [ 0 : 1 ] [ 0 : 1 ]

print (t[1][1][1:2]) o/p : 9

(2)

Vector Operations(2.1) vector addition and subtraction

Note: vectors must be the same size  
(same number of dimensions).

i) - vector Addition:

$$\vec{a} = [2, 4, 6], \vec{b} = [1, 3, 5]$$

$$\vec{a} + \vec{b} = [2+1, 4+3, 6+5]$$

$$\vec{a} + \vec{b} = [3, 7, 11]$$

ii) - vector subtraction:

$$\vec{a} - \vec{b} = [2-1, 4-3, 6-5]$$

$$\vec{a} - \vec{b} = [1, 1, 1]$$

(2.2) scalar multiplication.

- multiplication of vectors by single number

(scalar).

if:  $\vec{v} = [v_1, v_2, v_3, \dots, v_n]$

and:  $k = \text{scalar}$

then:  $k \cdot \vec{v} = [k \cdot v_1, k \cdot v_2, k \cdot v_3, \dots, k \cdot v_n]$

e.g.:  $\vec{v} = [2, 4, 6]$

(for  $k=3$ ) resulting vector (P.S.)

$$\text{above } \vec{v} \rightarrow 3 \cdot [2, 4, 6] = [6, 12, 18]$$

Means that:

The length (magnitude) of the vector increases or decreases.

if  $k > 1$ : vector gets longer

if  $0 < k < 1$ : vector gets shorter

if  $k < 0$ : vector reverse direction.

### (2.3) Dot product (Inner product)

- The dot product of two vectors is a single number (scalar) you get by multiplying their corresponding elements and adding the results.

if:  $\vec{a} = [a_1, a_2, a_3]$ ,  $\vec{b} = [b_1, b_2, b_3]$

$$\vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2 + a_3 b_3$$

$$\therefore \vec{a} \cdot \vec{b} = \sum_{i=1}^N a_i b_i$$

eg:  $\vec{a} = [2, 3, 4]$ ,  $\vec{b} = [5, 6, 7]$

$$\vec{a} \cdot \vec{b} = 2 \times 5 + 3 \times 6 + 4 \times 7$$

$$\vec{a} \cdot \vec{b} = 10 + 18 + 28$$

$$\vec{a} \cdot \vec{b} = 56$$

- measure how much two vectors align with each other.

- If dot product = 0 then vectors are perpendicular.

### (2.4) cross product (only in 3D)

- The cross product of two 3D vectors is a new vector (not a scalar) that is perpendicular to both original vectors.

if:  $\vec{a} = [a_1, a_2, a_3]$ ,  $\vec{b} = [b_1, b_2, b_3]$

$$\vec{a} \times \vec{b} = [(a_2 b_3 - a_3 b_2), (a_3 b_1 - a_1 b_3), (a_1 b_2 - a_2 b_1)]$$

eg:  $\vec{a} = [1, 2, 3]$ ,  $\vec{b} = [4, 5, 6]$

$$\vec{a} \times \vec{b} = [(2 \times 6 - 3 \times 5), (3 \times 4 - 1 \times 6), (1 \times 5 - 2 \times 4)]$$

$$\vec{a} \times \vec{b} = [-3, 6, -3]$$

usecase:

- finding a vector that perpendicular to two vectors.
- used in physics & 3D graphics  
(not much in deep learning)

(2.5) (L1, L2 Norms/ magnitude/ length)

→ L1 Norms

This is the sum of absolute values of the vector elements

$$\|\vec{v}\|_1 = |v_1| + |v_2| + \dots + |v_n|$$

eg:  $\vec{v} = [3, -4]$

$$\|\vec{v}\|_1 = |3| + |-4| = 7$$

→ L2 Norms (Euclidean Norm)

This is the most common norm -

It tells you how long the vector is  
(like distance from origin)

$$\|\vec{v}\|_2 = \sqrt{v_1^2 + v_2^2 + \dots + v_n^2}$$

eg:  $\vec{v} = [3, 4]$

$$\|\vec{v}\|_2 = \sqrt{3^2 + 4^2} = \sqrt{9+16} = 5$$

(2.6) Unit Vector

A unit vector is a vector with a length of 1, pointing the same direction as the original.

formula :  $\hat{v} = \frac{\vec{v}}{\|\vec{v}\|_2}$

$$\text{eg: } \vec{v} = [3, 4]$$

$$\|\vec{v}\|_2 = 5$$

$$\therefore \hat{v} = \left[ \frac{3}{5}, \frac{4}{5} \right]$$

$$\therefore \hat{v} = [0.6, 0.8]$$

$$\rightarrow \|\hat{v}\| = \sqrt{0.6^2 + 0.8^2} = 1$$

(3)

### Matrix operations:

$$\text{level 1: } A + B + C = \|ABC\|$$

#### (3.1) Matrix addition & subtraction

- You can add or subtract two matrices only if they have same dimensions.

$$\text{eg: } A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$$

$$A+B = \begin{bmatrix} 1+5 & 2+6 \\ 3+7 & 4+8 \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix}$$

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

#### (3.2) Scalar multiplication

- Multiply every element of matrix by scalar value.

$$\text{eg } A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, k=3$$

$$\therefore k \cdot A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot 3 = \begin{bmatrix} 3 & 6 \\ 9 & 12 \end{bmatrix}$$

## (3.3) Matrix Multiplication (Dot product):

Rule :

- you can multiply two matrices

 $A(m \times n)$  and  $B(n \times p)$  if:

The number of columns in A

= The number of rows in B.

- Multiply row of A with column of B and then sum the products.

e.g.:  $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}_{(2 \times 2)}$ ,  $B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}_{(2 \times 2)}$

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

$A \cdot B = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$

## (3.4) Transpose of matrix : i)

- flip rows into columns (and vice versa)

e.g.:

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

 $(2 \times 3)$ 

$$A^T = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}_{(3 \times 2)}$$

## (3.5) Element wise operation in matrix:

## (i) Element wise Addition

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

$$\therefore A + B = \begin{bmatrix} 1+5 & 2+6 \\ 3+7 & 4+8 \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix}$$

## (ii) Element wise subtraction

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

## (iii) Element wise multiplication

(Hadamard product)

$$\therefore E = A \circ B$$

$$\therefore A \circ B = \begin{bmatrix} 1 \times 5 & 2 \times 6 \\ 3 \times 7 & 4 \times 8 \end{bmatrix} = \begin{bmatrix} 5 & 12 \\ 21 & 32 \end{bmatrix}$$

## (iv) Element wise division

$$\therefore A / B = \begin{bmatrix} 1/5 & 2/6 \\ 3/7 & 4/8 \end{bmatrix} = \begin{bmatrix} 0.2 & 0.33 \\ 0.42 & 0.5 \end{bmatrix}$$

It's all operations useful in:

- Image processing : Adjust Brightness etc
- ML : Apply activation Funct<sup>n</sup>, loss funct
- Data Analysis
- Scientific computation.

## (4) Special matrix and properties :

(4.1) - Identity matrix and zero matrix

- main diagonal (From top left  $\rightarrow$  bottom right) are 1, and other elements are 0.

$$\text{eg: } I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, J = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$I = A$  - Zero matrix have all elements are 0.

$$O_{2 \times 3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Properties:  $\rightarrow A + 0 = A$  &  $0 + A = A$

$\rightarrow AO = 0$  &  $OA = 0$

(4.2) Diagonal matrix and Orthogonal matrix

- main diagonal have different values and other elements are 0.

$$\text{eg: } D = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

- Orthogonal matrix:

properties:  $A^T \cdot A = I$  or  $A^T = A^{-1}$

(Transpose x matrix = Identity)

$$\text{eg: } A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, A^T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\therefore A \cdot A^T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

## (4.3) Symmetric Matrix &amp; Inverse Matrix

- symmetric matrix :

A square matrix where  $A = A^T$

matrix  $A$  is equal to its transpose  $A^T$ .

$$\text{eg : } A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}, A^T = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$$

This is symmetric matrix

- Inverse matrix :

$A \cdot A^{-1} = I$  &  $A^{-1} \cdot A = I$

How to find Inverse matrix

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

Now Interchange 1 & 4 and

change sign of 2 & 3.

$$D = AD : \therefore A^{-1} = \begin{bmatrix} 4 & -3 \\ -2 & 1 \end{bmatrix}$$

Not divided by It's determinant.

$$\therefore A^{-1} = \frac{1}{-2} \begin{bmatrix} 4 & -3 \\ -2 & 1 \end{bmatrix}$$

## (4.4) Rank of matrix &amp; Trace of matrix

- Rank of matrix :

The rank of matrix is the number of linearly independent rows or columns.

- Independent = "not a copy or

combination of other"

- Rank tell us how much actual information is in the matrix.

→ Methods to find Rank of matrix.

(i) Row Echelon Form (REF):

row operations:

$$(i) R_2 \rightarrow R_2 - 2R_1$$

$$(ii) R_3 \rightarrow R_3 - R_1$$

$$(iii) R_3 \leftrightarrow R_2$$

→ use row operation to convert the matrix into upper triangular

$$\begin{bmatrix} & & \\ & & \\ \vdots & & \end{bmatrix}$$

e.g.  $A =$

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 1 & 1 \end{bmatrix}$$

Now, performing row operation.

$$(i) R_2 \rightarrow R_2 - 2R_1$$

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

$$(ii) R_3 \rightarrow R_3 - R_1$$

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

$$(iii) R_3 \leftrightarrow R_2$$

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

(upper triangular form).

→ Total non-zero rows = 2,  $\therefore \text{Rank} = 2$

(ii) using determinants.  
(For square sub-matrices)

- Find largest square submatrix
- with non-zero determinant.
- The size of the submatrix  
= rank.

Eg:  $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 1 & 1 \end{bmatrix}$

- largest square submatrix is  $3 \times 3$ .
- $\therefore d(3 \times 3) = 0 \therefore \text{Rank} \neq 3$
- Now  $2 \times 2$  square submatrix  
(at least one) with determinant  
of non-zero means it's  
rank is 2.

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

$$\therefore d(A) = (1 \times 4) - (2 \times 2)$$

$$\therefore d(A) = 4 - 4 = 0$$

→ Now change  $(2 \times 2)$  matrix.

$\therefore A = \begin{bmatrix} 2 & 3 \\ 4 & 6 \end{bmatrix}$

$$\therefore d(A) = (2 \times 6) - (4 \times 3) = 0$$

→ Again change  $(2 \times 2)$  matrix

$\therefore A = \begin{bmatrix} 4 & 6 \\ 1 & 1 \end{bmatrix}$

$$\therefore d(A) = (4 \times 1) - (6 \times 1) = (-2)$$

(Non zero)  $\therefore \text{Rank} = 2$

- Trace of matrix

It's sum of all diagonal elements of square matrix.

$$\text{eg : } A = \begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix} \therefore \text{Trace} = 2+5 = \underline{\underline{7}}$$

## (5) System of Linear equations :

(5.1) Representing Linear system as Matrix equations

- Let's say we have

$$\begin{cases} 2x + 3y = 8 \\ 4x - y = 2 \end{cases}$$

- coefficients :  $\begin{bmatrix} 2 & 3 \\ 4 & -1 \end{bmatrix}$

- Variables :  $\begin{bmatrix} x \\ y \end{bmatrix}$

- Constants :  $\begin{bmatrix} 8 \\ 2 \end{bmatrix}$

- Matrix equation form :

$$A \cdot X = B$$

$\therefore A$  = coefficient

$\therefore X$  = variables

$\therefore B$  = constants

$$\text{so, } \begin{bmatrix} 2 & 3 \\ 4 & -1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 8 \\ 2 \end{bmatrix}$$

eg: 3 equations, 3 variables

$$\left\{ \begin{array}{l} x + 2y + 3z = 14 \\ 4x + 5y + 6z = 32 \\ 7x + 8y + 9z = 50 \end{array} \right.$$

Matrix form of 3 equations.

$$\left[ \begin{array}{ccc|c} 1 & 2 & 3 & x \\ 4 & 5 & 6 & y \\ 7 & 8 & 9 & z \end{array} \right] = \left[ \begin{array}{c} 14 \\ 32 \\ 50 \end{array} \right]$$

### (5.2) Row Echelon Form & Gaussian Elimination

→ what is Row Echelon Form?

- All zero rows (if any) are at the bottom.
- Leading entry from left side of row should be 1. (not for all matrix)

Example matrix in REF.

$$\left[ \begin{array}{ccc|c} 1 & 2 & 1 & 4 \\ 0 & 1 & 3 & 7 \\ 0 & 0 & 1 & 2 \end{array} \right]$$

$$\left[ \begin{array}{ccc|c} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \end{array} \right]$$



This is  
REF Form

$\Rightarrow$  Gaussian Elimination.

$$\left\{ \begin{array}{l} x + 2y + z = 9 \\ 2x + 3y + 3z = 21 \\ 3x + y + 2z = 14 \end{array} \right.$$

Step 1: Augmented matrix

$$\left[ \begin{array}{ccc|c} 1 & 2 & 1 & 9 \\ 2 & 3 & 3 & 21 \\ 3 & 1 & 2 & 14 \end{array} \right]$$

Step 2: convert (2,1) & (3,1) positioned values into (0).

$$\therefore R_2 \rightarrow R_2 - 2R_1$$

$$\therefore R_3 \rightarrow R_3 - 3R_1$$

$$\therefore \left[ \begin{array}{ccc|c} 1 & 2 & 1 & 9 \\ 0 & -1 & 1 & 3 \\ 0 & -5 & -1 & -13 \end{array} \right]$$

Step 3: convert (3,2)- positioned value into (0).

$$\therefore R_3 \rightarrow R_3 - 5R_2$$

$$\therefore \left[ \begin{array}{ccc|c} 1 & 2 & 1 & 9 \\ 0 & -1 & 1 & 3 \\ 0 & 0 & -6 & 2 \end{array} \right]$$

## Step 4: Back substitution.

→ Start from the last row.

$$\therefore -6z = 2$$

$$\therefore z = -\frac{1}{3}$$

→ now second last row

$$\therefore -y + z = 3$$

$$\therefore -y - \frac{1}{3} = 3$$

$$\therefore y = -3.33$$

Substituting (1, 8) & (1, 9) from row 1 & 2

→ Now first row

$$\therefore x + 2y + z = 9$$

$$\therefore x + 2(-3.33) + \left(-\frac{1}{3}\right) = 9$$

$$\therefore x + (-6.66) - \frac{1}{3} = 9$$

$$\therefore x - 6.66 - \frac{1}{3} = 9$$

Now by formula  $3x - (19.98 - 1) = 27$

$$\therefore 3x = 27 + 19.98 + 1$$

$$\therefore 3x = 48.98$$

$$\therefore x = \frac{48.98}{3} \quad \therefore x = \underline{\underline{16}}$$

## (5.3) Matrix Inverse Method.

suppose, we have equations like

$$\begin{cases} 2x + 3y = 8 \\ 4x + y = 10 \end{cases}$$

$$\text{matrix form: } \begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 8 \\ 10 \end{bmatrix}$$

(A)                    (x)                    (B)

$$\therefore A \cdot x = B$$

$$\text{formula: } x = A^{-1} \cdot B$$

find  $A^{-1}$

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

$$\therefore A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$\therefore A^{-1} = \frac{1}{-10} \begin{bmatrix} 1 & -3 \\ -4 & 2 \end{bmatrix}$$

$$\therefore A^{-1} = \begin{bmatrix} -0.1 & 0.3 \\ 0.4 & -0.2 \end{bmatrix}$$

$$\rightarrow x = A^{-1}B = \begin{bmatrix} -0.1 & 0.3 \\ 0.4 & -0.2 \end{bmatrix} \cdot \begin{bmatrix} 8 \\ 10 \end{bmatrix}$$

$$\therefore x = (-0.1)(8) + (0.3)(10) = 2.2$$

$$\therefore y = (0.4)(8) + (-0.2)(10) = 1.2$$

$$\therefore x = 2.2, y = 1.2$$

## (5.4) Cramer's Rule

- Cramer's Rule is a method to solve a system of linear equation using determinants.
- It works only when (square matrix)  $(2 \times 2)$  or  $(3 \times 3)$
- Determinant of the coefficient matrix is not zero.

eg :  $2x + 3y = 8$

$4x + y = 10$

$\therefore AX = B$

$\therefore A = \begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix}$  (coefficient matrix)

$\therefore B = \begin{bmatrix} 8 \\ 10 \end{bmatrix}$

Step 1 :- find Determinant of A.

$\therefore \det(A) = 2 \times 1 - 3 \times 4$

$\det(A) = 2 - 12$

$\det(A) = -10$

Step 2 :- Replace each column of A with B to get  $Ax$  &  $Ay$

$$\therefore Ax = \begin{bmatrix} 8 & 3 \\ 10 & 1 \end{bmatrix}$$

$$\therefore \det(Ax) = 8 \times 1 - 3 \times 10 = (-22)$$

$$\therefore Ay = \begin{bmatrix} 2 & 8 \\ 4 & 10 \end{bmatrix}$$

$$\therefore \det(Ay) = 2 \times 10 - 4 \times 8 = (-12)$$

Step 3: Apply cramer's rule

$$\therefore x = \frac{\det(Ax)}{\det(A)} = \frac{-22}{-10} = 2.2$$

$$\therefore y = \frac{\det(Ay)}{\det(A)} = \frac{-12}{-10} = 1.2$$

$$\rightarrow \therefore x = 2.2 \text{ & } y = 1.2$$

## (6) Matrix Inverse and Determinant :

### (6.1) Determinant of $(2 \times 2)$ and $N \times N$ matrices.

→ Determinant is a special number calculated from square matrix.

#### (1) Determinant of $2 \times 2$ matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \rightarrow \det(A) = ad - bc$$

$$\text{eg: } A = \begin{bmatrix} 3 & 4 \\ 2 & 5 \end{bmatrix} \rightarrow \det(A) = (3)(5) - (2)(4) \\ \therefore \det(A) = 7$$

(2) Determinant of  $3 \times 3$  or  $n \times n$  matrix

e.g. if  $A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 5 & 6 & 0 \end{bmatrix}$

$$\det(A) = 1 \cdot \begin{vmatrix} 1 & 4 \\ 6 & 0 \end{vmatrix} - 2 \cdot \begin{vmatrix} 0 & 4 \\ 5 & 0 \end{vmatrix} + 3 \cdot \begin{vmatrix} 0 & 1 \\ 5 & 6 \end{vmatrix}$$

(using cofactors rule)

$$\therefore \det(A) = 1 \times (-24) - 2 \times (-20) + 3 \times (-5)$$

$$\therefore \det(A) = -24 + 40 - 15$$

$$\therefore \det(A) = 1$$

## (6.2) Properties of Determinants.

(i) only for square matrices.

(ii) if  $\det(A) = 0$ , then  $A$  is non-invertible

(iii) Multiplying a row  $\downarrow$  multiplies (by  $k$ ) determinant by  $k$ .

