Mathematics - Basics to Advanced for Data Science & GenAI

INTRODUCTION TO LINEAR ALGEBRA

Linear Algebra is a branch of maths that focuses on the study of vectors, vector/linear spaces, linear transformations, and system of linear eqn.'s.

Applications: ⇒ Data Representation and Manipulation: {Linear Algebra works with higher dimension data}

⇒ ML & AI: → Model Train → Dimensionality Reduction
→ Neural Network → Computer Graphics → Optimization
(Fw & Bw propagation) ↳ (ex: gradient descent)

Scalar: A scalar is a single numerical value. It represents a magnitude (or) quantity & has no direction.

ex: Car Speed = 45 kmph → (magnitude), Simple Linear Regression: $y = mx + c$ (m: slope, c: intercept (scalar value))

Vector: A vector is an ordered list of no.'s. It can represent a point in space (or) quantify with both magnitude & direction.

ex: Speed of the car is 45 kmph & is moving towards East direction: → E (45 kmph)

a vector representing person's weight over time: [70, 72, 75, 73]

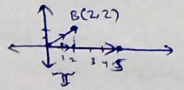
- $\hat{i}, \hat{j}$ ⇒ unit vector towards 'x' & 'y' axis

Addition of 2 vectors: $A = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$, $B = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$ → $A+B = \begin{bmatrix} x_1+x_2 \\ y_1+y_2 \end{bmatrix} = \begin{bmatrix} x_3 \\ y_3 \end{bmatrix}$

(Point in space) → ex: (-4, 3) → represented as $\begin{bmatrix} -4 \\ 3 \end{bmatrix}$

Multiplication of Vectors: 3 types: → Dot Product (Inner Product), Element wise multiplication, Scalar multiplication

- **Dot Product:** The dot prod. of 2 vectors results in a scalar & is calculated as the sum of the products of their corresponding components.

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

length of projected $\vec{B}$

$\vec{A} \cdot \vec{B} = 5 \times 2 + 0 \times 2 = 10$ (scalar val.)

$\vec{A} \cdot \vec{B}$ = (len. of projected $\vec{B}$) • (len. of vector $\vec{A}$)

ex: $A \cdot B^T = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 4 & 5 \end{bmatrix} = 2 \times 4 + 3 \times 5 = 23$

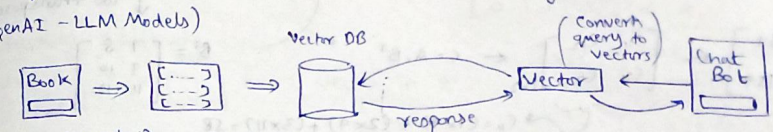
⇒ $\vec{A} \cdot \vec{B} = 0$ → project the vector to the origin

⇒ **Applications:** 1) Cosine Similarity: It's a measure used to determine how similar 2 vectors are. It calculates the cosine of the angle btw 2 vectors, providing a similarity score that ranges -1 (dissimilar) to 1 (complete similar)

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

2) **Vector Databases:** When we talk about Vector database, we basically design a RAG (Retrieval Augmented Generation)

(GenAI - LLM Models)

ex: 

(converted to vectors) (stored in vector DB) (cosine similarity search occurs) → (highest value gets returned)

- **Element Wise Multiplication:** In this, corresponding elements of 2 vectors are multiplied to form a new vector of the same dimension.

ex: $A = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, $B = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$ ⇒ $A \otimes B = \begin{bmatrix} 3 \\ 8 \\ 15 \end{bmatrix}$

⇒ **Applications:** Feature Engineering, Deep Learning

- **Scalar Multiplication:** It involves multiplying vector by a scalar, resulting in a vector where each component is scaled by the vector.

ex: $A = \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}$, $c = 4$ → $cA = \begin{bmatrix} 12 \\ 20 \\ 28 \end{bmatrix}$

⇒ **Applications:** Normalization, Standardization.

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

$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$; a_{ij} → i = row, j = col

ex: Data representation, images in computer vision, confusion matrix, nlp, neural nw

: $\begin{bmatrix} 50 & 10 \\ 5 & 35 \end{bmatrix}_{2 \times 2}$ (50 = true +, 10 = false -, 5 = false -, 35 = true -)

Matrices Operations: → Matrix addition & subtraction, → Scalar Matrix Multiplication, → Matrix Multiplication

- **Matrix Add. & Subtract.:** Add or subtract corresponding ele.'s of 2 matrices of the same dimension.

ex: $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$, $B = \begin{bmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 1 & 2 & 3 \end{bmatrix}$ → $A+B = \begin{bmatrix} 1+4 & 2+5 & 3+6 \\ 4+7 & 5+8 & 6+9 \\ 7+1 & 8+2 & 9+3 \end{bmatrix} = \begin{bmatrix} 5 & 7 & 9 \\ 11 & 13 & 15 \\ 8 & 10 & 12 \end{bmatrix}$

- **Scalar Multiplication**: Scalar multiplication involves multiplying every element of a matrix by a scalar value. ($B = cA$)

- **Matrix Multiplication**: It involves the dot product of rows of the first matrix with columns of the 2nd matrix.

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

ex: $\begin{bmatrix}1&2&3\end{bmatrix}\begin{bmatrix}2\\3\\4\end{bmatrix} = 1\times2 + 2\times3 + 3\times4 = 20$; $A = \begin{bmatrix}1&2&3\\4&5&6\end{bmatrix}_{2\times3}$, $B = \begin{bmatrix}7&9&11\\8&10&12\end{bmatrix}_{2\times3}$

↓ (do transpose)

$\begin{bmatrix}58&64\\139&154\end{bmatrix} \leftarrow \begin{bmatrix}c_{11}&c_{12}\\c_{21}&c_{22}\end{bmatrix}_{2\times2} \leftarrow C = A \cdot B^T$; $B^T = \begin{bmatrix}7&8\\9&10\\11&12\end{bmatrix}_{3\times2}$

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

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

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

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

INTRODUCTION TO FUNCTIONS & TRANSFORMATIONS

Functions: A func. is a mathematical relationship that uniquely associates elements of one set (~ domain) with elements of another set (~ codomain).