(iv)  $\det(AB) = \det(A) \times \det(B)$

(v)  $\det(A^T) = \det(A)$

## (6.3) Invertibility.

→ A square matrix  $A$  is invertible

(also called non-singular), if there exists another matrix  $A^{-1}$  such that:

$$A \cdot A^{-1} = A^{-1} \cdot A = I$$

$\therefore A^{-1}$  = Inverse of  $A$

$\therefore I$  = Identity matrix

→ When is a matrix Invertible?  
only if :

- It is a square matrix ( $2 \times 2$ ) ( $n \times n$ )
- Its determinant is not zero.

$$\therefore \det(A) \neq 0$$

When  $\det(A) = 0$  then

- The matrix is singular
- It has no inverse
- System of equations may have no or infinite solutions.

→ Diff. between singular vs Non-singular

Matrix.

Type	Condition	Inverse exists
singular	$\det(A) = 0$	No
Non-singular	$\det(A) \neq 0$	Yes

→ Compute Inverse of matrix.

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

$$\therefore A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$\text{Ex: } A = \begin{bmatrix} 2 & 1 \\ 5 & 3 \end{bmatrix}$$

$$\therefore A^{-1} = \frac{1}{(2)(3) - (1)(5)} \begin{bmatrix} 3 & -1 \\ -5 & 2 \end{bmatrix}$$

$$\therefore A^{-1} = \begin{bmatrix} 3 & -1 \\ -5 & 2 \end{bmatrix}$$

(7) Linear Transformations:

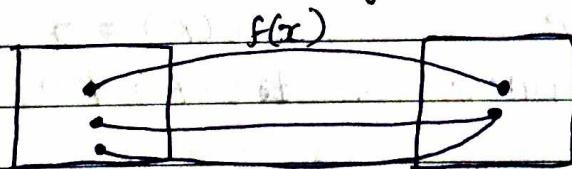
## (7.1) Function and vector transformation

Function: A function is a mathematical relationship that uniquely associates elements of one set (Domain) with elements of another set (codomain).

In simple, a function maps inputs to output in a specific way.

$$f : x \rightarrow y$$

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



$$f : \begin{bmatrix} x \\ y \\ z \end{bmatrix} \in \mathbb{R}^3 \rightarrow \begin{bmatrix} xy \\ yz \end{bmatrix} \in \mathbb{R}^2$$

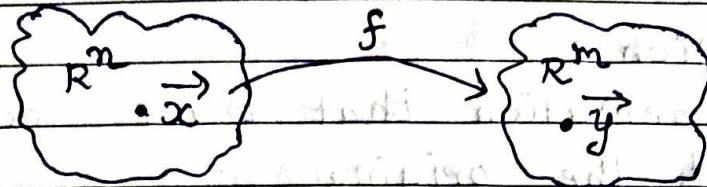
→ function ( $f$ ) convert 3 dimension to 2 dimension.

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

∴  $f(x) \Rightarrow$  Transformation.

vector transformation:

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



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

eg:  $\vec{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, f(x, y, z) = (x + y, 2z)$

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

→ This function  $\vec{x} \in \mathbb{R}^3$  transformed to  $\vec{y} \in \mathbb{R}^2$ .

$$\therefore f : \mathbb{R}^3 \rightarrow \mathbb{R}^2$$

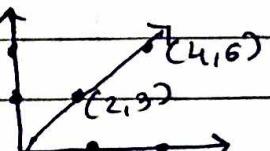
## definition of vector transformation:

- vector transformation refer to operations that map vectors from one space ( $\mathbb{R}^n$ ) to another space ( $\mathbb{R}^m$ ), often changing their magnitude, direction or both.

eg: (i) scaling:

scaling is a transformation that change the magnitude of vector but direction remain same.

$$v' = 2v = 2 \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$



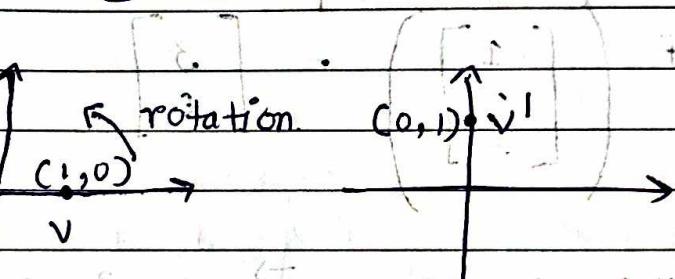
Application : (i) data normalization  
(ii) Resize objects  
(image, paint etc).

### (2) Rotation :

Transformation that turns vectors around the origin.

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

$$v' = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \in \mathbb{R}^2$$



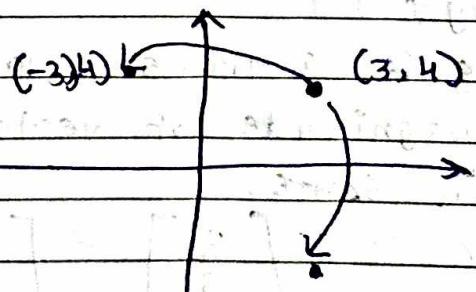
Application : (i) Image processing  $\Rightarrow$  Rotate Image.  
(ii) 3D graphics  $\Rightarrow$  Rotating object.

### (3) Reflection :

Transformation that flip vectors

over a specific axis or plane.

$$v = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \Rightarrow \text{across the } y\text{-axis}$$



Application :  
• Mirroring Images.

## (7.2) Linear Transformation.

- A Linear transformation is a function between two vector space that preserve the operation of vector addition and scalar multiplication
- This mean that if  $T$  is a linear transformation from a vector space  $V$  to a vector space  $W$ , then

$T: V \rightarrow W \Rightarrow$  Linear transformation.  
properties.

- (i) Additivity  $T(u+v) = T(u) + T(v)$
- (ii) Homogeneity  $T(cu) = c \cdot T(u)$

$\therefore u, v \in V$  and  $c$  is scalar value.

e.g.: Reflection

$$x = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2, \quad T(x) = \begin{bmatrix} -x \\ y \end{bmatrix}, \quad A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$T(x) = A \cdot x$$

$$T(x) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -x \\ y \end{bmatrix}$$

(i) check Additivity:

Let  $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$  be two vector

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

∴ LHS : <sup>For</sup> ~~Left hand side~~ ~~is~~ ~~left hand side~~

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

$$\text{Now } T(u+v) = A \cdot (u+v)$$

$$= \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1+v_1 \\ u_2+v_2 \end{bmatrix}$$

$$\text{LHS} = T(u+v) = \begin{bmatrix} -(u_1+v_1) \\ u_2+v_2 \end{bmatrix}$$

$$(v_1 T + u_1 T) + (v_2 T + u_2 T) \quad \text{(i)}$$

$$(v_1 T + u_1 T) + (v_2 T + u_2 T) \quad \text{(ii)}$$

∴ For RHS :

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

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

$$\text{RHS} = T(u) + T(v)$$

$$= \begin{bmatrix} -u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} -v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} -(u_1+v_1) \\ u_2+v_2 \end{bmatrix}$$

$$\therefore \text{LHS} = \text{RHS} \quad \text{(i)}$$

First condition satisfied.

(ii) check Homogeneity : real part

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

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

For LHS :

$$cu = c \cdot \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} cu_1 \\ cu_2 \end{bmatrix}$$

$$\therefore T(cu) = \bar{A} \cdot (cu)$$

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

For RHS :

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

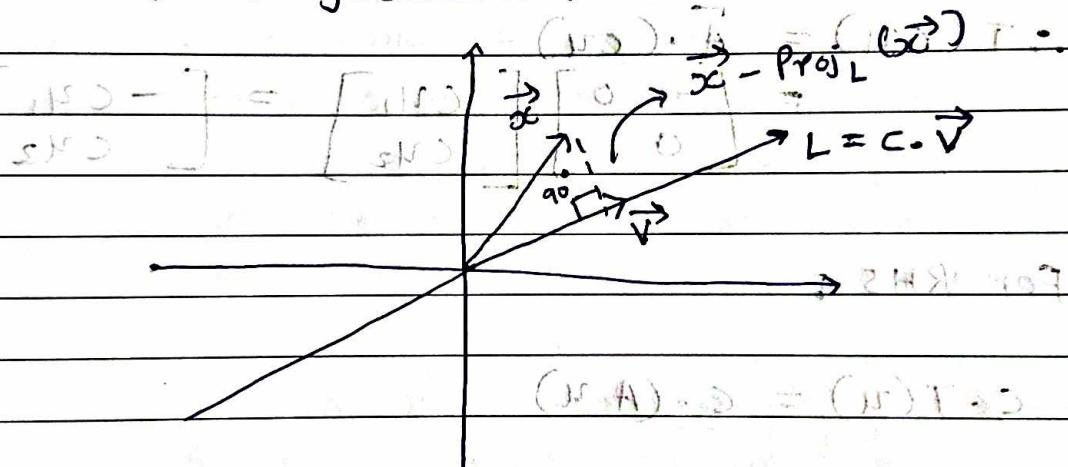
$$= c \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = c \begin{bmatrix} -u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} -cu_1 \\ cu_2 \end{bmatrix}$$

$\therefore$  LHS = RHS ∴ satisfied  
 $\therefore$  Second condition also satisfied.

## (7.3) Linear Transformation in Data Science.

- Dimensionality Reduction (PCA)
- Feature engineering
- Data Preprocessing  
Normalization and Standardization
- Neural Networks (Forward propagation, Activation Function)
- Image and signal processing.
- Optimization

## (7.4) Projections :



$\text{Proj}_L(\vec{x}) \Rightarrow$  Projection of  $\vec{x}$  on the Line L.

→ Note: Projected vector is perpendicular to  $\vec{x} - \text{Proj}_L(\vec{x})$ .

Formula: projection of  $\vec{x}$  on vector v.

$$\begin{aligned}\text{Proj}_v(\vec{x}) &= c \cdot \vec{v} \\ &= \left( \frac{\vec{x} \cdot \vec{v}}{\vec{v} \cdot \vec{v}} \right) \cdot \vec{v}\end{aligned}$$

OR simple formula :  
Projection of  $\vec{a}$  onto  $\vec{b}$  is :

$$\text{Proj}_{\vec{b}} \vec{a} = \frac{\vec{a} \cdot \vec{b}}{\|\vec{b}\|^2} \cdot \vec{b}$$

eg :  $\vec{a} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \vec{b} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

$$\therefore \vec{b} = \left\{ c \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} \right\}$$

$$\text{Proj}_{\vec{b}} \vec{a} = \frac{\vec{a} \cdot \vec{b}}{\|\vec{b}\|^2} \cdot \vec{b}$$

$$\therefore = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$= ((2)(2) + (3)(1)) \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 7 \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 14 \\ 5 \end{bmatrix}$$

$$\therefore \text{Proj}_{\vec{b}} \vec{a} = \begin{bmatrix} 2.8 \\ 1.4 \end{bmatrix}$$

(8)

## Inverse Function or Transformation

### (8.1) Inverse of Function

- Inverse of Function is a function that "Reverse" the effect of original Function.
- If you have a function  $f$  that maps an element  $x$  from set  $X$  to an element  $y$  in a set  $Y$ , the inverse function  $f^{-1}$  map  $y$  back to  $x$ .

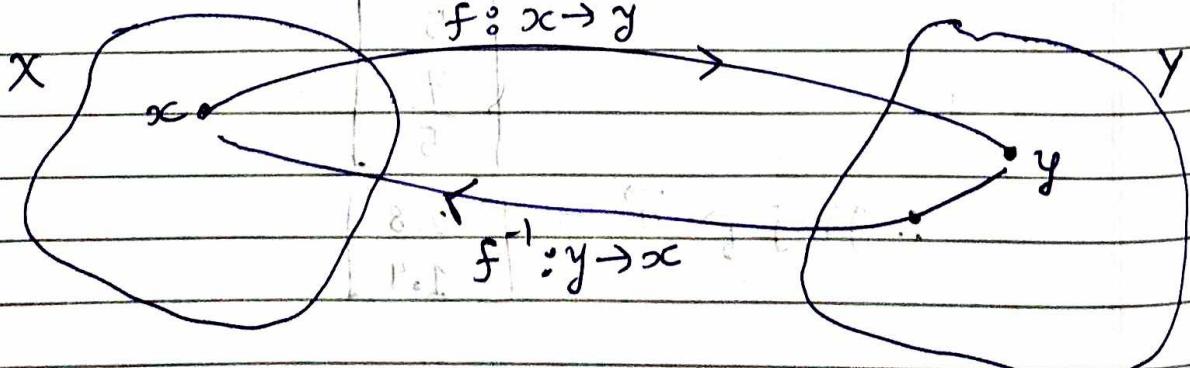
function  $f: X \rightarrow Y$

Inverse function  $f^{-1}: Y \rightarrow X$

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



→ Identity function :

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

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

$$Ix = x \rightarrow x \quad \text{and} \quad Ix(a) = a$$

where }  $a \in x$

→ For a set  $X$ , the Identity Function ( $I_x$ )

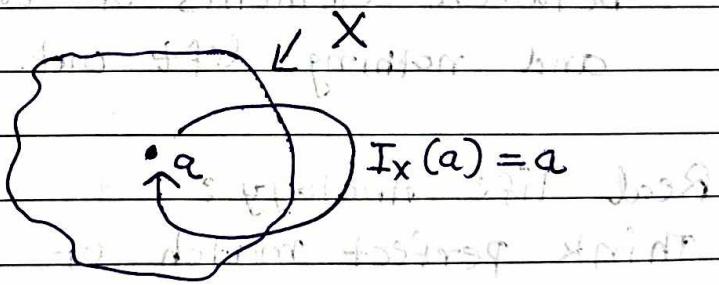
is defined as ( $=$ )

$I_x(a) = a$  for all  $a \in X$ .

→  $I_x$  is the identity function on the set  $X$

and it maps every element  $x$  in  $X$  to

itself.



→ Properties of Identity Function :

(i) Preservation : Does not alter any element.

If  $x$  is the domain, then the image of  $x$  under the identity fn is  $x$ .

(ii) Identity Function is a linear transformation

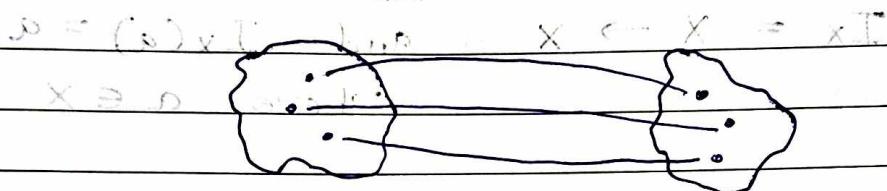
$$- I(u+v) = I(u) + I(v)$$

$$- I(cu) = c \cdot I(u)$$

→ Existence and Uniqueness:

- A function  $f$  has an inverse if and only if it is bijective.

(i) Injective (One to one): Different element in the domain map to different element in codomain.



(ii) Surjective (onto): Every element in  $S \ni s$  codomain is the image of atleast one element in the domain.

$X \ni x$  and no element left in  $S$  -

- Bijective function is a perfect pairing between elements of two sets - no duplicates, and nothing left out.

Real life Analogy:

Think perfect match of dance partners:

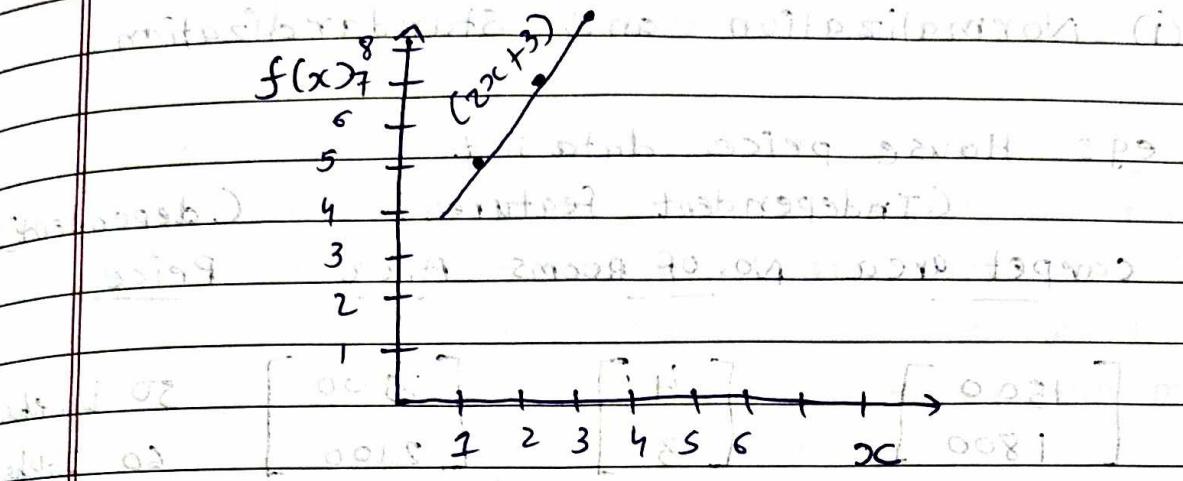
- Each person has exactly one partner

∴ No two people share same partner  $\rightarrow$  Injective

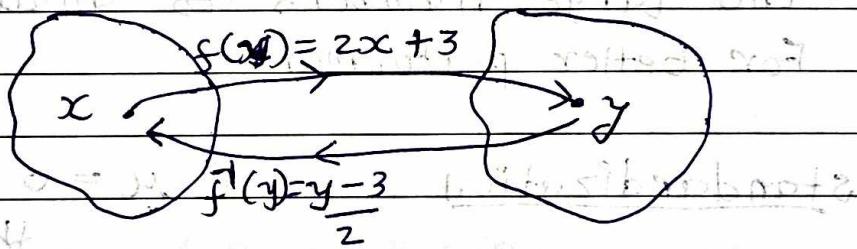
- Everyone has a partner (nobody is left

alone)  $\rightarrow$  surjective

Eg: Linear Functions  $f(x) = 2x + 3$   
 $x = [1, 2, 3]$



and graph of  $y = [5, 7, 9]$  is



find the Inverse

$$y = 2x + 3 \quad \text{for } x:$$

$$y - 3 = 2x \quad | :2$$

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

The Inverse Function

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

Verification:

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

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

## (8.2) Application of Inverse Function:

## (i) Normalization and Standardization:

eg: House price dataset.

(Independent features)

(Dependent)

carpet area	No. of rooms	Area	Price
-------------	--------------	------	-------

$\begin{bmatrix} 1500 \\ 1800 \end{bmatrix}$	$\begin{bmatrix} 4 \\ 3 \end{bmatrix}$	$\begin{bmatrix} 1800 \\ 2100 \end{bmatrix}$	50 Lakhs 60 Lakhs
--	--	--	----------------------

for the model training, we scaling down the large numbers to small numbers for better performance.

standardization

$$(M = 0 \text{ & } \sigma = 1)$$

Standard Normal distribution

No. of Rooms.

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

No. of Rooms.