- **Notation**: A func. 'f' mapping ele.'s from set X to set Y is denoted by $f: X \to Y$

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

ex: $f(x) = 2x + 3$

Vector transformations: It refer to operations that map vectors from one space to another, often changing their magnitude, direction (or) both. These transformations are typically described using matrices & are fundamental in various fields, including computer graphics & data science.

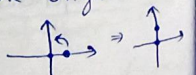
ex:

1) **Scaling**: Scaling is a transformation that changes the magnitude of vector while keeping their direction same.

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

2) **Rotation**: transformation that turns vectors around the origin.

ex: $v = \begin{bmatrix}1\\0\end{bmatrix} \in \mathbb{R}^2$ (2D, real no.) $\longrightarrow v' = \begin{bmatrix}0\\1\end{bmatrix}$ ↝ showing a 90° rotation



3) **Reflection**: transformation that flips vectors over a specified axis (or) plane.

ex: $v = \begin{bmatrix}3\\4\end{bmatrix}$ ⇒ across the Y-axis ↝ $\begin{bmatrix}3\\4\end{bmatrix} \cdot \begin{bmatrix}-1\\1\end{bmatrix} = \begin{bmatrix}-3\\4\end{bmatrix}$ (f(x))

4) **Shearing**: It refers to distortion of shape of an obj. by shifting some of its points in a particular direction.

Linear Transformation: It's a func. btw 2 vector spaces that preserves the operations of vector addition & scalar multiplication. This means that if 'T' is a linear transformation from a vector space V to a vector space W, then for any vectors :2 imp. prop.'s :
- Additivity: $T(u+v) = T(u) + T(v)$
- Homogeneity: $T(cu) = c \cdot T(u)$

where $u, v \in V$ and 'c' is a scalar value.

ex: **reflection**:

The reflec. tran. T across Y-axis maps a vector : $X = \begin{bmatrix}x\\y\end{bmatrix} \in \mathbb{R}^2$, $T(x) = \begin{bmatrix}-x\\y\end{bmatrix}$

∴ Transformation can be expressed as: $A = \begin{bmatrix}-1&0\\0&1\end{bmatrix}$ ↝ $T(x) = A \cdot x$

(we have to check whether it's LT or not)

$= \underset{n\times1}{[x\ y]}\underset{2\times2}{\begin{bmatrix}-1&0\\0&1\end{bmatrix}} = \underset{1\times2}{\begin{bmatrix}-x\\y\end{bmatrix}}$

① Checking additivity:

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

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

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

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

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

RHS ⇒ $T(u) + T(v) = \begin{bmatrix}-u_1\\u_2\end{bmatrix} + \begin{bmatrix}-v_1\\v_2\end{bmatrix} = \begin{bmatrix}-u_1-v_1\\u_2+v_2\end{bmatrix}$ → (RHS)

(LHS = RHS) ✓

② Checking Homogeneity:

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

$cu = c\begin{bmatrix}u_1\\u_2\end{bmatrix} = \begin{bmatrix}cu_1\\cu_2\end{bmatrix}$, $T(cu) = A(cu) = \begin{bmatrix}-1&0\\0&1\end{bmatrix}\begin{bmatrix}cu_1\\cu_2\end{bmatrix} = \begin{bmatrix}-cu_1\\cu_2\end{bmatrix}$ → LHS

$c \cdot T(u) = c(A \cdot u) = c\begin{bmatrix}-1&0\\0&1\end{bmatrix}\begin{bmatrix}u_1\\u_2\end{bmatrix} = c\begin{bmatrix}-u_1\\u_2\end{bmatrix} = \begin{bmatrix}-cu_1\\cu_2\end{bmatrix}$ → RHS

⇒ LHS = RHS

Similarly there are examples which don't follow Linear Transformation.

Why? ↝ used in :

- Dimensionality Reduction (PCA)
- Feature Engg.
- Data preprocessing (Normalization & Standardization)
- Neural Networks
- Image & Signal Processing (CNN)
- Optimizing & Solving sys. of eqn.'s

Also, in LT ↝ Origin must be fixed
↝ All lines must remain lines

(The word 'transformation' means the same thing as func.)

Magnitude & Unit Vector:

ex: $\vec{B} = \begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}$

$\|\vec{B}\| = \sqrt{x_1^2 + x_2^2 + x_3^2}$ } vector length

∴ $\|\vec{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$

Unit Vector: $\|\hat{u}\| = 1$, denoted by $\hat{u}$.

$\vec{u} = \frac{\vec{v}}{\|\vec{v}\|} = \frac{1}{\|\vec{v}\|} \cdot \vec{v}$ → scalar multiplication

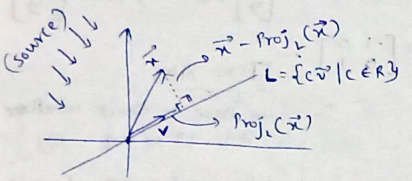
ex: $\vec{v} = \begin{bmatrix}1\\2\\0\end{bmatrix} \in \mathbb{R}^3 \rightarrow \|\vec{v}\| = \sqrt{1^2+2^2+0} = \sqrt{5}$

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

$\|\vec{u}\| = \sqrt{\left(\frac{1}{\sqrt{5}}\right)^2 + \left(\frac{2}{\sqrt{5}}\right)^2 + 0^2} = 1 = \hat{u}$

- A point in space → ex: (2,3) can also be written as: $2\hat{i}+3\hat{j}$.

Introduction to Projections:



; $\text{Proj}_L(\vec{x})$ = projection of vector 'x' to 'l'.
↳ (some vector in line where $\vec{x} - \text{Proj}_L(\vec{x})$ is perpendicular to L)

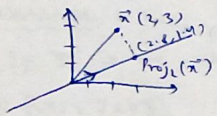
- Dot product of 2 ⊥ vectors = 0.

$\text{Proj}_L(\vec{x}) \Rightarrow (\vec{x} - c\vec{v}) \cdot \vec{v} = 0 \rightarrow \vec{x}\cdot\vec{v} - c\vec{v}\cdot\vec{v} = 0 \rightarrow c = \frac{\vec{x}\cdot\vec{v}}{\vec{v}\cdot\vec{v}}$

$\therefore \text{Proj}_L(\vec{x}) = c\vec{v} = \left(\frac{\vec{x}\cdot\vec{v}}{\vec{v}\cdot\vec{v}}\right)\cdot\vec{v}$ ⟺ $\text{Proj}_b(a) = \left(\frac{a\cdot b}{b\cdot b}\right)\cdot b$

↳ It gives us the point on $\vec{v}$ where projection of $\vec{x}$ is intersected.

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



$\text{Proj}_L(\vec{x}) = \left(\frac{\begin{bmatrix}2\\3\end{bmatrix}\cdot\begin{bmatrix}2\\1\end{bmatrix}}{\begin{bmatrix}2\\1\end{bmatrix}\cdot\begin{bmatrix}2\\1\end{bmatrix}}\right)\cdot\begin{bmatrix}2\\1\end{bmatrix} = \frac{7}{5}\cdot\begin{bmatrix}2\\1\end{bmatrix} = \begin{bmatrix}14/5\\7/5\end{bmatrix}$

$\therefore \text{Proj}_L(\vec{x}) = \begin{bmatrix}2.8\\1.4\end{bmatrix}$

INVERSE FUNCTIONS OR TRANSFORMATIONS

Inverse of a function:

If func. 'f' maps an ele 'x' from 'set x' to an ele. 'y' in 'set y'. $(f: X \rightarrow Y)$, the inverse func. f^{-1} maps 'y' back to 'x' $(f^{-1}: Y \rightarrow X)$, for every $y \in Y$, there is a unique $x \in X$ such that $f(x) = y$.

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

Identity function: $I_X : X \rightarrow X$ - for a set X, the identity func. I_X is defined as: $I_X(a) = a$, for all $a \in X$.

(identity func. on set 'X' & it maps every element 'x' in X to itself.)

Properties:

1) Preservation: doesn't alter any ele.
2) Linearity: Identity func. is a linear transformation.
3) Identity Matrix: all diag.'s = 1 & rest = 0's (n×n) ex: $I_3 = \begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}_{3\times3}$
4) Inverse: Identity func. is it's own inverse.

Existence & Uniqueness: A func. 'f' has an inverse if & only if it is bijective.

- Injective (1-to-1): diff. ele.'s in the domain map to diff. ele.'s in the codomain.
- Surjective (Onto): every ele. in codomain is the image of atleast 1 ele. in the domain.

Find the Inverse: ex: $y = 2x+3$ for 'x' ; $\therefore y = 2x+3 \Rightarrow x = \frac{y-3}{2}$

$\therefore$ Inverse func.: $f^{-1}(y) = \frac{y-3}{2}$

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

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

Applications of Inverse func.'s:

1) Normalization & Standardization (mean) (std. deviation)
↳ Standard Normal Distribution: $\mu = 0$ & $\sigma = 1$
↳ Standard transformation $\Rightarrow z = \frac{x_i - \mu}{\sigma}$
↳ Inverse transformation $\Rightarrow x = z\sigma + \mu$

Use case: After training model on standardized data, the predictions are often rescaled back to the org. scale to interpret the results.

- Min-Max Normalization: → Org. Transformation: $Z = \frac{x - \min(x)}{\max(x) - \min(x)} \Rightarrow T: X \rightarrow Y$
↳ Inverse Transformation: $x = Z(\max(x) - \min(x)) + \min(x)$

2) Distribution of Data :  → Normal distribution

3) Data encryption & decryption

Find inverse of a matrix

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

Determinant: → It's a scalar value.

$\begin{bmatrix}a&b\\c&d\end{bmatrix} \Rightarrow ad - bc \Rightarrow$ scalar, determinant

- If $\det(A)$ (matrix) $\neq 0$ → Inverse of the matrix exists.

EIGEN VECTORS AND EIGEN VALUES

Eigen Value (λ): 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).

For a square matrix 'A' $= \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$, an eigen vector & its corresponding eigen value 'λ' satisfies : $\boxed{Av = \lambda v}$

1) **Find Eigen Values:** $\det(A - \lambda I) = 0$

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

$= \begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix}$ → $\det. = (4-\lambda)(3-\lambda) - (2)(1)$

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

2) **Solve the eqn.:** $\lambda = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$ ⇝ $\therefore \lambda_1 = 5,\ \lambda_2 = 2$ } eigen values of 'A'

3) **Find eigen vectors:** $(A - \lambda I)v = 0$

For $\lambda_1 = 5$ ⇝ $A - 5I = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} - \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}$ → $\begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = 0$ → $-x + y = 0 \Rightarrow x = y$

For $\lambda_2 = 2$ ⇝ $A - 2I = \begin{bmatrix} 4-2 & 1 \\ 2 & 3-2 \end{bmatrix}$ → $\begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 0$ ⇝ $2x + y = 0 \Rightarrow y = -2x$

$\therefore A = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$, $\lambda_1 = 5$, $\lambda_2 = 2$ → $v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $v_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$

$(A\vec{v} = \lambda\vec{v})$ ← eigen value

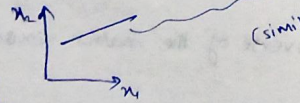
These eigen vectors & values describe how a matrix A scales & rotates vectors in its transformation. EigenValues indicates the factor by which the eigenvectors are stretched (or) compressed, and eigenvectors provide the directions in which this stretching (or) compression occurs.