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

$$(M = 2.5 \text{ & } \sigma = 2)$$

Original Transformation:  $z = \frac{x_i - \mu}{\sigma}$  (P)

Inverse Transformation:  $x_i = z\sigma + \mu$

→ use case: after deletion of outliers

After training a machine learning model on standardized data, the prediction after rescaled back to the original scale to interpret the result in meaningful way.

### Normalization

feature scaling with min max Normalization.

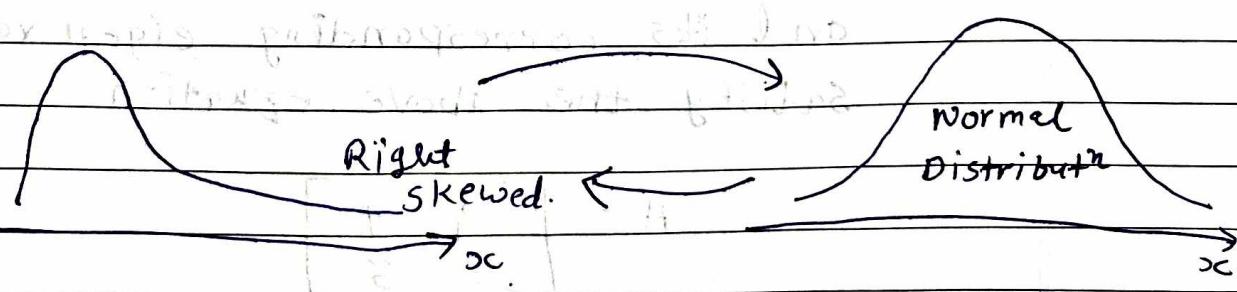
(Crosses the zero in original value)

Original transformation:  $z = \frac{x - \min(x)}{\max(x) - \min(x)}$

Inverse transformation:  $x = z(\max(x) - \min(x)) + \min(x)$

### (ii) Distribution of data:

#### - Logarithmic distribution.



Original Transformation  $y = \log(x)$

Invers Transformation  $x = e^y$

Visit ↴

(9) Eigen Value and Eigen Vector ([shad.io/MatVis](#))

→ Eigenvector of a matrix is a special vector that doesn't change direction when the matrix is applied to it. It only gets stretched or squished by some factor (called eigenvalue).

Eigen value ( $\lambda$ ): A scalar that indicates how much an eigen vector is stretched or compressed during Linear transformation.

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

$\because A = \text{square matrix}$

$\therefore \lambda = \text{Eigen value}$

$\therefore v = \text{Eigen vector}$

→ For a square metric  $A$ , an eigen vector and its corresponding eigen value  $\lambda$  satisfy the above equation

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

(i) Find Eigen values.

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

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

$$\begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix} = I_2 - A$$

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

$$\begin{aligned} \det(A - \lambda I) &= (4-\lambda)(3-\lambda) - (2)(1) \\ &= 12 - 4\lambda - 3\lambda + \lambda^2 - 2 \end{aligned}$$

$$\det(A - \lambda I) = \lambda^2 - 7\lambda + 10$$

Solve this equation.

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

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

$$\begin{bmatrix} 1 & 8 \\ 1 & 5 \end{bmatrix} \lambda = \frac{7 \pm \sqrt{49 - 40}}{2} = 7 \pm 3$$

$$\lambda_1 = \frac{7+3}{2}, \lambda_2 = \frac{7-3}{2}$$

$$\therefore \lambda_1 = 5, \lambda_2 = 2$$

This is Eigen value,  $\lambda_1 = 5$  or  $\lambda_2 = 2$

(ii) Find Eigen vector

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

For  $\lambda_1 = 5$ :  $\begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} - 5I = IA - A$

$$A - 5I = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} - \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$$

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

$$(D - 5I) - (\lambda - 5)(E - 5I) = 0$$

$$-x + y = 0$$

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

$$-x + y = 0$$

$$\therefore x = y$$

For  $\lambda_2 = 2$

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

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

$$\therefore y = -2x$$

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

For  $\lambda_1 = 5$ ,  $v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

For  $\lambda_2 = 2$ ,  $v_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$

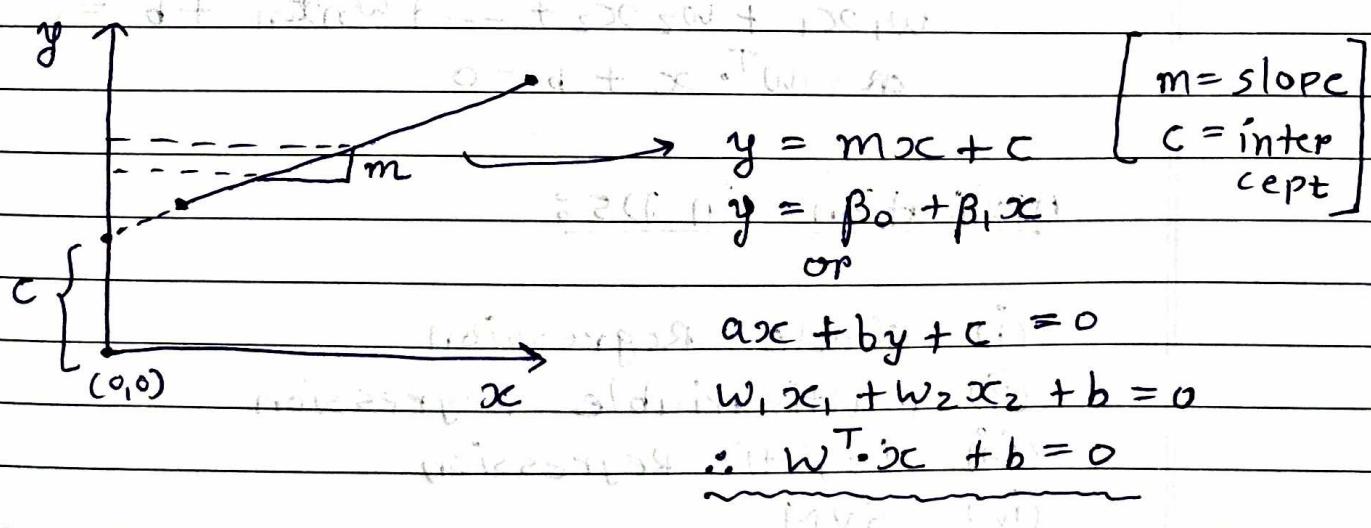
### Application :

(i) PCA (Dimensionality Reduction)

(ii) Latent Semantic Analysis (LSA) in NLP

### (10) Equation of line, plane and Hyperplane

- Line : (2D)



→ Linear Regression model prediction (salary based on Experience, in years) (in)

$$\text{Salary} = 5000 \cdot \text{Experience} + 20000$$

$$y = m \cdot x + c$$

- Plane (3D):

$$ax + by + cz = d$$

$$\text{or } w_1x_1 + w_2x_2 + w_3x_3 + b = 0$$

$$\therefore \underline{w^T \cdot x + b = 0} \quad S = \text{set}$$

$(a, b, c)$  is normal vector perpendicular to  
the plane

$d$  is a scalar.

(Gitterstruktur mit anschaulichem Aspekt)

e.g.: multiple linear regression (with 2 Features)

$$\text{Price} = 2000 \cdot \text{size} + 1500 \cdot \text{Rooms} + 10000$$

$$y = m_1 \cdot x_1 + m_2 \cdot x_2 + \dots + c$$

- Hyperplane ( $n$ -D):

$$w_1x_1 + w_2x_2 + \dots + w_nx_n + b = 0$$

$$\text{or } \underline{w^T \cdot x + b = 0}$$

Application in DS:

(i) simple Regression

(ii) Multivariable Regression

(iii) Logistic Regression

(iv) SVM

(v) PCA (Principal Component Analysis)

(vi) clustering (k-mean)

# Statistics

## Index :

### 1. Introduction to statistics

1.1 Introduction and types of statistics

1.2 Population and sample data

1.3 Types of Sampling

1.4 Types of data and scale of measurement

### 2. Descriptive statistics

2.1 Measure of central tendency

2.2 Measure of Dispersion

2.3 Percentile and Quartile.

2.4 5 Number summary

2.5 Histogram and skewness

2.6 Correlation and covariance

### 3. Inferential statistics and hypothesis testing

3.1 Hypothesis testing and its mechanism

3.2 P value and hypothesis testing

3.3 z test and hypothesis testing

3.4 t test and hypothesis testing

3.5 Type 1 & type 2 error

3.6 Baye's Theorem

3.7 Chi square test

3.8 Anova test

(1)

## Introduction to Statistics :

### (1.1) Introduction and Types of Statistics

- Statistics is the science of collecting, organizing and analyzing data.

#### Application :

- i) Data Exploration and Summarize
- ii) Model Building and validation.
- iii) Statistical Analysis
- iv) Hypothesis testing
- v) Optimization and efficiency.
- vi) Reporting.

- Types of statistics

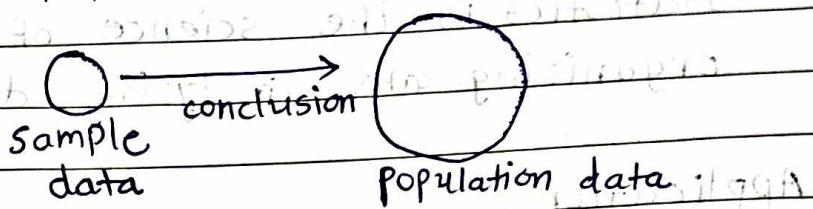
#### (i) Descriptive Statistics :

It involves methods for summarizing and organizing data to make it understandable.

- Measure of central tendency
- Measure of Dispersion
- Data distribution
  - i) Histogram
  - ii) Box plot
  - iii) Pie chart
- Summary statistics  
(Five number summary)

## (ii) Inferential Statistics.

It involves methods for making prediction about population based on sample data.



- Hypothesis testing

- P. Values

- Confidence interval

- Analysis test

- z test

- T-test

- Anova/F test

- chi square

### (1.2) Population and Sample data

#### (i) Population data:

A population is a entire set of individuals of interest in particular study.

characteristics

i) complete set : contain all the observations of interest

ii) Parameter : A numerical value summarizing the entire population

- Population mean ( $\mu$ )

- Population variance ( $\sigma^2$ )

Example : i) ~~Population in school study (a)~~

(i) Population in school study

- all students enrolled in a school.

- determine Avg. height of students.

(ii) Population in market study

- all consumers in a city

- To understand purchasing behavior

(iii) Sample data

- A sample is a subset of the population that is used to represent the entire group.

characteristics:

i) subset : Represent a portion of the population.

ii) statistic : A numerical value summarizing characteristics of the sample data.

[sample mean, sample variance]

iii) Random sampling : Sample should randomly

selected to avoid bias.

Example :

i) Sample in school study

- A group of 50 students from school

- Estimate the Avg. height of students in school.

### (1.3) Types of Sampling :

(1) Probability sampling

(2) Non probability sampling

→ (1) Probability sampling :

(a) simple random sampling :

- Every member has an equal chance of being selected.

eg: selecting people randomly.

(b) Systematic sampling :

- Select every  $n^{\text{th}}$  member from population.

eg: selecting every even number student

$2^{\text{nd}}, 4^{\text{th}}, 6^{\text{th}}$  etc.

(c) Stratified sampling :

- Divide the population into strata (group)

based on specific characteristics and

then randomly sampling from each strata

eg: Divide employees by department

then randomly select a specific

number of employee from each department.

(d) cluster sampling :

- Divide the population in clusters

than select random clusters,

Its mean all the members of selected clusters are sample.

## → (2) Non probability sampling: (प्र.)

### (a) Convenience sampling:

- selecting individual who are easier to reach.

eg: surveying people at mall.

### (b) Judgmental (purposive) sampling:

- Select individual based on researcher's purpose.

eg: choosing experts in data science.

### (c) Snowball Sampling:

- Existing study subjects recruit Future subject From their network.

eg: Internal employee select candidate From their network for job position.

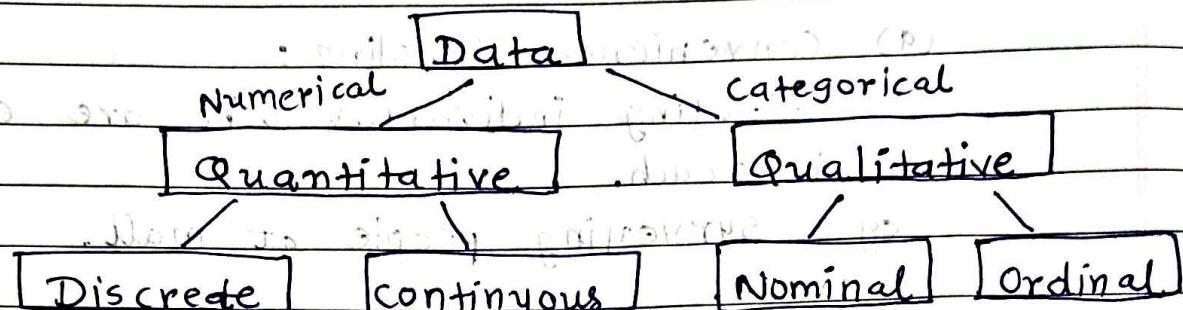
### (d) Quota sampling:

- Divide population into categories like (age, gender, income etc) and then Select a fixed number (quota) of subjects From each category.

eg: total 100 seats for vacancy

and 20 seats quota for disabled.

## (1.4) Types of Data and Scale of measurement



- Whole number Any number + No Rank Rank.
- eg: No. of eg: weight eg: gender eg: costume
  - Bank Acc height M, F Feedback
  - NO. OF Tempredure Blood grp. Good 1  
children speed. Moderate 2 Bad 3
  - Better 2

### Scale of measurement of data.

(i) Nominal scale

(ii) Ordinal scale

(iii) Interval scale

(iv) Ratio scale

- (i) Nominal scale

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

- Qualitative / categorical data.

### Characteristics

- i) Data is categorized based on labels, names
- ii) categories are mutually exclusive
- iii) No logical order among categories (Rank)

eg: Types of cuisines

- Italian

- Chinese

- Mexican

(ii) ordinal scale

- This scale classifies data into categories that can be ranked or ordered.

characteristics

- i) Data is categorized and ranked in specific order

- ii) The interval between ranks are not necessarily equal.

eg: education level, customer feedback

High school      0

Bachelor	1	Poor	0
Master	2	good	1
Doctorate	3	Excellent	2

(iii) Interval scale

- Numerical data with meaningful differences.

- No true (zero) value.

eg: Temperature in celsius

- $20^\circ\text{C}$  is  $10^\circ$  more than  $10^\circ\text{C}$ .

- $0^\circ\text{C}$  is not mean no temperature

→ you can add/subtract, but ratios don't make sense.

(eg:  $20^\circ\text{C}$  is not "twice as hot" as  $10^\circ\text{C}$ )

## (iv) Ratio scale :

- Like interval, but have true zero value
- Allow all maths operation like add, subtract, multiply, divide.

eg: Age, height, weight, income, distance

→ you can say 20kg is twice heavy as 10kg.

(2) Descriptive Statistics :(2.1) Measurements of central tendency:

- measures of central tendency are statistical metrics that describe the center point or typical value of a dataset. They provide a single value that summarize a set of data by identifying the central position within that dataset.

(i) mean :

- mean is the sum of all values divide by the number of values.

Population mean ( $\mu$ )      Sample mean ( $\bar{x}$ )

population (N)

sample (n)

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

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

$N = 5$

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

- characteristics
  - Affected by extreme outliers.
  - used for interval and ratio data

### (ii) median :

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

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

$$\text{Median} = 3$$

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

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

- characteristics

- Not affected by extreme outliers.
- used for ordinal, interval & Ratio data.

### (iii) mode :

- The most frequent value in a dataset.

eg : - Dataset : {2, 4, 4, 6, 7, 7, 7, 9}

$$\therefore \text{Mode} = 7$$

$$- \text{Dataset} : \{3, 5, 5, 6, 6, 8\}$$

$$\therefore \text{Mode} = 5, 6 \text{ (Bimodal)}$$

- characteristics

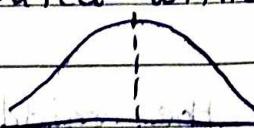
- Not affected by extreme outlier.

- used for Nominal, ordinal, interval & Ratio.

- choosing the Appropriate Measure

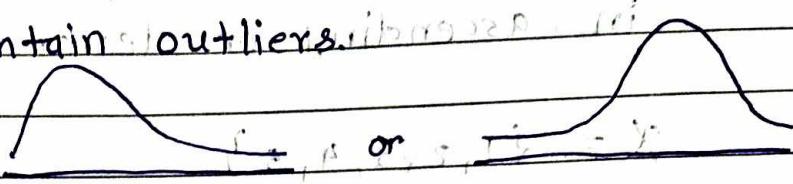
positions 3003659 p. 1-3105474

i) mean : Best use when data is symmetric distributed without outliers.



+ minimum (ii)

bar (symmetric) ii) median : Best use when data is skewed or contains outliers. (ignore outliers)



iii) mode : Best use for categorical data to identify the most common category.

- Real world Application. (EDA & Feature eng.)

Age	weight	Salary	Gender	Degree
-----	--------	--------	--------	--------

24	70	40K	M	BE
----	----	-----	---	----

25	180	100K	F	-
----	-----	------	---	---

27	95	45K	F	-
----	----	-----	---	---

24	-	50K	M	PHD
----	---	-----	---	-----

32	90K	60K	F	BE
----	-----	-----	---	----

-	60	-	-	-
---	----	---	---	---

-	65	55K	F	BSC.
---	----	-----	---	------

40	72	-	M	BE
----	----	---	---	----

→ There are some missing values in dataset, we use mean or median for age and weight, mode for gender & degree.

## (2.2) Measure of Dispersion:

- Measure of dispersion describe the spread or variability of a dataset. They indicate how much the values in a dataset is differ from the central tendency.

- common measures of dispersion.

(i) Range

(ii) Variance

(iii) Standard deviation.

(iv) Interquartile Range.

- (i) Range

- Range is a difference between the maximum and minimum value in dataset.

- Range = max. value - min. value

e.g.: Ages = { 14, 13, 10, 20, 25, 75, 15 }

$$\text{Range} = 75 - 10 = 65$$

- characteristics:

- simple to calculate

- sensitive to outliers

- Rough measure of dispersion.

- (ii) Variance

- variance measures the average squared deviation of each value from the mean.