Application: 'PCA' in dimensionality reduction.

EQN. OF A LINE, PLANE, HYPERPLANE

Line eqn.: $y = mx + c$ (or) $ax + by + c = 0$ (or) $w_1x_1 + w_2x_2 + b = 0$ ← eqn. of straight line.

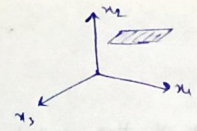
(2D) ; (slope)



(similar) $(w^Tx + b = 0)$

- If line passes via origin → $b = 0$ } ⇝ $\therefore w^Tx = 0$

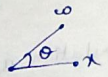
For 3D plane :



$w_1x_1 + w_2x_2 + w_3x_3 + b = 0$ ⇒ $w^Tx + b = 0$

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

for plane (line via origin) : $\Pi_n : w^Tx = 0$



$w.x = w^Tx = |w||x|\cos\theta$, 'w' will always be ⊥ to plane if $\theta = 90°$

INTRODUCTION TO STATISTICS

Statistics is the science of collecting, organizing & analyzing data.

Application:
- Data Exploration & Summarize
- Model building & Validation
- Statistical Analysis
- Hypothesis testing
- Optimization & Efficiency
- Reporting

Types of Statistics :

⇒ Descriptive Statistics	Inferential Statistics
• Involves summarizing & organizing data, to make it understandable.	• Involves methods for making predictions or inferences about a population based on a sample of data.
• Measures of central Tendency → Mean, Median, Mode	• Hypothesis Testing
• Measure of dispersion (variance, std. deviation)	• P-value
• Data Distributions : Histograms, Box plot, Pie Chart	• Confidence Interval
• Summary Statistics : Five No. Summary Q1, Q2, Q3, Maximum	• Statistical Analysis Test: → Z-test, ANOVA, Chi Square test

ex: Let say there are 20 students class in your college, & you have collected the height of students in the class.

↳ Descriptive : "What is the avg. height of the entire classroom"

↳ Inferential : "Are the heights of the sample students in classroom similar to what you expect in the entire college?"

⇒ Population	Sample Data
• It's the entire set of objects of interest in a particular study.	• A sample is a subset of the population that is used to represent the entire group.
• Contains all the observation of interest	• represents a portion of population.
- parameter: a numerical value summarizing the entire population.	- parameter: a numerical value summarizing the sample data.
↳ σ^2, μ → mean, ↳ variance	↳ (mean, variance)
	- samples should be randomly selected to avoid bias.
ex: all consumers in a city	ex: 500 consumers from the city
⇓ to understand the purchasing behaviour of all customers.	⇓ Behaviour → population.

Types of Sampling Techniques:

1) Probability Sampling:

a) **Simple Random Sampling**: every member of the population has an equal chance of being selected.

ex: selecting people randomly

b) **Systematic Sampling**: select every n^{th} member of the population after a random starting point.

ex: Feedback Survey → selected every 7th person's feedback.

c) **Stratified Sampling**: divide the population into strata (groups) based on specific characteristics & then randomly sampling from each strata.

d) **Cluster Sampling**: divide the population into clusters, randomly selecting clusters, then sampling all the members from the selected clusters.

e) **Multi Stage Sampling**: combining several sampling methods. Usually involves selecting clusters, then randomly sampling within those cluster.

2) Non Probability Sampling:

Select individuals who are easiest to reach. ex: surveying people at mall.

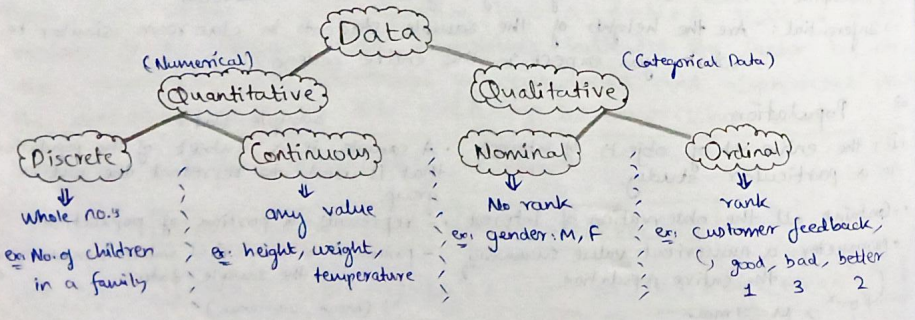
a) **Convenience Sampling**: selecting individual who are easiest to reach.

b) **Judgemental Sampling**: select individual based on the researcher's judgement
ex: choose experts in a field to participate.

c) **Snowball Sampling**: existing study subjects, recruite future subjects from among their acquiantances.

ex: survey members of a rare disease.

d) **Quota Sampling**: sampling based on age, group, gender, etc.



Scales of Measurement of Data:

The scales of measurement describe the nature of info. within the values assigned to variables. We have 4 primary scales:

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

1) Nominal Scale:

This scale classifies data into distinct categories that don't have an intrinsic order.

Characteristics:
- → Data is categorized based on labels, names or qualities
- → Categories are mutually exclusive
- → No logical order among categories [No rank].

ex: Gender → M, F

2) Ordinal Scale:

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

Characteristics:
- → Data is categorized & ranked in a specific order
- → The interval btw ranks are not necessarily equal.

ex: customer feedback → Satisfied (1), Very Satisfied (2), Not satisfied (3)

3) Interval Scale:

The interval scale not only categories & orders, but also specify the exact difference btw intervals. It lacks a true zero point.