## Population variance

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

where all observations  $N$  are included

$x_i \Rightarrow$  Data points of  $x_i \Rightarrow$  Data point  
 $\mu \Rightarrow$  population mean  $\bar{x} \Rightarrow$  sample mean  
 $N \Rightarrow$  population size  $n \Rightarrow$  sample size

e.g.: size of flower petals

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

$$N = 5$$

$$\mu = 5 + 8 + 12 + 15 + 20 = 12$$

$$5$$

$$\sigma^2 = (5-12)^2 + (8-12)^2 + (12-12)^2 + (15-12)^2 + (20-12)^2$$

$$= 35 + 16 + 0 + 9 + 64 = 124$$

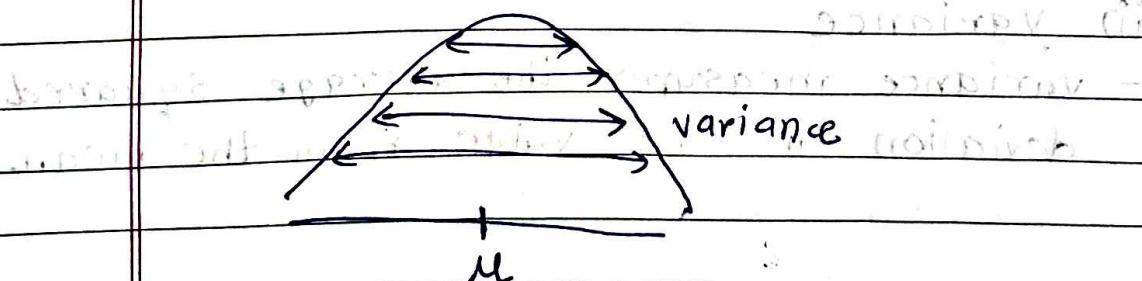
$$\sigma^2 = 27.6$$

$$\{31, 37, 28, 25, 31, 36, 34\} = 29.8$$

$$\sigma^2 = 31 - 29.8 = 1.6$$

- characteristics.

- provide a precise measure of variability
- units are squared off the original data unit
- more sensitive to outliers than the Range



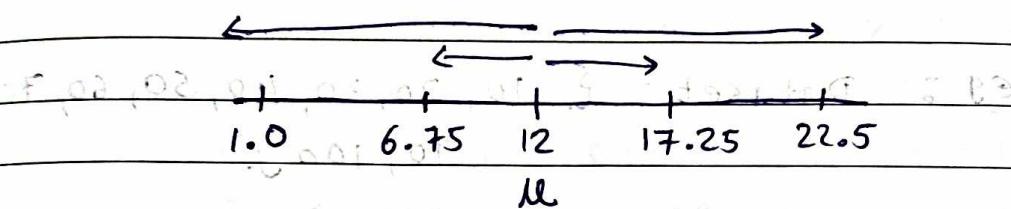
- (iii) standard Deviation:

The standard deviation is the square root of the variance.

$$\sigma^2 = \sqrt{27.6}$$

$$\therefore \sigma = 5.25$$

(Ans) 5.25



$$(2 + 18) \times 9 = 180$$

- characteristics

provide a clear measure of spread in the same unit as the data.

- Sensitive to outliers

Big Question ?

Why sample variance is divided by  $N-1$ .

### (2.3) Percentile and Quartile.

$$P_{10} = 11 \times 2 + 0 = 11 \times 2 = 22$$

(i) percentile :

- A percentile tells you the relative standing of a value in a dataset.

e.g.: in a class of 100 students, if you scored at 90<sup>th</sup> posi percentile, it means you scored better than 90 students.

(ii) Quartile : definition (ii) -

- quartile is a specific percentile that divides your data into four equal parts.

 $Q_1$  $1^{st}$  quartile ( $25\text{-P}$ ) $Q_2$  $2^{nd}$  quartile ( $50\text{-P}$ ) $Q_3$  $3^{rd}$  quartile ( $75\text{-P}$ )

e.g. : Dataset  $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$

$$\text{Rank} = \frac{P}{100} \times (n+1)$$

∴ Range  $\therefore \text{Rank} = \frac{25}{100} \times (10+1) = 0.25 \times 11$

∴ 25% part of 100 items sample  $\therefore 2.75^{\text{th}}$  position

$$\therefore Q_1 = 30$$

{ no item  $\therefore$  pick

$$\therefore Q_2 = \frac{50}{100} \times 11 = 0.5 \times 11 = 5.5$$

$$\therefore Q_2 = 55$$

{ item 5th item from starting (E.S.)

$$Q_3 = \frac{75}{100} \times 11 = 0.75 \times 11 = 8.25$$

∴  $Q_3 = 80$  { item 8th item from starting  
∴ item 8th item is 80th item

∴  $Q_1 = 30, Q_2 = 55, Q_3 = 80$  { item 8th item is 80th item

{ item 5th item from starting (E.S.)

∴  $Q_1 = 30, Q_2 = 55, Q_3 = 80$  { item 8th item is 80th item

{ item 8th item is 80th item

## (2.4) Five number summary.

- Five number summary is a quick overview of a dataset's distribution.
- components of 5 number summary :
  1. minimum - the smallest data value.
  2.  $Q_1$  (First Quartile) - 25<sup>th</sup> percentile
  3. Median ( $Q_2$ ) - 50<sup>th</sup> percentile
  4.  $Q_3$  (Third Quartile) - 75<sup>th</sup> percentile
  5. maximum - the largest data value.

eg:  $\{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27\}$

still,  $N = 19$ . dt 70 observations (dt 21 + C)

$$\text{Lower fence} = Q_1 - 1.5 \times (\text{IQR})$$

$$\text{Higher Fence} = Q_3 + 1.5 \times (\text{IQR})$$

$$\therefore Q_1 = \text{Percentile } \times (n+1) = 25 \times (20) = 5^{\text{th}} \text{ position}$$

$$\therefore Q_3 = \frac{75}{100} \times 20 = 15^{\text{th}} \text{ position} = 7$$

$$\text{IQR} = Q_3 - Q_1 = 7 - 3 = 4$$

$$\text{Lower Fence} = Q_1 - 1.5 \times (\text{IQR})$$

$$= 3 - 1.5(4)$$

$$= 3 - 6 = -3$$

$$\text{Higher Fence} = Q_3 + 1.5 \times (\text{IQR})$$

$$= 7 + 1.5(4)$$

$$= 13$$

→ 27 will be removed because it is higher than Higher Fence.

$$\text{minimum} = 7$$

$$1^{\text{st}} \text{ Quartile} = 3$$

$$\text{Median} = 5$$

$$3^{\text{rd}} \text{ Quartile} = 7.5 - \text{minimum}$$

$$\text{Maximum} = 19 + (7.5 - 7) = 20$$

$$19 + (7.5 - 7) = 20$$

### C.2.5) Histogram and Skewness

- Histogram is a graphical representation of the distributions of numerical data.

It is an estimate of the probability distribution of a continuous variable and is used to visualize the shape, central tendency and variability of dataset.

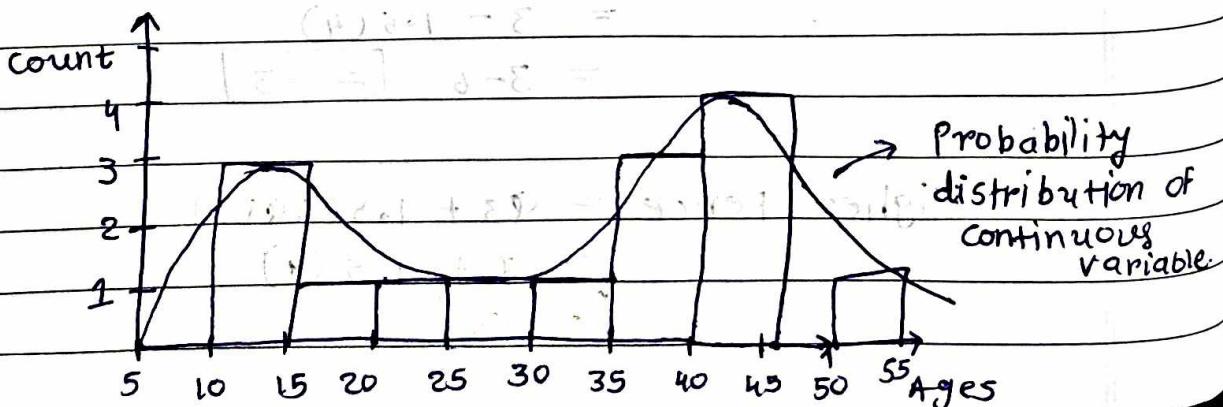
$$\text{Ages} = \{11, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50\}$$

$$\rightarrow \text{No. of bins} = 10$$

$$\rightarrow \text{Range} = [0-50]$$

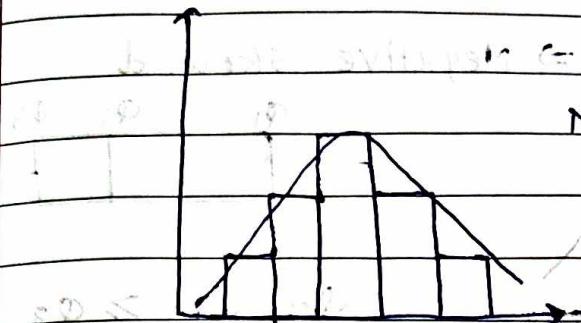
$$\rightarrow \text{bin size} = \frac{50}{10} = 5$$

$$\text{Bins} \rightarrow [0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35, 35-40, 40-45, 45-50]$$



## - Skewness

bimodal -

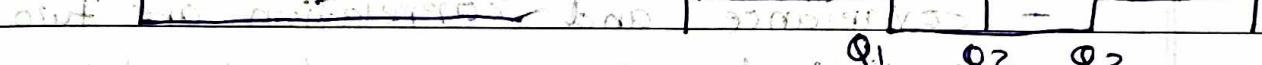
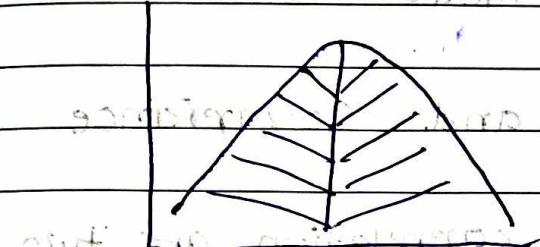


Normal / Gaussian distribution

symmetrical distribution  
(No skewness)

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

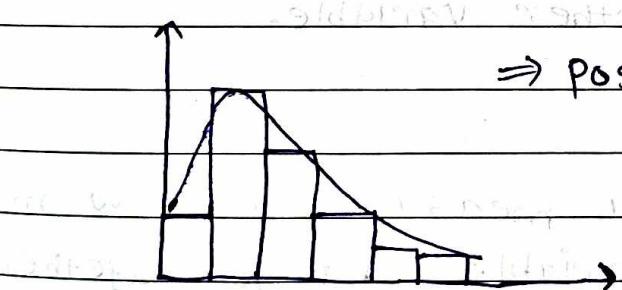
The mean, median & mode  
are perfectly at the center



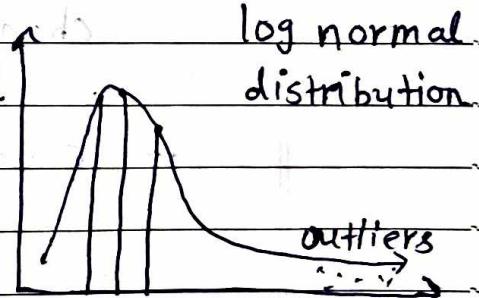
Mean = Median = Mode and  $Q_3 - Q_1 \approx Q_2 - Q_1$

mean = mode = median =  $Q_1 + Q_3 / 2$

## - Right skewed



$\Rightarrow$  positive skewed

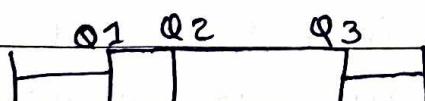


log normal  
distribution

outliers

mean  $>$  median  $>$  mode

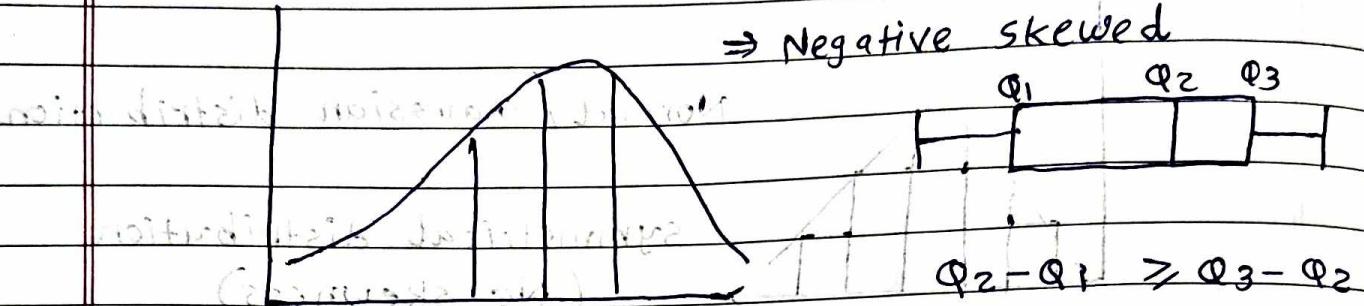
Box plot:  $Q_3 - Q_1 \geq Q_2 - Q_1$



$Q_3 - Q_1 \geq Q_2 - Q_1$

- Left skewed.

$\Rightarrow$  Negative skewed



mean < median < mode

### (2.6) Correlation and Covariance

- covariance and 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 changes in another variable.

#### i. covariance:

covariance is a measure of how much two random variables change together.

If the variable tend to increase and decrease together, the covariance is positive.

If one tends to increase and other decreases then the covariance is negative.

size of house  $\downarrow$  Price  $\uparrow$

1200

$x \uparrow y \uparrow$

1300

$x \downarrow y \uparrow$

1500

$x \uparrow y \downarrow$

45 lakhs

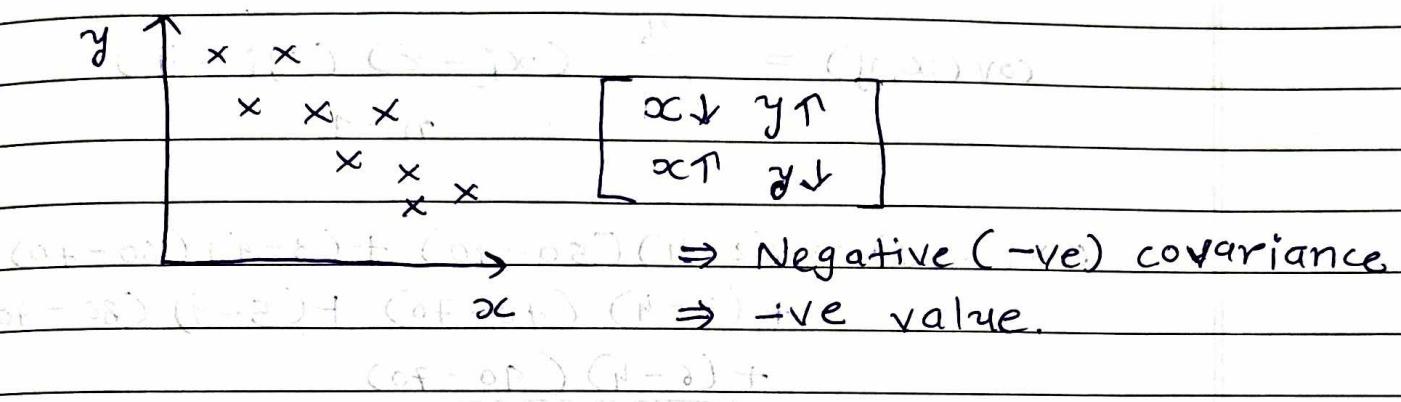
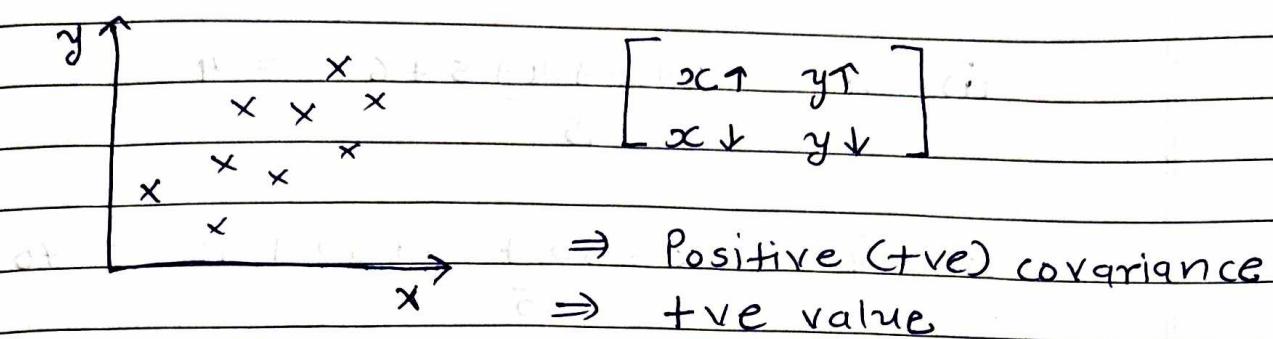
$x \downarrow y \uparrow$

50 lakhs

$x \uparrow y \downarrow$

75 lakhs

$x \downarrow y \downarrow$



Formula:

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

To find cov add const of it. if sample size is increasing

$x_i$  → datapoint of random variable  $x$

$\bar{x}$  → sample mean of  $n$  no's

$y_i$  → Data points of random variable  $y$

$\bar{y}$  → Sample mean of  $y$

Eg: Students

Hour of studied (x)      Exam score (y)

2	50
3	60
4	70
5	80
6	90

$x \uparrow y \uparrow$

$x \downarrow y \downarrow$

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

$$(ii) \bar{y} = \frac{50+60+70+80+90}{5} = 70$$

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

$$\begin{aligned} \text{cov}(x, y) &= (2-4)(50-70) + (3-4)(60-70) \\ &\quad + (4-4)(70-70) + (5-4)(80-70) \\ &\quad + (6-4)(90-70) \\ &= 40 + (-20) + 0 + 20 + 40 \\ &= 20 \end{aligned}$$

$$\therefore \text{cov}(x, y) = 20 \quad \Rightarrow \text{Cov}(x, y)$$

positive covariance indicates the no. of hours studied increased the exam score also.

- Advantage: easier to interpret
- Quantify the relationship between  $x$  &  $y$ .
- Disadvantage: (a) tribute to unit
- covariance does not have a specific unit value

$$\text{cov}(x, y) \Rightarrow -\infty \text{ to } +\infty$$

- correlation :

→ Pearson correlation coefficient  
 → Spearman Rank correlation

(i) Pearson correlation coefficient.  $\Rightarrow [-1 \text{ to } 1]$

- Linear relationship between two continuous variable

- It checks how well a straight line fits the data.

Formula:  $r = \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y}$

$$r = \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y}$$

Standard deviation of  $x$  &  $y$

- The more the value toward +1 the more +ve correlated  $(x, y)$ .
- The more the value toward -1 the more -ve correlated it is  $(x, y)$ .
- IF  $r = 1$  then all variables perfectly fit to straight line.

(ii) Spearman Rank correlation

- Good when data is not linear or ordinal
- use rank instead of raw values.