Characteristics:
- → Data is ordered with consistent interval btw values
- → Allows for meaningful comparison of differences (Ratio can't be measured.)
- → No true zero point.

ex: IQ Scores: 90, 100, 110 ⇝ difference = 100−90 = 10 (IQ ≠ 0)

4) Ratio Scale:

- → The order matters
- → Differences are measurable.
- → (Ratio can be measured).
- → Contains a 0 starting point.

ex: Student marks
- ⇝ 0, 90, 60, 30
- ⇝ asc. 0, 30, 60, 90

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

DESCRIPTIVE STATISTICS

Measures of Central Tendency:

These are statistical measures which provides a single value that summarizes a set of data by identifying the central position within that dataset.

1) Mean: sum of all values divided by the no. of values.

→ Population Mean (μ), Population size (N): $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

→ Sample mean ($\bar{x}$), sample size (n), ⇝ $n \leq N$: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Here 'x' = random variable.

Characteristics:
- → Affected by extreme outliers.
- → Used for interval & ratio data.

ex: $x = \{1,2,3,4,5\} \rightarrow \mu = \frac{1+2+3+4+5}{5} = 3$

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

2) Median: It's the middle value in a dataset when the values are arranged in ↑ (or) ↓ order.

Characteristics:
- → Not affected by extreme outliers
- → Used for ordinal, interval & ratio data.

ex: $x = \{1,2,3,4,5\} \rightarrow$ No. of ele. = 5 (odd) ∴ Median = 3

$x = \{1,2,3,4,5,10\} \rightarrow$ No. of ele. = 6 (even) ∴ Median = $\frac{3+4}{2} = 3.5$

3) **Mode**: the value that appears most frequently in a dataset.

Chars: → Not affected by extreme values. → Used for nominal, ordinal, interval & ratio data.

Choosing the appropriate measure:

1) **Mean**: Best used when data is symmetrically distributed without outliers. Provides a mathematical avg., which is useful for further statistical calculations.

2) **Median**: Best used when data is skewed or contain outliers. Provides the middle value, which better represents the center of a skewed dataset.

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

Measure of Dispersion:

These describe the spread or variability of a dataset. They indicate how much the values in a dataset differ from the central tendency.

1) **Range**: It's the difference btw the max. & min. value in a dataset.

Chars: → Simple to calculate
→ Sensitive to outliers
→ Rough measure of dispersion.

• Range = Max. Value − Min. Value

2) **Variance**: It measures the avg. squared deviation of each value from the mean. It provides a sense of how much the values in a dataset vary.

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

(x_i = data points, μ = population mean, N = population size)

→ Sample Variance: $S^2 = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{n-1}$

(x_i = data points, $\bar{x}$ = sample mean, n = sample size)

Chars:
→ Provides a precise measure of variability.
→ Units are squared of the original data units.
→ More sensitive to outliers than the range.

Variance: determines the spread of the data.

3) **Standard Deviation**: It's the square root of the variance.

Chars: → Provides a clear measure of spread in the same units as the data.
→ Sensitive to outliers.

• If we have std. deviation, we can find out variance & vice-versa.

• Variance = Provides a measure of dispersion of data points in squared units, which can be difficult to interpret directly.

Standard deviation = provides a measure of the dispersion in the same units as the original data, making it easier to interpret & understand.

Why 'n-1' in sample variance?

When we calc. var. using just the sample data, we're only working with a small part of the whole population. This can lead to underestimating the true variability, which is the bias we're correcting. By using 'n-1', we adjust for this & get a more accurate estimate.

$S^2 = \sum_{i=1}^{n} \frac{(x - \bar{x})^2}{n-1}$ → Bessel's Correction

Random Variables: X: it's a func. which derives it's values from process or experiments.

- **Discrete Random**
  ex: tossing a coin, rolling a dice
- **Continuous Random Variable**
  ex: tmrw how many inches it's going to rain [0, 1.1, 5.5, 10-5, 10.75]

• **Percentage**: ex: {1,2,3,4,5,6} → percentage of odd no.'s in this group $= \frac{3}{6} \times 100 = 50\%$

• **Percentile**: A percentile is a value below which certain % of observations lie

ex: {2,2,3,4,5,5,6,7,8,8,8,9,9,10}, x = 9

Percentile of value x = $\frac{\text{No. of values below } x}{n} \times 100 = \frac{11}{14} \times 100 = 78.57\%$ of value 9

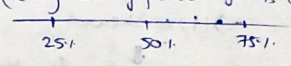
(means 78.57% of the entire distribution is less than 9.)

If percentile is given & we need the value;

ex: 25.5 → Value = $\frac{\text{Percentile}}{100} \times (n+1) \Rightarrow$ Val. = $\frac{25}{100} \times 15 = 3.75$

(position) $\frac{3+4}{2} \approx 3.5$ ∴ 25% is 3.75 (gives us position in dataset)

• **Quartiles**: 25% = 1st Quartile
50% = 2nd Quartile
75% = 3rd Quartile
... : (every 25% of percentage is defined by quartile)



5 Number Summary: (to identify & remove the outliers)

1) Minimum 2) Q1 (25%) → 1st Quartile 3) Median 4) 3rd Quartile (75%) 5) Maximum

ex: 1,2,2,2,3,3,4,5,5,5,6,6,6,7,8,8,9,27 ← outlier

1st Step: (to remove outliers): define lower fence & higher fence value. Anything higher than higher fence value will be removed as an outlier & anything lower than lower fence value will be removed as an outlier.