Formula:  $P = \frac{\text{cov}(\text{rank}(x), \text{rank}(y))}{\sigma(\text{rank}(x)) \cdot \sigma(\text{rank}(y))}$

eg:  $x = [1, 2, 5, 6, 0], y = [2, 3, 4, 1, 7]$

$\text{Rank}(x) = [1, 2, 3, 4, 5], \text{Rank}(y) = [1, 2, 3, 4, 5]$

$\text{Cov} = \frac{1}{5} \sum_{i=1}^5 (\text{Rank}(x)_i - \bar{x})(\text{Rank}(y)_i - \bar{y}) = 1.2$

$\sigma(\text{rank}(x)) = \sqrt{\frac{1}{5} \sum_{i=1}^5 (\text{Rank}(x)_i - \bar{x})^2} = 1.58$

$\sigma(\text{rank}(y)) = \sqrt{\frac{1}{5} \sum_{i=1}^5 (\text{Rank}(y)_i - \bar{y})^2} = 1.58$

$P = \frac{1.2}{1.58 \cdot 1.58} = 0.47$

### - Real world Application in datascience.

#### (i) Feature selection.

- Remove irrelevant features (correlation)

#### (ii) Linear regression.

- Detecting multicollinearity

- If two features are highly co-related  
then remove one feature or combine  
them using PCA. (correlation)

#### (iii) PCA.

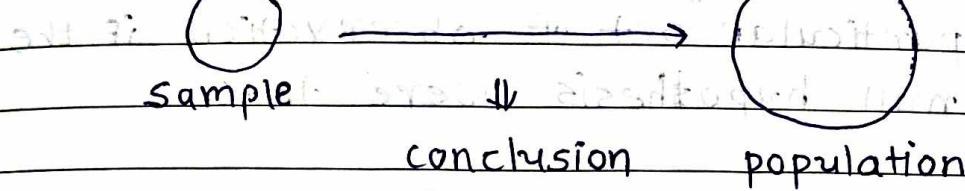
- reduce the number of features  
while retaining most of the variance  
in the data. It relies on the  
covariance matrix.

### (3) Inferential Statistics and hypothesis testing.

#### (3.1) Hypothesis testing and its mechanism

hypothesis testing mechanism

Sample result at sample level → Conclusion



##### - Hypothesis testing mechanism :

###### (1) Null hypothesis ( $H_0$ )

- person is not guilty

- The assumption you are beginning with

###### (2) Alternate hypothesis ( $H_1$ )

- The person is guilty

- opposite of null hypothesis

###### (3) Experiment → Statistical Analysis

Different type of test.

→ Accept or Reject the hypothesis

###### (4) Accept the null hypothesis or reject the null hypothesis

rejecting  $H_0$ , i.e.,

sample statistic

sample statistic

sample statistic

accepting  $H_0$

$10.0 > 9$

$8.0 > 9$

$8.0 < 9$

### (3.2) P value and hypothesis testing

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

#### Simple example:

A company claims their battery lasts 10 hours on average. You test 30 batteries and find the average life is 9.5 hours.

Question ?

Is this difference (10 vs 9.5) due to random chance or is claim not true?

$H_0$  (Null hypothesis):  $\mu = 10$  hours.

$H_1$  (Alt Hypothesis):  $\mu \neq 10$  hours.

You calculate P-value = 0.03

P-value

$P < 0.01$

$0.01 < P < 0.05$

$P \geq 0.05$

Interpretation

Strong evidence (Reject  $H_0$ )

Moderate evidence (Reject  $H_0$ )

Weak evidence (Accept  $H_0$ )

In our example  $\alpha = 0.03$

therefore Reject  $H_0$

conclusion: Average battery life is likely not 10 hours.

### - Example:

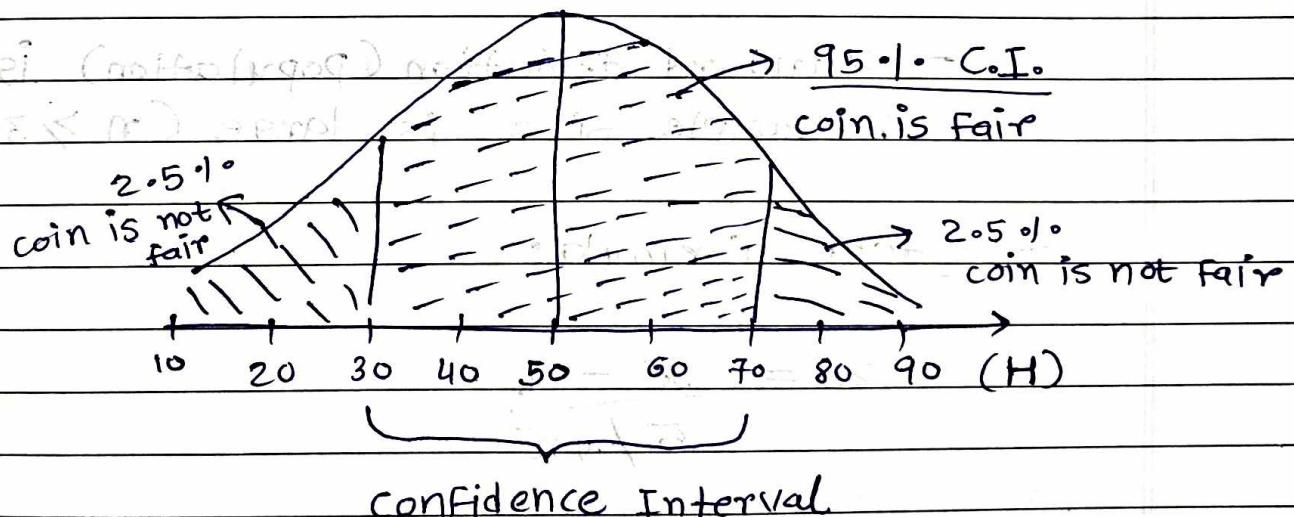
coin is fair or not, we want to test.

if we toss 100 times

(i) Null Hypothesis ( $H_0$ ) = coin is fair

(ii) Alternative hypothesis ( $H_1$ ) = coin is not fair

(iii) Experiment: 100 times



(iv) Significance value:  $\alpha = 0.05$

$$\text{C.I.} = 1 - 0.05$$

$$\therefore \text{C.I.} = 0.95$$

(iv) conclusion:  $P < \alpha$  (significance value) then

Reject the Null hypothesis

also called to reject null hyp.

### (3.3) z test and hypothesis testing.

off from continuous

- (i) z test } Average  $\Rightarrow$  z table
- (ii) t test } Sample size  $\leq 30$   $\Rightarrow$  t table
- (iii) chi square  $\Rightarrow$  Categorical data
- (iv) Anova  $\Rightarrow$  variance.

$\Rightarrow$  significant

#### - z test:

z test is a statistical test used to determine whether there is a significant difference between two sample means.

- A sample mean and population mean
- or two sample means.

- Standard deviation (population) is known & sample size is large ( $n \geq 30$ ).

#### - z test Formula:

$$(H_1) \text{ of } Z_3 = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$\bar{x}$  = sample mean

$\mu$  = population mean

$\sigma$  = population standard deviation

$n$  = sample size.

Problem: company claim that their energy drink increase alertness and people stay focused for 8 hours on average. You think it's less. You test it with a sample of 36 peoples and get:

- Sample mean = 7.5 hours.

- population std = 1.2

- population mean = 8 hours.

- sample size = 36

You want to check if people actually stay focused less than 8 hours. (left tailed test)

$\rightarrow$  (i) Null hypothesis ( $H_0$ ) =  $\mu = 8$

(ii) Alt. hypothesis ( $H_1$ ) =  $\mu \neq 8 < 8$

(iii) calculate z score.

APRIL 2020

$$Z = \frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}} = \frac{7.5 - 8}{1.2 / \sqrt{36}} = \frac{-0.5}{0.2}$$

$$\therefore Z = -2.5$$

(iv) Now find p value from z table.

(look at -2.5 & 0.00 value in z table)

16.3

16.5

$$\therefore p(Z < -2.5) = 0.0062$$

$$\therefore p \text{ value} = 0.0062$$

(v) compare p value with significance level ( $\alpha$ )

let take  $\alpha = 0.05$

$$\therefore 0.0062 < 0.05 \Rightarrow \text{We Reject } H_0$$

( $p < \alpha$ )  $\Rightarrow$  People stay focused less than 8 hr.

→ Two types of hypothesis tests

(i) Left tailed or Right tailed test (one tailed)

- we test, if the mean is less than or greater than the population mean.
- use only one side of distribution
- No need to multiply p value by 2
- Left tailed : use  $P(Z < z)$
- Right tailed : use  $1 - P(Z < z)$

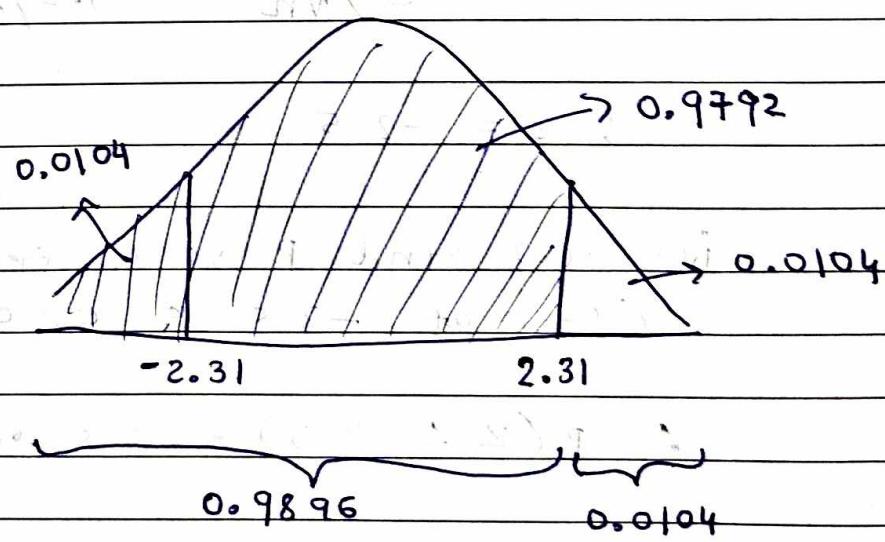
(ii) Two tailed test

- we test, if the mean is simply different
- you have to check both side of the curve
- So you multiply by 2 to get full p value.

Suppose  $\therefore z = 2.31$ .

& z table shows 0.9896

$$\therefore P(Z < 2.31) = 0.9896$$



1

$$1 - 0.9896 = 0.0104$$

area is  $0.0104$  area of Right side (tail)  
 & also  $0.0104$  area of left tail.  
 ∴ both is  $2 \times 0.0104$

$$P = \underline{0.0208}$$

∴  $P = 0.0208$  -

→ two method use for conclusion in z test.

- (i) use confidence Interval (C.I.) & find critical z value and compare with original z value.

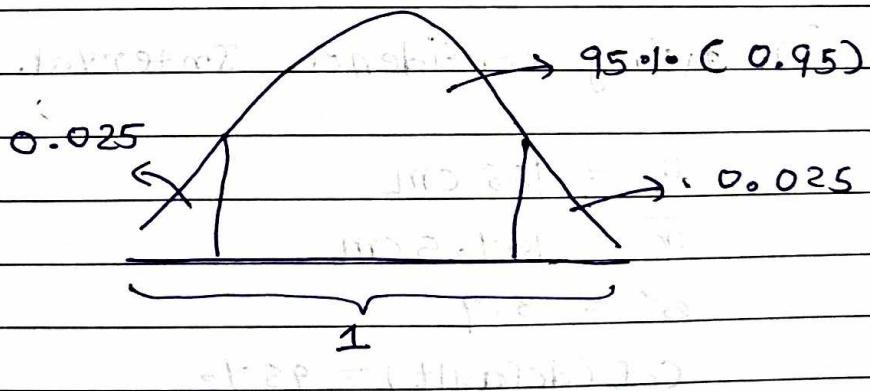
In many case default C.I. = 95%.

which  $\Rightarrow \alpha = 0.05$  out of 0.05

∴ two tailed test : critical  $z = \pm 1.96$

∴ one tailed test : critical  $z = + 1.645$ .

→ how can i find z using C.I.



$$1 - 0.025 = 0.9750 \Rightarrow \frac{\text{Now check}}{\text{Area}} \quad z\text{-table}$$

∴  $0.9750$  is not in table  $+ 1.96$

hence proved.

(ii) calculate the P-value and compare

with  $\alpha$ .

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

- Find p value using z table.

if p value  $< 0.05 \Rightarrow$  Reject null hypothesis

if p value  $\geq 0.05 \Rightarrow$  do not reject null.

- problem : how many S units?

A average heights of all residents in a city is 168 cm with standard deviation  $\sigma = 3.9$

A doctor believes the mean to be different.

he measure the height of 36 individuals

and found the average height to be 169.5 cm

Is doctor's belief real or not?

we solve using both methods :-

(i) using confidence Interval.

$$\mu = 168 \text{ cm}$$

$$\bar{x} = 169.5 \text{ cm}$$

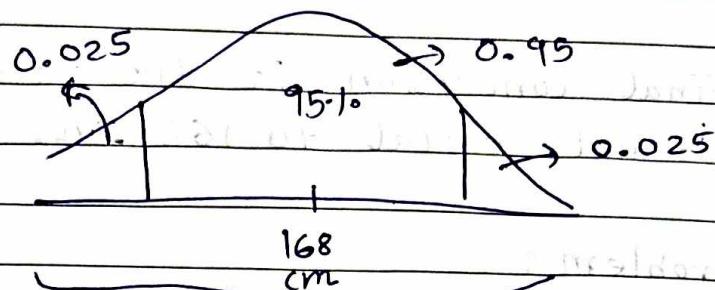
$$\sigma = 3.9$$

$$\text{C.I (default)} = 95\%$$

$$\alpha = 0.05 \quad (\alpha = 1 - \text{C.I.} = 1 - 0.95 = 0.05)$$

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

Alt hypothesis  $H_1 = \mu \neq 168 \text{ cm}$ .



$$\therefore 20 \cdot 68 - 1 - 0.9025 = 0.9750$$

∴ z score for 0.9750 in z-table is  $+1.96$   
with given critical z score  $\neq \pm 1.96$

Now calculate z score for given data.

Given  $\bar{x} = 169.5$ ,  $\mu = 168$ ,  $\sigma = 3.9$

$$\therefore z = \frac{\bar{x} - \mu}{\sigma} = \frac{169.5 - 168}{3.9} = \frac{1.5}{3.9}$$

$$= \frac{1.5}{3.9} = \frac{1}{3} = \frac{1}{3} = 0.3333$$

$$z = \frac{\bar{x} - \mu}{\sigma} = \frac{1.5}{0.65} = 2.31$$

Conclusion :

$z > \text{critical } z$  → test →

$$\therefore 2.31 > 1.96$$

∴ Reject the null hypothesis

(ii) using P value method :

$$P(z < 2.31) = 0.98956 \quad (\text{Right tail})$$

$$\text{Now, } 1 - 0.98956 = 0.01044$$

(Right tail area).

$$\therefore \text{Both tail} = 2 \times 0.01044$$

$$\text{Both tail} = 0.02088$$

$\rightarrow 0.02088 < 0.05 \Rightarrow \therefore \text{Reject Null Hypothesis (H}_0)$ .

$\therefore$  Final conclusion is the average height is not equal to 168 cm.

- Problem:

A factory manufactures bulbs with avg. warranty of 5 years with std of 0.50. A worker believes that the bulb will defected in less than 5 years. he tests 40 bulbs and find the average time of bulb is 4.8 years.

- (i) state null & alt hypothesis
- (ii) at a 2.1.0. significant level, is there enough evidence to support the idea that the warranty should be revised?

$\rightarrow$  Null hypothesis ( $H_0$ ) =  $\mu = 5$

Alt hypothesis ( $H_1$ ) =  $\mu < 5$  (left tailed)

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

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

$$\therefore z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \text{putting values (i)} \\ \therefore z = \frac{4.8 - 5}{0.5/\sqrt{40}}$$

$$\therefore z = (-2.53)$$

now look up in Z table

prob. of getting  $z < -2.53$  is 0.0002

prob. of getting  $z < -2.53$  is 0.0002

prob. of getting  $z < -2.53$  is 0.0002

which is very small

$$z \text{ score} = (-2.53)$$

see z table and get p-value (ii)

$$P\text{-value} = 0.00570$$

(iii) H.A

$$\text{since here } \alpha = 2.1\% \text{ & } \beta = 0.025\% \text{ (ii)}$$

∴ compare p-value with  $\alpha$  and  $\beta$

$$\text{option } f \neq 0.00570 < 0.025\% \text{ (iv)}$$

∴ we reject the null hypothesis

Conclusion: company should not withdraw to get

company should Revise their

warranty timeline.

### (3.4) t test and hypothesis testing

as the population standard deviation is unknown

~~t~~ test uses when:

- sample size is small ( $n \leq 30$ )

- population standard deviation ( $\sigma$ ) is unknown

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

$\bar{x}$  = sample mean

$\mu$  = population mean

$s$  = sample standard deviation

$n$  = sample size

## Steps For performing t-test :

(i) state hypotheses

null ( $H_0$ )

Alt ( $H_1$ )

(ii) calculate t-score using formula

(iii) find critical t-value from t-table.

(based on degrees of freedom :  $df = n - 1$ )

(iv) compare your t-score with critical value or compute the P-value.

### Problem:

In a population, the average IQ is 100.

A team of researchers to test a new meditation to see if it has either positive or negative effect on IQ, or no effect at all.

A sample of 30 participants who have taken the meditation has a mean of 140 IQ with a standard deviation of 20.

Did the meditation effect intelligence?

(or > 10) Null hypothesis =

$\rightarrow \mu = 100, n = 30, \bar{x} = 140, s = 20,$

CI = 95.1%,  $\alpha = 0.05$

(i) null hypothesis ( $H_0$ ) =  $\mu = 100$

Alt hypothesis ( $H_1$ ) =  $\mu \neq 100$  (2 tail test)

(ii)  $\alpha = 0.05$

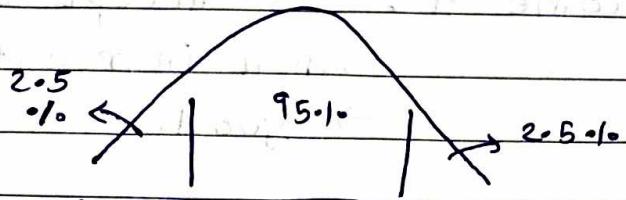
(iii) Degree of freedom =  $n - 1$

$$\begin{aligned} df &= 30 - 1 \\ &= 29. \end{aligned}$$

i → Now use t-table,

critical dof = 29 &  $\alpha = 0.05$  (two tail test)

∴ critical t-score =  $\pm 2.045$



If t-score is less than  $-2.045$  or greater than  $+2.045$  than reject the null hypothesis.