• Lower fence = Q1 − 1.5(IQR) • Higher fence = Q3 + 1.5(IQR)

(IQR = Inter quartile range (Q3 − Q1))

∴ Q1 = $\frac{\text{Percentile}}{100} \times (n+1) = \frac{25}{100} \times 20 = 5^{th}$ position = 3

Q3 = $\frac{75}{100} \times 20 = 15^{th}$ position = 7 → IQR = 7 − 3 = 4

∴ [Lower fence (3 − 1.5(4)) ⟷ Higher fence (7 + 1.5(4))] = [−3 ⟷ 13]

∴ '27' eliminated as an outlier

These 5 values define 5 no. summary & based on this we can define a box plot → {used to visualize the outliers}.

Histogram & Skewness.

(Graphical representation of the distribution of numerical data.)

ex: ages = {11, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50} → (histogram)

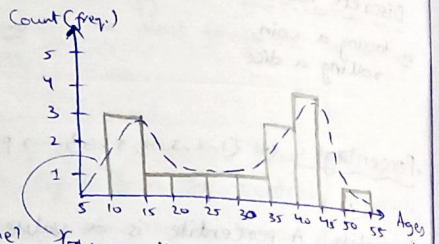
range : 0 - 50

let no. of bins = 10 ∴ $\frac{50}{10} = 5$ → bin size

(no. of intervals into which my data range is divided.)

(width of these intervals)

∴ Bins → [0-5, 5-10, 10-15, ---, 45-50] : (≥5, <10)

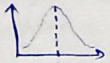


How it's done? → (Kernel density estimator)

This smoothening of graph is known as: probability distribution of a continuous variable.

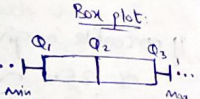
• Skewness.

→ (Normal/Gaussian Distribution) → Symmetrical Distribution → No Skewness



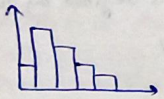
(Mean = Median = Mode) ← mean, median, mode all perfectly at the center.

Box plot:

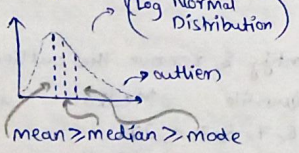


$Q_3 - Q_2 \approx Q_2 - Q_1$

→ Right Skewed:

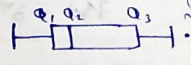


⇒ (+)ve Skewed



mean ≥ median ≥ mode

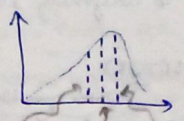
⇒ Box plot:



may have outliers, so denoted like this.

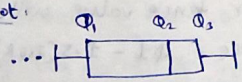
$Q_3 - Q_2 \geq Q_2 - Q_1$

→ Left Skewed:



mean ≤ median ≤ mode

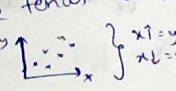
⇒ (-)ve Skewed → Box plot:

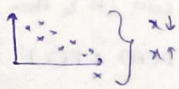


$Q_2 - Q_1 \geq Q_3 - Q_2$

• Covariance & Correlation are 2 statistical measures used to determine the relationship btw 2 variables. Both are used to understand how changes in 1 variable are associated with changes in another variable.

Covariance: It's a measure of how 2 random variables change together. If the variables tend to ↑ & ↓ together, the covariance is (+)ve. If one tends to ↑ when the other ↓, the covariance is (-)ve.

ex:  {x↑ = y↑, x↓ = y↓} (+) covariance

 {x↓ = y↑, x↑ = y↓} (-) covariance

$\rightarrow Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} \Rightarrow Cov(x,x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{n-1} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

$\therefore Cov(x,x) = Var(x)$

[x_i = datapoint of random variable 'x'.
$\bar{x}$ = sample mean of 'x'
y_i = datapoint of random variable 'y'.
$\bar{y}$ = sample mean of 'y'.]

Adv.: Quantify the relationship btw 'x' & 'y'

Disadv.: Co-var. doesn't have a specific limit value.

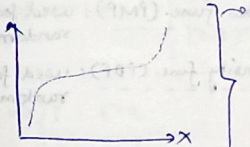
$Cov(x,y) \rightarrow -\infty$ to $+\infty$

Correlation: → Pearson Correlation Coefficient; → Spearman Rank Correlation

1) Pearson Correlation Coefficient: $\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$ (std. deviation) } range: [-1 to 1]

→ The more the val. towards 1, the more (+)ve correlated 'x' & 'y' are!

→ The more the val. towards -1, the more (-)ve correlated 'x' & 'y' are!

ex: 

Pearson Corr. = 0.88
Spearman Corr. = 1

→ Here we can see: x↑ → y↑ & x↓ → y↓. So in this case pearson corr. must come as +1, but we are getting as 0.88. This is because of the slower growth in middle of graph.

∴ Pearson's Corr. cannot capture correlation for non-linear data. → { → 1 }

2) Spearman Rank Correlation:

$\rho_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) \cdot \sigma(R(y))}$; R → rank ; ex:

x	y	R(x)	R(y)
1	2	2	1
3	4	3	2
5	6	4	3
7	8	5	5
0	7	1	4

Application: Feature Selection.

INTRODUCTION TO PROBABILITY

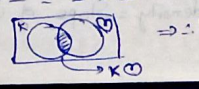
(it's about determining the likelihood of an event.)

Mutual Exclusive Event: 2 events are mutually exclusive if they cannot occur at the same time.

ex: tossing a coin. (H) (T) $P(H) = \frac{1}{2}$, $P(T) = \frac{1}{2}$

$P(H \text{ or } T) = P(H) + P(T)$ {Additive rule for mutual exclusive event}
$= \frac{1}{2} + \frac{1}{2} = 1$

Non-mutual exclusive event: ex: taking a card from a deck.