(iv) calculate t-score

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

Conclusion: (i) strong association

$t >$  critical t ( $10.96 > 2.045$ )

∴ we Reject the null hypothesis

∴ IQ is increasing with meditation.

### (3.5) Type 1 & Type 2 Error

- 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.

C  
30

- outcome 3 : we accept the null hypothesis when in reality it is false  
→ Type 2 error.

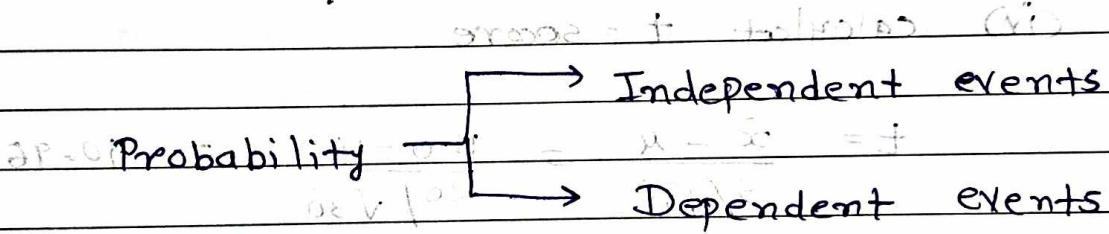
- outcome 4 : we accept the null hypothesis when in reality it is true  
→ good.

Note: we will discuss more about in  
aspects to SPSS-confusion matrix topic. (Stay tuned)

More info to follow from experts to math

probability

### (3.6) Baye's Theorem :



(i) Independent Events (ii) Dependent Events

eg: Rolling dice

toss an coin

$$\text{p}(1) = \frac{1}{6}, \text{ p}(2) = \frac{1}{6}$$

one box have two

Red & 3 white balls

- we picking ball one

etc... & repeat & by one.

$$\text{P(Red)} = \frac{2}{5}$$

Now

$$\text{p(Yellow)} = \frac{3}{4}$$

Next event depends on  
Previous events.

$$\therefore P(\text{Red and white}) = P(R) * P(W/R)$$

$$\therefore P(A \text{ and } B) = P(B \text{ and } A)$$

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

$$\therefore P(B/A) = \frac{P(B) * P(A/B)}{P(A)}$$

Baye's theorem

$$\therefore P(A/B) = \frac{P(A) * P(B/A)}{P(B)}$$

→ It is used in Naïve Baye's algorithm  
 we will deep dive in that algorithm  
 in ML part.

### (3.7) CHI Square test

- The chi-square test is used to check if there is significant relationship between two categorical values or
- If the observed data fits the expected data.
- It is used for ordinal & nominal data.

Example: Chi-square goodness of fit.

In a science class of 75 students, 11 are left handed. Does this class fit the theory that 12.1% of people are left handed?

$$(81\%) + (8)9 = (12.1\%)9 * (A) :$$

	Observation	Expectation
Left hand	11	9
Right hand	$\frac{64}{75}$	$\frac{66}{75}$

- chi-square test formula:

$$\chi^2 = \sum \frac{(Observat^n - Expect^n)^2}{Expectation}$$

Hence, standard value of  $\chi^2$  is 3.89

- Solved Example : book line 81  
from pg 100

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

Weight of the individuals	$< 50$ kg	$50-75$	$> 75$
percentage	20.10	30.10	50.10

(as per book for strengthen result)

In 2020, weights of 500 individuals were sampled, below is result.

	$< 50$ kg	$50-75$	$> 75$
frequency	140	160	200

using  $\alpha = 0.05$ , would you conclude the population difference of weight has changed in the last 10 years?

- Ans :-

Expected	< 50	50 - 75	> 75
	100	150	250

Observed.	< 50	50 - 75	> 75
	140	160	200

- (1) Null hypothesis ( $H_0$ ) : The data meets the null hypothesis or it fulfills the expectation.

Alt. ( $H_1$ ) : The data doesn't meet expectation.

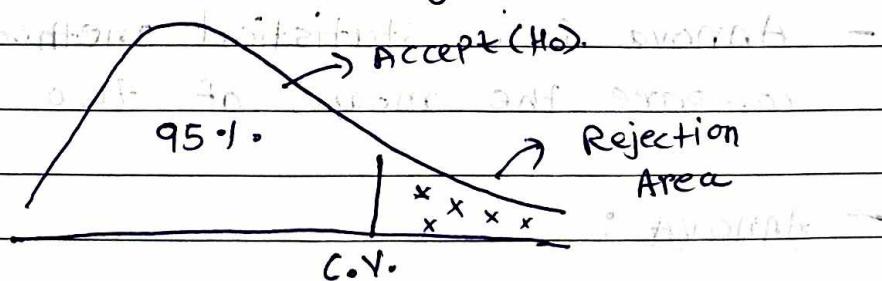
(2)  $\alpha = 0.05$ , CI = 95%

(3) Degree of freedom

$$df = (K-1) \times (Categories - 1)$$

$$\therefore df = 3-2 = 2$$

(4) Decision Boundary



check in chi-square table

with  $df = 2$  &  $\alpha = 0.05$

$$\therefore C.V. = @. 5.991$$

$\therefore$  IF  $\chi^2 > 5.99$ , Reject  $H_0$ .

else failed to Reject  $H_0$ .

(5) calculate chi-square test score

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

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

$$= \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250}$$

$$\chi^2 = 26.66.$$

2010	2020	2020	bayrsado
OBS	OBS	OBS	OBS
2FC	2F-02	022	011

conclusion :

$$\chi^2 > 5.99$$

since  $\chi^2 > 5.99$  we Reject ( $H_0$ )

The weight of 2020 population are different than those expected in the 2010 population.

### (3.8) ANNOVA Test :-

(Analysis of variance) :

- Anova is a statistical method used to compare the means of two or more groups.

- Anova :

(1) Factors (variable)

(2) Levels

e.g.: Medicine (factor)

[Dosage] → 5 mg, 10 mg, 15 mg → Levels

## mode of payment (factor)

[Levels] → Gpay | phonePay | IMPS | NEFT

### Assumption of ANOVA

#### - Assumption in ANOVA

##### (i) Normality of sampling distribution of mean

- The data in each group should be approximately normally distributed.

##### (ii) Absence of outliers

- Outlying score need to be removed from the dataset

##### (iii) Homogeneity of variance

- All groups should have equal variance (spread of data). This is called homoscedasticity

$$[ \sigma_1^2 = \sigma_2^2 = \sigma_3^2 ]$$

##### (iv) Independence of samples

- samples must be independent of each other (no overlap in group)

#### - Types of ANOVA :

- There are main three types of ANOVA depending on the number of factors and group involved.

##### (i) One way ANOVA

##### (ii) Two way ANOVA

##### (iii) Repeated measures ANOVA.

## (i) one way ANOVA:

- use to compare means of 3 or more groups (Levels) for 1 independent variable (factor).

eg: compare weight loss across

3 different diets.

weight loss (Independent variable or factor)

keto      Vegan      Protein

5.1      7.1      6.2

6.0      3.5      7.0

→ Now, compare mean

between 3 groups

## (ii) Two way ANOVA:

- use when we have 2 independent variables (factors), and want to check:-

- effect of each group

- interaction effect between them.

eg: Does diet type & exercise type affect weight loss?

factor (Diet)

Exercise (factor)



keto      vegan      Protein

cardio      Strength

(level 1)    (level 2)    (level 3)

(level 1)    (level 2)

Diet type	Exercise type	weight loss(Kg)
keto	cardio	5.0
keto	Strength	6.1
vegan	cardio	4.0
Vegan	strength	4.8
Protein	cardio	6.3
protein	Strength	6.9

### (iii) Repeated Measures ANOVA:

→ Measure the same group or individual multiple times over times or under different condition.

e.g.: does a weight loss program change weight over time?

Time	weight(kg)
Before	80
After 1 months	78
After 2 months	75

⇒ solved Example of ANOVA Test :  
(One way ANOVA).

A Doctor want to test a new medication which reduce headache. They split the participant into 3 condition (15 mg, 30mg , 45 mg). Later on the doctor ask the patient to rate the headache between [1-10]. Are there any difference between the three conditions,  $\alpha = 0.052$

Ans :

	15 mg	30 mg	45 mg
1.2	6.9	7.14	
1.3	8.0	6.93	
0.1	7.1	6.92	
8.2	8.1	7.03	
2.3	8.0	8.14	
3.2	9.2	7.03	
4.2	8	6	2

: ANOVA 2-way random block design

(i)  $H_0 = \mu_{15} = \mu_{30} = \mu_{45}$

H<sub>1</sub> : not all  $\mu_i$  are equal or any  
relation to assert one or different.

within + between

(ii) Significant  $\alpha = 0.05$ , CI = 0.95

spreads morphology triticeal &amp; rye SPS

(iii) calculate degrees of freedom

(N = 21), n = 7, a = 3

df<sub>between</sub> = a - 1 = 3 - 1 = 2

df<sub>within</sub> = N - a = 21 - 3 = 18

df<sub>total</sub> = N - 1 = 20

: test which is known level

Now, df<sub>1</sub>, df<sub>2</sub> (2, 18) &  $\alpha = 0.05$ 

check F table and find

critical value (C.V.)

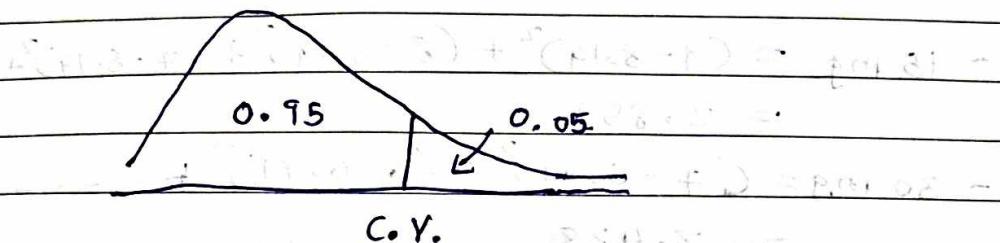
to obtain C.V. of F table

normal (0.05, 0.01, 0.001) relation is

should not show any relationship with

block and with each other.  $F = 2.7$  means

there is no significant difference and it is same

(iv) Decision Boundary,  $H_0$  to use Given

if  $F$  is greater than 3.5547, then  
reject the null hypothesis.

## (v) Group mean &amp; overall mean

$$\bar{x}_1 (15 \text{ mg}) = (9+7+8+8+8+9+8)/7 = 8.14$$

$$\bar{x}_2 (30 \text{ mg}) = (7+6+6+7+8+7+6)/7 = 6.71$$

$$\bar{x}_3 (45 \text{ mg}) = (4+3+2+3+4+3+2)/7 = 3$$

$$\bar{x} (\text{overall}) = \frac{57 + 47 + 21}{21} = \frac{125}{21} = 5.95$$

## (vi) sum of square between group (SSB).

$$SSB = \sum n_i (\bar{x}_i - \bar{x})^2$$

$$- 15 \text{ mg} = 7 \times (8.14 - 5.95)^2 = 7 \times 4.82 = 33.74$$

$$- 30 \text{ mg} = 7 \times (6.71 - 5.95)^2 = 7 \times 0.58 = 4.06$$

$$- 45 \text{ mg} = 7 \times (3 - 5.95)^2 = 7 \times 8.70 = 60.89$$

$$SSB = 33.74 + 4.06 + 60.89$$

$$SSB = 98.69$$

## (vii) sum of square within groups (SSW)

$$\begin{aligned}
 - 15 \text{ mg} &= (9-8.14)^2 + (8-8.14)^2 + (7-8.14)^2 + \dots \\
 &= 2.853 \\
 - 30 \text{ mg} &= (7-6.71)^2 + (6-6.71)^2 + \dots \\
 &= 3.428 \\
 - 45 \text{ mg} &= (4-3)^2 + (3-3)^2 + (2-3)^2 + \dots \\
 &= 4
 \end{aligned}$$

$$\therefore SSW = 2.853 + 3.428 + 4 = 10.281$$

## (viii) Mean Squares.

$$MSB = SSB / df_{FB} = 98.69 / 2 = 49.345$$

$$MSW = SSW / df_{FW} = 10.281 / 18 = 0.571$$

## (ix) F-statistics.

$$F = \frac{MSB}{MSW} = \frac{49.345}{0.571} = 86.42$$

Conclusion:  $F = 86.42 > 3.55$

$\therefore$  we, Reject the null hypothesis ( $H_0$ )

All medicines effects are different.

Index :

1. Introduction to probability
  - 1.1 Addition and Multiplication rule
2. Probability distribution Function.
  - 2.1 Probability Mass function (PMF)
  - 2.2 cumulative distribution Function (CDF)
  - 2.3 Probability density Function (PDF)
3. Types of Probability distribution.
  - 3.1 Bernoulli distribution
  - 3.2 Binomial distribution
  - 3.3 Poisson distribution
  - 3.4 Normal Gaussian distribution
  - 3.5 Standard Normal distribution
  - 3.6 Uniform distribution
  - 3.7 Log-Normal distribution
  - 3.8 Power Law distribution
  - 3.9 Pareto distribution.
  - 3.10 Central Limit theorem
  - 3.11 Estimates.

(1) Introduction to Probability :

## (1.1) Addition and Multiplication Rule

(i) Intro

(ii) Addition rule (for mutually exclusive event)

(iii) Addition rule (for non mutual exc. event)

(iv) Multiplication rule (Independent &amp; Dependent events)

(i) Probability: It is about determining the likelihood of an event.

eg.: toss a coin  $\{H, T\}$  (Algorithm (v))

$$P(H) = \frac{1}{2} = 50\%$$

$$P(T) = \frac{1}{2} = 50\%$$

(ii) Mutually Exclusive Events

- Two events are mutually exclusive if they cannot occur at the same time

eg.: tossing a coin (Head & tail cannot both come at same time).

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

$$= \frac{1}{2} + \frac{1}{2} = 1 \quad \text{mutually exclusive}$$

- Rolling a dice:  $\{1, 2, 3, 4, 5, 6\}$

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

$$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

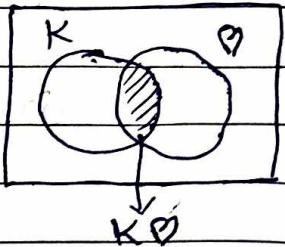
## (iii) Non-mutually Exclusive events.

eg: Taking a card from the deck.

$$P(K \text{ or } Q) = P(K) + P(Q) - P(K \text{ and } Q)$$

$$\therefore P(K \text{ or } Q) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$

$$\therefore P(K \text{ or } Q) = \frac{16}{52} = \frac{4}{13}$$



## (iv) Multiplication Rule (Independent and Dependent events)

- Two events are Independent, if they do not affect one another.

eg: Tossing a coin {H or T}

Rolling a dice {1, 2, 3, 4, 5, 6}

- Two events are dependent, if they affect each other

eg: Take a king from the deck and then take a queen from the deck.

$$P(K) = \frac{4}{52}, P(Q) = \frac{4}{51}$$

( $\leftarrow$  multiplication rule : ~~utilizing ages~~)

(1) Independent Event : {Tossing coin}

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

$$= \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$$

a card containing 3 red cards and 4 blue cards  
what is the probability of drawing a red card  
and then a blue card?

(2) Dependent Event:

$$(x) \rightarrow \text{conditional prob.}$$

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

for example if  $K = X$  with a different set.

$$= \frac{4}{52} * \frac{3}{51}$$

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

(2) Probability Distribution Function:

- Age = {5, 7, 9, 12, 11, 25, ...}  $\Rightarrow$  continuous

$\rightarrow$  Random variable



(i) Probability mass function (PMF): used for discrete random variable.

(ii) Probability density function (PDF): used for continuous random variable.

(iii) cumulative distribution function (CDF): used for both discrete & continuous rand. var.

## (2.1)(i) Probability Mass Function : (PMF) [Discrete Random variable]

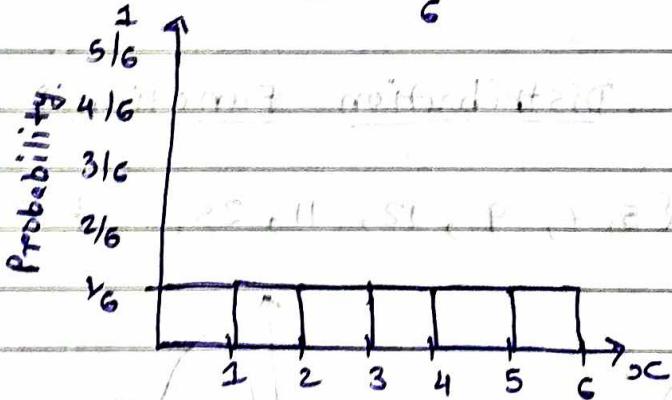
- Discrete random variable takes countable value like  $\{0, 1, 2, 3, \dots\}$
- PMF gives the probability that a discrete random variable equal to a specific value.

$$P(X = x) = \text{PMF}(x).$$

e.g Rolling a dice  $X = \{1, 2, 3, 4, 5, 6\}$

$$P(X = 1) = \frac{1}{6}$$

$$P(X = 5) = \frac{1}{6} \text{ etc. ---}$$



## (2.2)(i) cumulative Distribution Function : (CDF)

- used for both discrete and continuous random variable
- CDF gives the probability that a variable is less than or equal to a value  $x$ .

$$F(x) = P(X \leq x)$$

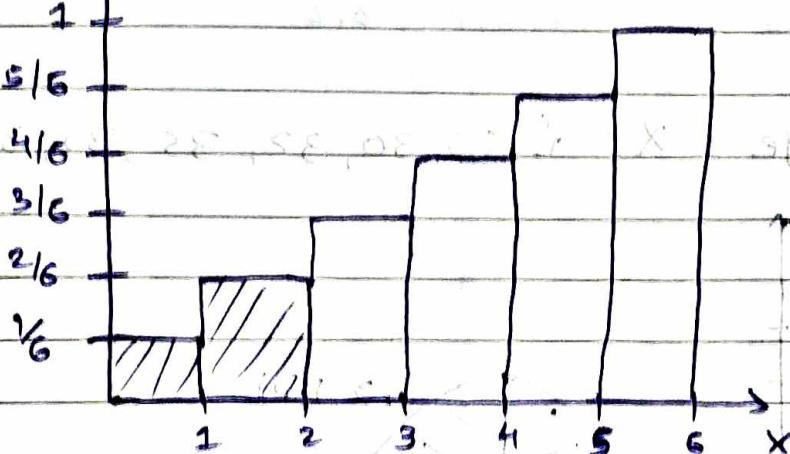
Ex: Rolling a dice.

$$\text{P}(X \leq 3) = P(1) + P(2) + P(3)$$

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$

$$\text{P}(X \leq 3) = \frac{3}{6} = 0.5$$

cumulative probability



$$\text{P}(X \leq 2) = P(1) + P(2)$$

$$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

### (2.3) (iii) Probability Density Function : (PDF)

- used for continuous random variables.
- PDF is a function such that the area under the curve between two values gives the probability that the variable lies in the range.