K, ♡ ~ 

$\Rightarrow \therefore P(K \text{ or } ♡) = P(K) + P(♡) - P(K \text{ and } ♡)$
$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$

(non-mutual exclusive event.)

Multiplication Rule: {Independent & Dependent Events}

→ **Independent**: 2 events are independent if they don't affect one another.
ex: tossing a coin {H & then Tails} ↳ P(H) = 1/2, P(T) = 1/2

• P(H and T) = P(H) * P(T) = 1/2 × 1/2 = 1/4

→ **Dependent**: 2 events are dependent if they affect each other.
ex: take a king from the deck & then the queen card from the deck

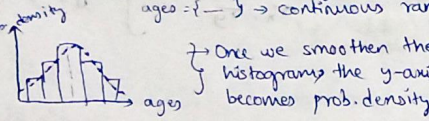
↳ P(K) = 4/52, P(Q) = 4/51

• P(K and Q) = P(K) * P(Q/K) → {Conditional probability}
= 4/52 × 4/51

PROBABILITY DISTRIBUTION FUNCTION & TYPES OF DISTRIBUTION

Probability Distribution Function:

↳ describes the likelihood of different outcomes in a random event.

ex: 

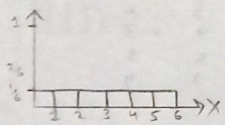
Types of probability dist. func. : → Probability Mass func. (PMF): used for discrete random variables
Cumulative distribution func. (CDF)
Probability Density func. (PDF): used for continuous random variables

→ PMF: [Discrete Random Variables]

ex: rolling a dice X = {1,2,3,4,5,6}

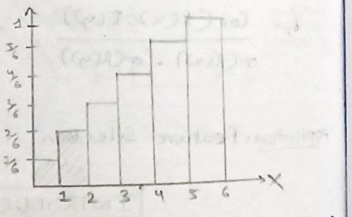
P(1) = P(2) = ... = P(6) = 1/6

PMF:

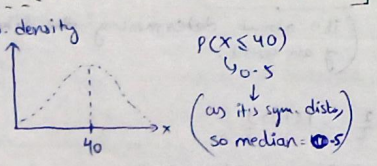


P(X ≤ 2) = P(X=1) + P(X=2)
= 1/6 + 1/6 = 2/6 = 1/3

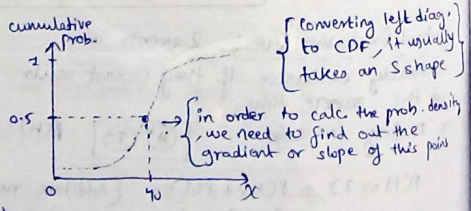
Cumulative Density func. (CDF)



→ PDF: [Continuous Random Variable]



P(X ≤ 40) ↳ 0.5 (as it's sym. dist., so median = 0.5)



{converting left diag. to CDF, it usually takes an S shape}

→ {in order to calc. the prob. density, we need to find out the gradient or slope of this point}

Slope = $\frac{y_2 - y_1}{x_2 - x_1}$ } gradient ⇒ prob. density

∴ Prob. density = Slope of cumulative density func.

{How we made this? We created a hist. & then smoothened it.}

Properties: → Non negativity $f(x) \geq 0$ for all x {func. which converts prob. density to CDF}

→ Total area under the PDF curve is equal to 1. ⇒ $\int_{-\infty}^{\infty} f(x)\,dx = 1$

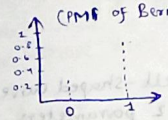
* wrt diff. distribution, f(x) func. is going to change.

→ Types of probability distribution:

1) Bernoulli dist. → outcomes are binary (pmf) → discrete random variables
2) Binomial dist. → pmf
3) Normal / Gaussian dist. → pdf
4) Poisson dist. → pmf
5) Log Normal dist. → pdf
6) Uniform dist. → pmf

→ Bernoulli Distribution:

It's the simplest discrete prob. dist. which represents the probab. dist. of a random variable that has exactly 2 possible outcomes: success (with prob. 'p') & failure (with prob. '1-p'). It is used to model binary outcomes.

• 

--: P(X=0) = 0.2, P(X=1) = 0.8

• Parameters: $0 \leq p \leq 1$, $q = 1 - p$, $k = \{0, 1\}$ ⇒ 2 outcomes
P(success) ⇒ k = 1
P(failure) ⇒ k = 0

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

• Mean of bern. dist. = 'p'

• Median of bern. dist. = $\begin{cases} 0 & ; \text{if } p < 1/2 \\ [0,1] & ; \text{if } p = 1/2 \\ 1 & ; \text{if } p > 1/2 \end{cases}$ ≡ $\begin{cases} \text{median} = 0 & \text{if } q > p \\ \text{median} = 0.5 & \text{if } q = p \\ \text{median} = 1 & \text{if } q < p \end{cases}$

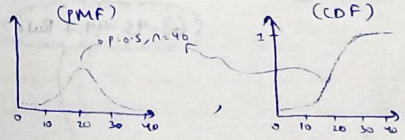
• Mode ⇒ $\begin{cases} p & ; p > q \\ q & ; q > p \end{cases}$

• Variance (σ^2) = pq

• Std. dev. (σ) = $\sqrt{pq}$

→ Binomial Distribution:

It's a discrete prob. dist. that gives the no. of successes in 'n' independent trails, each with a success prob. 'p' & failure prob. 'q' (q = 1-p). Every outcome of the experiment is binary. It's like collection of bernoulli distribution.