$$\text{P}(a \leq X \leq b) = \int_a^b f(x) dx$$

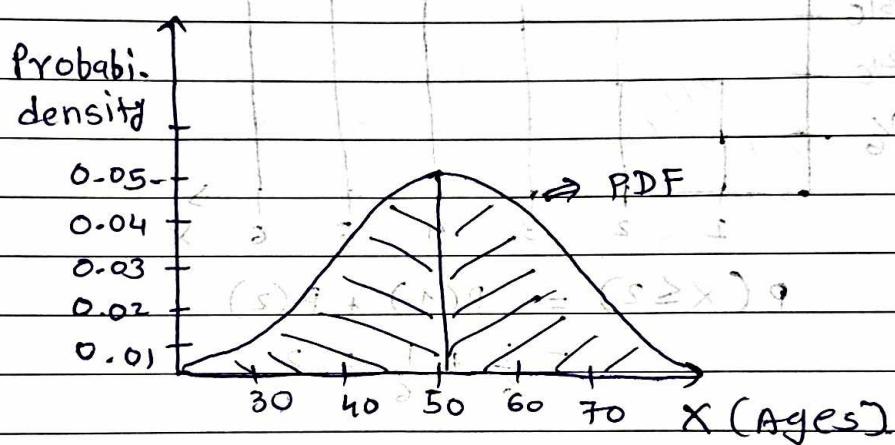
e.g.: Let  $X$  be a continuous variable

(i) with uniform distribution between 0 & 1.

$$f(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$P(0.2 \leq X \leq 0.5) = \int_{0.2}^{0.5} 1 dx = 0.5 - 0.2 = 0.3$$

- Age =  $X = \{25, 30, 32, 35, 38, 40, 45, \dots\}$



$$(Q19) : \text{Find } P(X \leq 50) = 0.5 = 50\% \quad (\text{Ans})$$

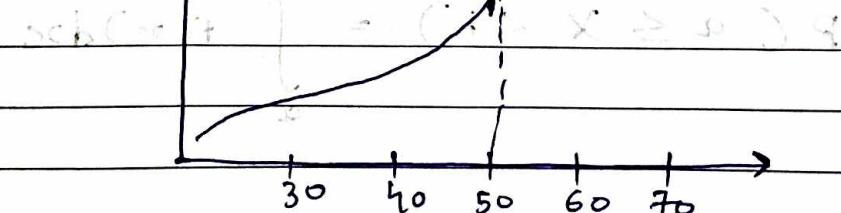
↳ cumulative CDF graph: probability that there -

↳ cumulative distribution function or CDF -  
gives prob. of getting less than or equal to x

↳ if both distributions are similar  
then their CDF's will also be similar

↳ here we have to find  $P(X \geq 50)$

↳  $P(X \geq 50) = 1 - P(X \leq 50)$



- PDF Properties:

(i) Non negativity  $f(x) \geq 0$  for all  $x$ .

(ii) The total area under the PDF curve is equal to 1.

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

### (3) Types of Probability distribution.

There are many types of probability distribution:

(i) Bernoulli distribution

(ii) Binomial distribution

(iii) Poisson distribution

(iv) Normal gaussian distribution

(v) Standard Normal distribution

(vi) Uniform distribution

(vii) Log normal distribution

(viii) Power Law distribution

(ix) Pareto distribution

(x) Central limit theorem

(xi) Estimates.

$$(9-1) \cdot 89 = 801$$

$$801 - 89 = 712$$

$$9 < \frac{1}{(9-1)} \cdot 89 + (1-1)^2$$

$$89 - 89 = 0$$

$$9 < \frac{1}{(9-1)} \cdot 89 + (1-1)^2$$

## (3.1) Bernoulli distribution :

The Bernoulli 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 questions.

- Tossing a coin → Head (1), Tail (0).
- Student passes (1) or fails (0) an exam.
- Customer buys (1) or doesn't buy (0).
- If a random variable  $X$  follows Bernoulli distribution, it means:  $X \sim \text{Bernoulli}(p)$ .
- $p$  = Probability of success ( $X=1$ )
- $(1-p)$  = probability of failure ( $X=0$ )
- Probability Mass Function (PMF):

$$P(X=x) = \begin{cases} p & \text{if } x=1 \\ 1-p & \text{if } x=0 \end{cases}$$

$$\text{PMF} = p^k \cdot (1-p)^{1-k}$$

if  $k=1$  (success)

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

if  $k=0$  (failure)

$$P(k=0) = p^0 \cdot (1-p)^1 \Rightarrow (1-p) \Rightarrow q$$

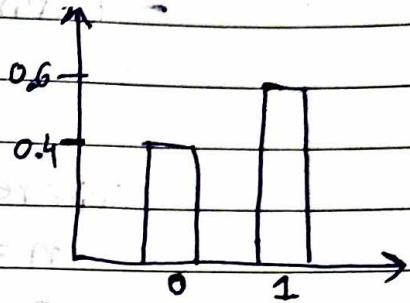
- Mean of Bernoulli distribution :-

$$E(X) = \sum_{k=0}^1 k \cdot P(k) \rightarrow k = \{0, 1\}$$

$$= (0) \cdot (0.40) + (1) \cdot (0.60)$$

$$= 0 + 0.60$$

$$E(X) = 0.60 = p$$



$$\text{and } q = 1 - p = 1 - 0.60 = 0.40$$

Want to find variance =  $\sigma^2$

- Variance :

$$P(K=0) = 0.4 = q$$

$$P(K=1) = 0.6 = p$$

$$\sigma^2 = 0.4 * (0 - 0.6)^2 + 0.6 * (1 - 0.6)^2$$

$$\sigma^2 = 0.4 * 0.36 + 0.6 * 0.16$$

$$\sigma^2 = 0.24 (p * q)$$

$$\therefore \text{variance } \sigma^2 = p \cdot q \text{ (or) } p \cdot (1-p)$$

### (3.2) Binomial Distribution :

- The binomial distribution is used to model the number of successes in a fixed number of independent experiments (called trials), where each trial has only two outcomes:

- success(1)

- failure(0)

- eg:
- flipping a coin 10 times (H or T)
  - Sending 100 emails (opened or not)

- Formula for binomial distribution:

$$X \sim \text{Binomial}(n, p)$$

where,

-  $n$  = number of trials.

-  $p$  = probability of success in one trial

$X$  = number of success in  $n$  trial

- PMF:

$$P(X = k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$$

$$P(k, n, p) = {}^n C_k \cdot p^k \cdot (1-p)^{n-k}$$

where, : written in terms of binomial distribution (X)

-  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$  is binomial coefficient

-  $k$  = number of success ( $0 \leq k \leq n$ )

eg: what is probability of getting exactly 3 heads in 5 coin tosses?

-  $n = 5$

-  $p = 0.5$

-  $k = 3$

$$\therefore P(X=3) = \binom{5}{3} (0.5)^3 \cdot (0.5)^2 \\ = 10 \times 0.125 \times 0.25 \\ P(X=3) = 0.3125$$

so, there are 31.25% chance of getting exactly 3 heads in 5 tosses!

- Mean :  $E(X) = n \cdot p$
- Variance :  $\text{Var}(X) = n \cdot p \cdot q$
- Std :  $\sqrt{n \cdot p \cdot q}$

### (3.3) Poisson Distribution :

- The poisson distribution is used to model the number of times an event happens in a fixed interval of time, space, distance, when events happen randomly and independently at constant average rate.

- Poisson answers the question:  
"How likely is it that an event happens k times in fixed period, given the average number of times it usually happens?"

- Eg :
- Calls at a call center per hour
  - Emails received per day
  - Road accidents per month in a city
  - Typos per page in book.

- If random variable  $X$  follows a poisson distribution, we write.

$$X \sim \text{poisson}(\lambda)$$

where,

- where  $\lambda = \text{avg. number of events per interval}$   
 $X = \text{actual number of events in that interval}$

- Probability Mass Function (PMF):

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

• Applications: modeling (ex. 2)

- $k = \text{no. of events}$

•  $e = 2.718$  (natural base)

•  $\lambda = \text{avg. rate (mean no. of events)}$ .

• Eg. A call center receives 5 calls per hour

• on average. To find probability

• what is the probability, it receives exactly

• 3 calls in one hour?

• Let, we have  $\lambda = 5$  and  $k = 3$

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

$$P(X=3) = \frac{5^3 \cdot e^{-5}}{3!} = \frac{125 \cdot e^{-5}}{6} \approx 0.1404$$

• So, there is a 14.04% chance of getting exactly 3 calls in an hour.

• So, there is a 14.04% chance of getting exactly 3 calls in an hour.

## (3.4) Normal gaussian distribution :

- Normal distribution is a bell shaped curve that represents how data is distributed around the mean (avg).

Real life example :

- Height of people. (most people are avg. height)
- Exam score. (most student score around the avg)

Shape:

- symmetrical around the mean
- has a peak at the mean
- Tails of equally on both sides.

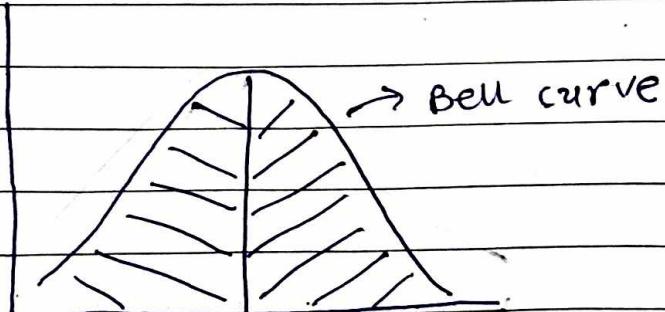
Formula of Normal distribution :

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

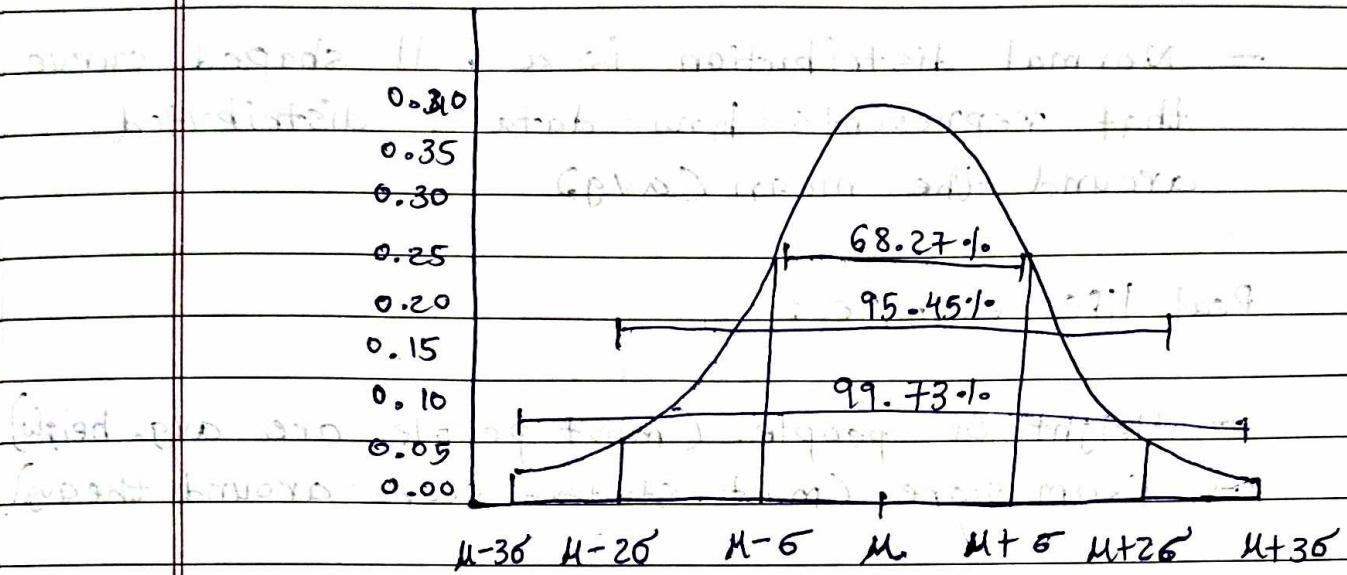
where,  $x$  = value

$\mu$  = mean

$\sigma$  = standard deviation



## Empirical rule of Normal distribution.

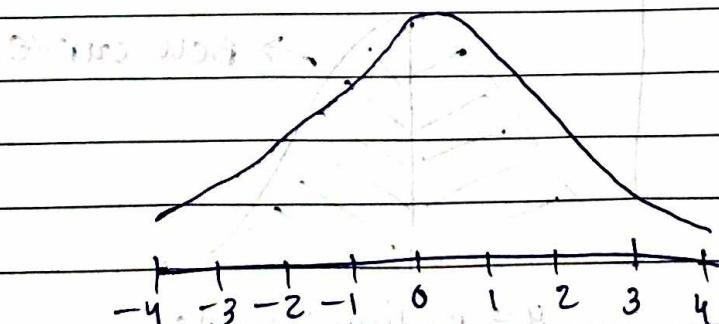


- 68.27% of data lies within  $1\sigma$  of the mean
- 95.45% within  $2\sigma$
- 99.73% within  $3\sigma$

### (3.5) Standard Normal Distribution

- Standard normal distribution is just a special case of the normal distribution where:
- mean  $\mu = 0$
- standard deviation  $\sigma = 1$

It also bell-shaped, but centered at 0.



- Why we use standard Normal?

To compare different datasets or values, we standardize them using the z-score formula:

$$z = \frac{x - \mu}{\sigma}$$

↳ Computing z-score is to convert data to standard deviation.

↳ In this  $x$  = your actual data point

$\mu$  = mean of the data

$\sigma$  = Standard deviation.

eg: Exam scores are normally distributed.

mean = 70, SD = 10

You scored 85, what is your Z-score?

$$z = \frac{85 - 70}{10} = 1.5$$

Your score is 1.5 standard deviation above mean.

### (3.6) Uniform Distribution.

- Uniform distribution is a type of probability distribution where all outcomes are equally likely.

eg: - Tossing a fair dice

- Picking a card randomly

- choosing random number between 1 and 10.

## - Types of uniform distribution-

### (i) Discrete uniform distribution

- Finite set of outcomes

- eg: Rolling a die  $\rightarrow \{1, 2, 3, 4, 5, 6\}$

### (ii) continuous uniform distribution

- Infinite number of outcomes within range

eg: Picking real numbers between 0 & 1

### (i) Discrete uniform distribution

- every one of  $n$  value has equal probability  $\frac{1}{n}$ .

If a random variable  $X$  can take  $n$  value equally likely, then:

$$P(X = x_i) = \frac{1}{n}, \text{ for all } x_i$$

$f(x)$

$\frac{1}{n}$

equally likely probability

(a, b)

(PMF)

(CDF)

with this we get -

discrete form of probability

continuous random variable

of this p

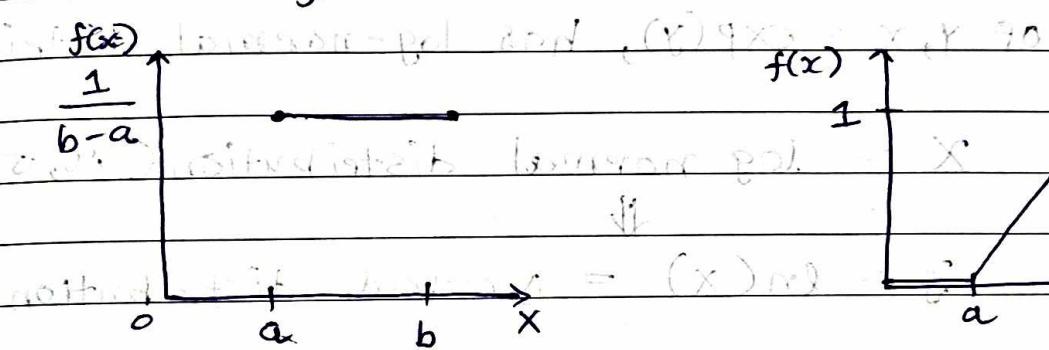
## (ii) Continuous uniform distribution

IF  $X$  is a continuous random variable between  $a$  and  $b$ , then the probability density function (PDF) is:

$$\text{PDF} : f(x) = \frac{1}{b-a}, \text{ for } a \leq x \leq b$$

mean & std. dev. of  $(X)$  are  $\mu = \frac{a+b}{2}$  &  $\sigma^2 = \frac{(b-a)^2}{12}$

$$\text{and, } P(a \leq X \leq b) = 1$$



(PDF)  $f(x)$  (CDF)  $F(x)$

eg: choose random number between 0 & 10.  
what is prob. picking number between 2 & 6?

$$a=0, b=10$$

$$\text{PDF} : f(x) = \frac{1}{10-0} = 0.1$$

$$P(2 \leq X \leq 6) = (6-2) \times 0.1 = 0.4$$

$\therefore$  So, there are 40% chance the number lies between 2 & 6.

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

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

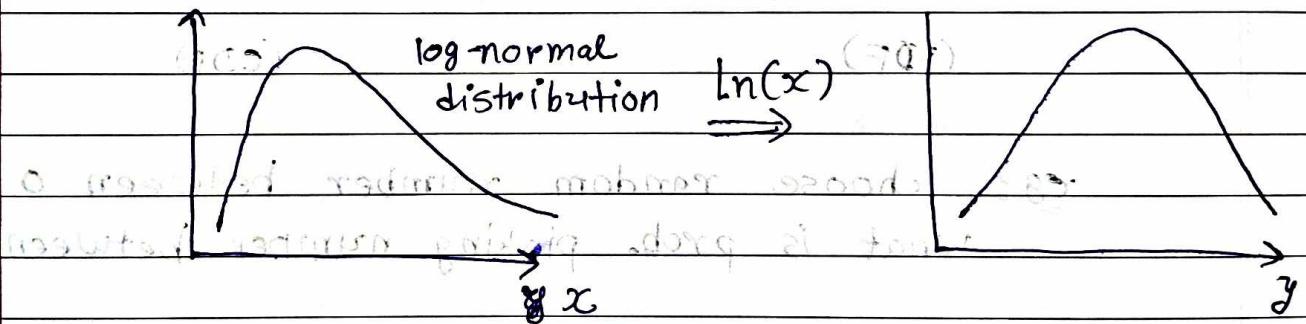
## (3.7) Log normal distribution

- A Log-normal distribution is a probability distribution of a random variable whose logarithm is normally distributed
- If the 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 log-normal distribution.

$x = \text{log normal distribution}(\bar{\mu}, \sigma^2)$



$y = \ln(x) = \text{Normal distribution}$



$$\text{OP} = d, \text{ OAS}$$

- eg:
- Wealth distribution of the world.
  - length of comments (only few long)
  - Salary of employees in a company.

$$P(X = x) = e^{-x} (e^{-x})^x = (e^{-x+1})^x$$

Volume of sounds about 50dB

$$(d+b)^{\frac{1}{2}} = \text{mean} -$$

### (3.8) Power law distribution.

- Power law is a mathematical relationship where small occurrences are extremely common and large instances are extremely rare, but still more likely than in normal distribution.

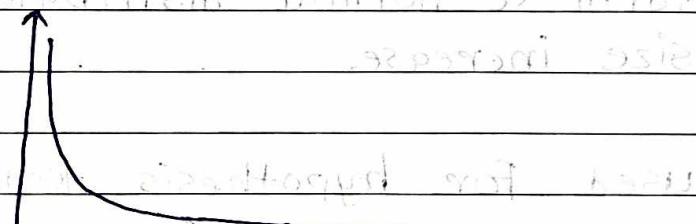
$$P(x) \propto x^{-\alpha} \quad (\text{for } x > 1)$$

~~calculus~~  $x = \text{Value of interest}$

~~calculus~~  $\alpha = \text{exponent} (\text{also called: pareto index})$

~~calculus~~  $\propto = \text{proportional to}$

~~calculus~~  $\rightarrow \text{tail is very long}$



### (3.9) Pareto distribution.

- Pareto distribution is a special case of power law, used to describe situations where a small number of things account for a large portion of the effects.

- 80/20 Rule (Pareto principle).

- 80% result outcome from 20% of causes.

### - Real life examples:

- 20% peoples have 80% of wealth
- 80% of peoples living in 20% cities
- 80% of entire project is done by 20% of the team.

### (3.10) Central Limit Theorem

- If you take many random samples from any population (no matter its shape), the average (mean) of those samples will form a normal distribution as the sample size increase.
- used for hypothesis testing, confidence interval etc, even if the original data is not normal

#### Real life example:

- Let's say brushing time of 1000 people =
- you take random groups of 30 peoples
  - calculate the average brushing time for each group.
  - Repeat this 1000 times.
  - The distribution of these average will look like bell curve.

$$\text{standard Error} = \frac{\sigma}{\sqrt{n}}$$

where: ~~estimate of the population mean~~

~~estimate~~  $\sigma$  = population standard deviation

$n$  = sample size; ~~number of observations~~

As  $n$  increase, standard error gets smaller, meaning sample means get closer to the true population mean.

- eg:
- you have population data of people's monthly spending (not normal distribution)
  - population mean  $\mu = 10,000 \pm$
  - standard deviation  $\sigma = 2,000 \pm$

Now,

- Take 50 samples ( $n=50$ ), each of 30 peoples.
- Find mean spending of each sample.
- plot the 50 means.
- Distribution will be normal distribution.

$\Rightarrow x = x =$

Index :

1. Differential calculus

2. Power rules and derivative rules

3. Product rules

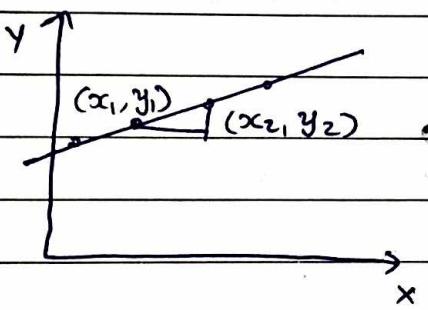
4. Chain rules of derivatives

(1) Differential calculus :

- slope : a unit of change

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.

The slope indicates the ratio of the vertical change (rise) to the horizontal change (run) between two points on a line.



(x<sub>1</sub>, y<sub>1</sub>)

(x<sub>2</sub>, y<sub>2</sub>)

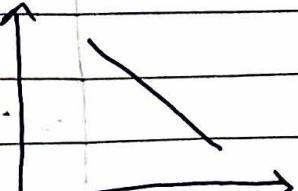
$$\text{slope} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x} = \frac{\text{rise}}{\text{run}}$$

since the line is straight  
Rate of change = constant.

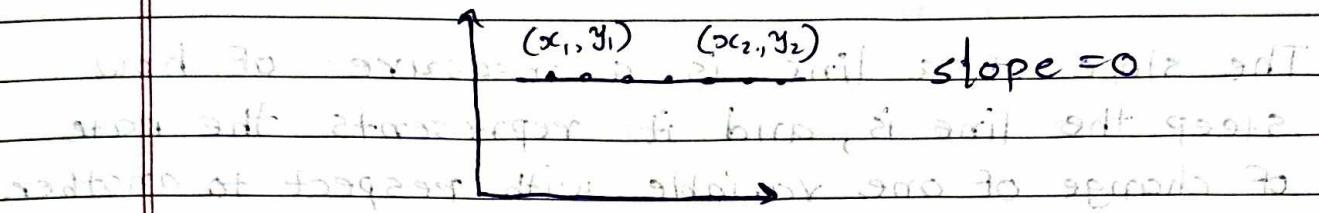
- Interpretation of slope :

(i) Positive slope : if slope  $> 0$ , the line rise as it moves from left to right, the larger the slope, the steeper the line.

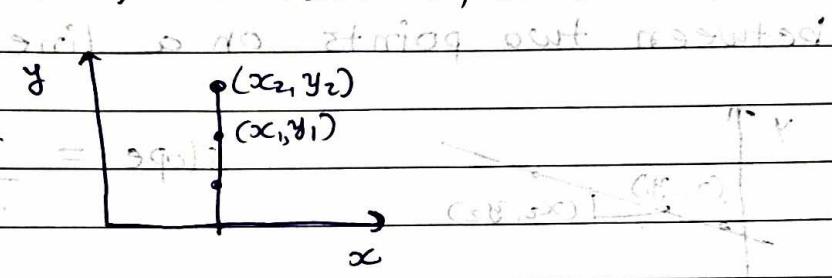
(ii) Negative slope : if slope  $< 0$ , the line falls as it move from left to right, the more negative the slope, the steeper the line in downward direction.



(iii) zero slope: if slope = 0, the line is horizontal, means there is no vertical change as the line moves from left  $\rightarrow$  Right.

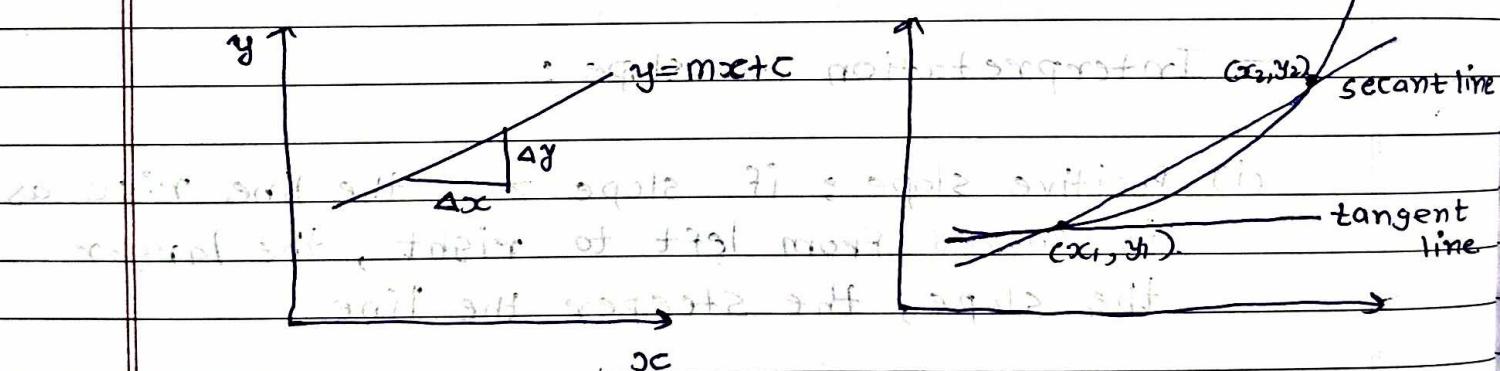


(iv) Undefined slope: if  $x_2 = x_1$ , the line is vertical and the slope is undefined.

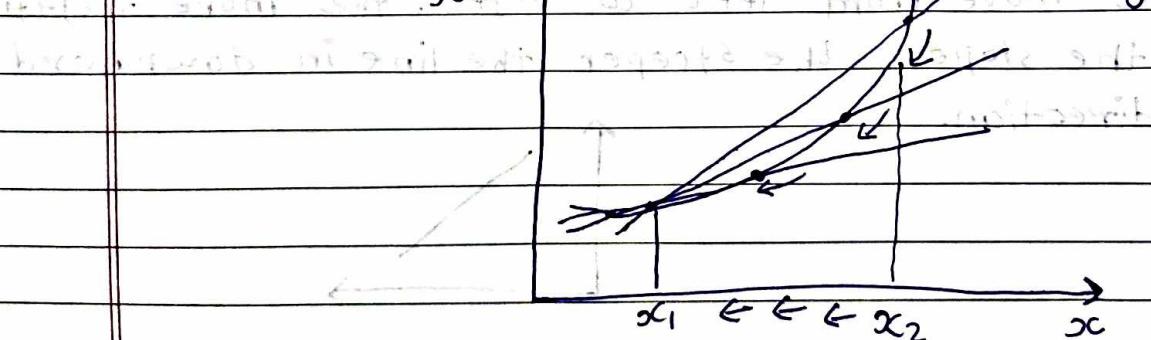


if  $y = mx + c$  is a straight line

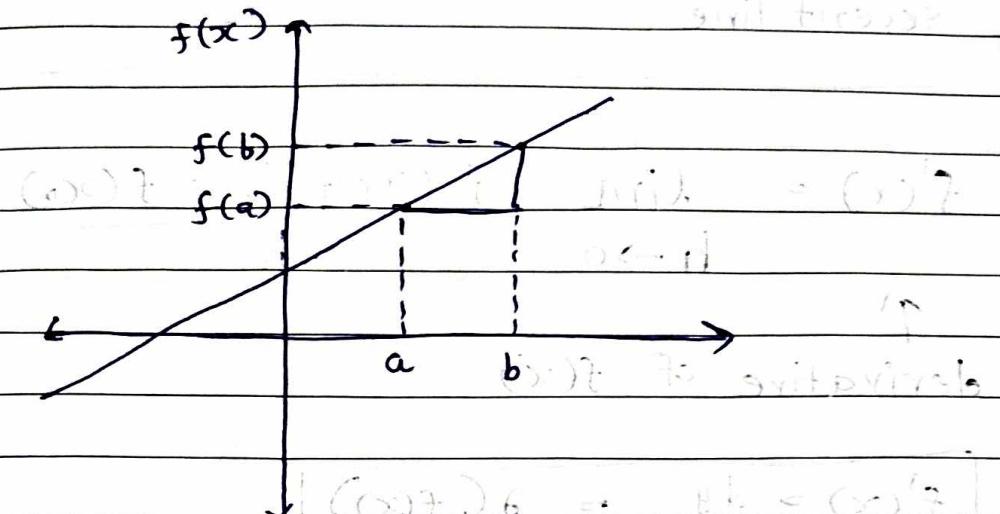
then  $m = \underline{\text{Derivative}}$



so if  $y = mx + c$ , then slope =  $\frac{\Delta y}{\Delta x}$



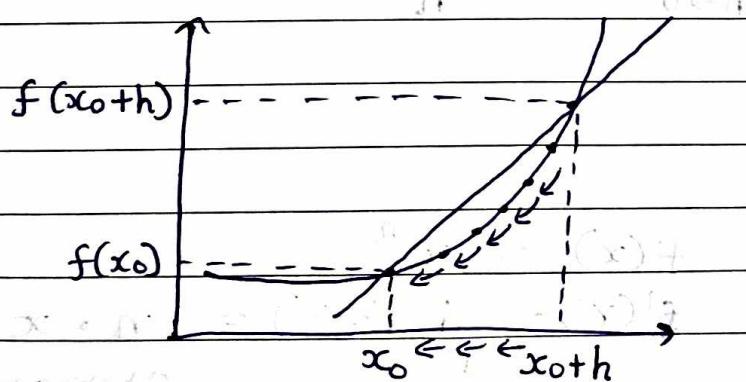
## - Mathematical Notation of Derivative with Limits



$$f(x) = mx + c \quad (m = \text{slope}, c = \text{intercept})$$

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

$$\boxed{\frac{\Delta y}{\Delta x} = \frac{f(b) - f(a)}{x_2 - x_1}}$$



Now, ( $h \rightarrow 0$ )

$$\text{slope} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{f(x_0 + h) - f(x_0)}{(x_0 + h) - x_0} = \frac{\Delta y}{\Delta x}$$

slope of Secant line =  $\frac{f(x_0 + h) - f(x_0)}{h}$

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

derivative of  $f(x)$

$$f'(x) = \frac{dy}{dx} = \frac{d(f(x))}{dx}$$

## (2) Power rules and derivative rules:

- Power rule:

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$$

Formula:

$$\text{If: } f(x) = x^n$$

$$\text{Then: } f'(x) = \frac{d(x^n)}{dx} = n \cdot x^{n-1}$$

if:  $n = 0$  then:  $f(x) = x^0 = 1 = \text{constant}$

$\rightarrow \text{slope} = 0$

$$f'(x) = 0$$

(Derivative of constant = 0)

e.g. : (1)  $f(x) = x^5$

$f'(x) = 5x^4$

(2)  $f(x) = x^{-2}$

$f'(x) = -2x^{-3}$

(3)  $f(x) = x^{1/2}$

$f'(x) = \frac{1}{2} x^{-1/2} = \frac{1}{2\sqrt{x}}$

(4)  $f(x) = 7x^3$

$f'(x) = 21x^2$

- sum of two functions ( $f(x), g(x)$ )

$$\frac{\partial}{\partial x} [f(x) + g(x)] = \frac{\partial f(x)}{\partial x} + \frac{\partial g(x)}{\partial x}$$

$$\frac{\partial}{\partial x} [x^4 + x^{-2}] = \frac{\partial (x^4)}{\partial x} + \frac{\partial (x^{-2})}{\partial x}$$

"  $= 4x^3 + (-2x^{-3})$

"  $= 4x^3 - 2x^{-3}$

- Derivative of trigonometric, logarithmic, exponential, power and constant function.

(i) Trigonometric function:

function	derivative
----------	------------

$\sin(x)$   $\cos(x)$

$\cos(x)$   $-\sin(x)$

$\tan(x)$   $\sec^2(x)$

$\cot(x)$   $-\operatorname{cosec}^2(x)$

eg:  $f(x) = \sin(x) + \cos(x)$   
 $f'(x) = \cos(x) - \sin(x)$

## (ii) Logarithmic function:

$$f(x) = \ln(x) \rightarrow f'(x) = \frac{1}{x}$$

$$f(x) = \ln_a(x) \rightarrow f'(x) = \frac{1}{x \cdot \ln(a)}$$

## (iii) Exponential function.

$$f(x) = e^x \rightarrow f'(x) = e^x$$

$$(e^{x_1})^6 + (e^{x_2})^6 = [e^{x_1} + e^{x_2}]^6$$

## (iv) constant function

$$\frac{(e^{x_1})f(x)}{x_1^6} + \frac{(e^{x_2})f(x)}{x_2^6} = \frac{[e^{x_1} + e^{x_2}]f(x)}{x_1^6}$$

## (v) Polynomial function.

$$f(x) = e^{x^n} \rightarrow f'(x) = n x^{n-1}$$

Identified, differentiate to derivative

then sum up using addition

Logarithmic differentiation (i)  
with respect to what?

$(x)^{\ln a}$	$(x)^{\ln b}$
$(x)^{\ln c}$	$(x)^{\ln d}$
$(x)^{\ln e}$	$(x)^{\ln f}$
$(x)^{\ln g}$	$(x)^{\ln h}$

(3) Product Rule : When multiply?

$$\frac{\partial}{\partial x} [h(x) \cdot f(x)] = h'(x) \cdot f(x) + h(x) \cdot f'(x)$$

e.g.:  $\frac{\partial}{\partial x} (x^2 \cos x) = \frac{\partial}{\partial x} (x^2) \cdot \cos x + x^2 \cdot \frac{\partial}{\partial x} (\cos x)$

$$\frac{\partial}{\partial x} (x^2 \cdot \cos x) = 2x \cdot \cos x + x^2 \sin x$$

$$(2x + x^2 \sin x) = x(2 + x \sin x)$$

(4) chain Rule of Derivative : (use in NN)

- The chain rule is a fundamental theorem in calculus that is used to find the derivative of a composite function.

- When a function is composed of other functions, the chain rule allows us to differentiate it with respect to the innermost variable.

If  $y = f(g(x))$

when  $y = f(u)$  and  $u = g(x)$ , then derivative of  $y$  with respect to  $x$ .

$$\boxed{\frac{\partial y}{\partial x} = \frac{\partial f}{\partial u} \cdot \frac{\partial u}{\partial x}}$$

In simple term:

$$\frac{dy}{dx} = f'(g(x)) \cdot g'(x)$$

This mean that to differentiate a composite function, you first differentiate the outer function with respect to inner function and then multiply by derivative of inner function.

$$\text{eg: } (i) y = (3x^2 + 2x + 1)^5$$

Ques 1. Identify the outer, and Inner function

$$\text{outer } f^n = f(u) = u^5$$

$$\text{Inner function: } u = g(x) = 3x^2 + 2x + 1$$

diff. both of base w.r.t. x

Differentiate the outer function.

$$\text{So, } f'(u) = \frac{df}{du} = u^4 = 5u^4$$

diff. of ex. w.r.t. u (i.e. 5u<sup>4</sup>)

diff. of inner function w.r.t. x (i.e. 3x<sup>2</sup> + 2x + 1)

Differentiate the inner function.

$$g'(x) = \frac{du}{dx} = \frac{d(3x^2 + 2x + 1)}{dx} = 6x + 2$$

$$\therefore \frac{dy}{dx} = f'(g(x)) \cdot g'(x)$$

$$= 5u^4 \cdot (6x + 2)$$

$$\frac{dy}{dx} = 5(3x^2 + 2x + 1)^4 \cdot (6x + 2)$$

$$(ii) \quad y = \sin(4x^3 + x)$$

Outer function =  $\cos(4x^3 + x)$ .  
 Inner function =  $12x^2 + 1$

$$\frac{\partial y}{\partial x} = \cos(4x^3 + x) \cdot (12x^2 + 1)$$

$$(iii) \quad y = \sqrt{\sin(3x)}$$

$$\therefore \text{outer } f^n : \sqrt{u} = u^{\frac{1}{2}} = \frac{1}{2} \frac{u}{\sqrt{u}}$$

$$\therefore \text{middle } f^n : \sin(3x) \rightarrow \cos(3x)$$

$$\therefore \text{Inner } f^n : 3x \rightarrow 3$$

$$\frac{\partial y}{\partial x} = \frac{1}{2\sqrt{\sin(3x)}} \cdot \cos(3x) \cdot 3$$

$$\therefore \frac{\partial y}{\partial x} = \frac{3 \cos(3x)}{2\sqrt{\sin(3x)}}$$

- Application of chain rule :

(i) Backpropagation In NN

(ii) Gradient decent optimization

(iii) Regularization technique (Overfitting, Underfitting)