• 

• Notation: B(n, p)

• Parameters: $n \in \{0, 1, 2, \ldots\}$ ⇒ no. of trails (or) experiments
$p \in [0, 1]$
$q = 1 - p$

• Support: $k \in \{0, 1, 2, \ldots, n\}$ ⇒ no. of successes

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

for $k = 0, 1, 2, \ldots, n$ where ${}^nC_k = \frac{n!}{k!(n-k)!}$ → binomial coefficient

• Mean = np • Variance = npq • Std. dev. = $\sqrt{npq}$

ex: Quality Control: Scenario: inspecting 10 items in a factory where each item has a 10% chance of being defective.

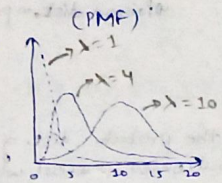
n = 10, prob. of success (p) = 0.1 (defective item), no. of successes (k): varies from 0 to 10.

(d) What is the prob. of finding exactly 2 defective items in a sample of 10?

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

→ Poisson Distribution:

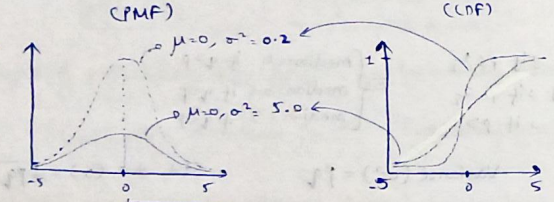
It's a discrete prob. dist. that expresses the prob. of a given no. of events occuring within a fixed interval of time (or space), assuming these events occur with a const. mean rate (λ) & independently of the time since the last event. (ex: counting how many cars pass through a checkpoint per hour.)



- λ = expected no. of events occuring at every time interval
- PMF $\Rightarrow P(x) = \frac{e^{-\lambda}(\lambda)^x}{x!}$
- Mean = Variance = $\lambda * t$ (time interval)

→ Normal / Gaussian distribution:

It's a continuous prob. dist. characterized by its bell-shaped curve, which is symmetric around its mean. It's defined by 2 parameters: the mean (avg.) μ & the std. dev. (σ), which measures the spread of the data.



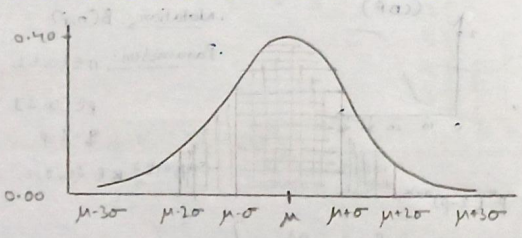
- Notation: $N(\mu, \sigma^2)$
- Parameters: $\mu \in R$ = mean, $\sigma^2 \in R > 0$ = Var., $x \in R$
- PDF $\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} . e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$

$\sigma^2 \uparrow\uparrow$ = spread $\uparrow\uparrow$ (mean = median = mode)

- Mean $\Rightarrow \mu = \sum_{i=1}^{n} \frac{x_i}{n}$
- Variance $\Rightarrow \sigma^2 = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{n}$
- Std. dev. = σ

. Empirical Rule of Normal / Gaussian Distribution:

Let size = 50, assuming the random var. follows → x ~ ... y → (Normal/Gaussian dist.)



68-95-99.7 Rule

$P(\mu-\sigma \le x \le \mu+\sigma) \approx 68\%$: (prob. of rand. var. to be btw '$\mu-\sigma$' & '$\mu+\sigma$')

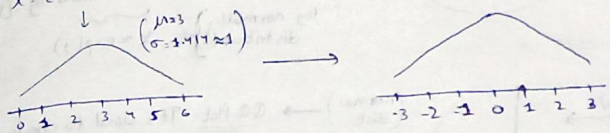
$P(\mu-2\sigma \le x \le \mu+2\sigma) \approx 95\%$

$P(\mu-3\sigma \le x \le \mu+3\sigma) \approx 99.7\%$

→ Standard Normal Distribution:

Let's say that I am having a gaussian distribution with some mean & std. dev. And now I want to convert this distribution into a different distribution where I get mean = 0 & std. dev. = 1. Then this distribution is specifically called as "standard normal distribution". For converting, we will use the formula called as Z-Score

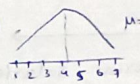
ex: $x = \{1, 2, 3, 4, 5\}$ → ($\mu = 3$, $\sigma = 1.414 \approx 1$) → ($\mu = 0$, $\sigma = 1$)



- Z-score = $\frac{x_i - \mu}{\sigma}$

(helps in bringing diff. features into 1 scale so that model can improve itself effectively.)

(Q) $\mu = 4$, $\sigma = 1$. How many std. dev. 4.25 is away from mean?



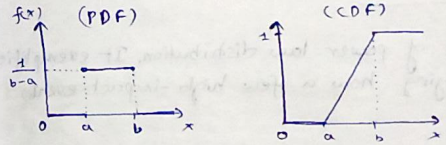
A) $x_i = 4.25 \rightarrow$ Z-score = $\frac{4.25-4}{1} = 0.25$ { means it's 0.25 away from mean

. Z-score is the result of standardization.

→ Uniform Distribution:

1) Continuous Uniform Distribution : [Continuous Random Variable]

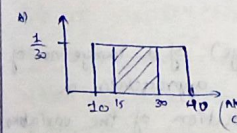
It's a probability distribution where all outcomes in a range are equally likely. ex: picking a random no. btw 0 and 1



- Notation: U(a, b)
- Parameters: $-\infty < a < b < \infty$
- PDF $\Rightarrow \begin{cases} \frac{1}{b-a} & ; x \in [a,b] \\ 0 & ; \text{otherwise} \end{cases}$
- Mean = Median = $\frac{1}{2}(a+b)$
- Variance = $\frac{1}{12}(b-a)^2$
- CDF $\Rightarrow \begin{cases} 0 & ; x < a \\ \frac{x-a}{b-a} & ; x \in [a,b] \\ 1 & ; x > b \end{cases}$

ex: The no. of candies sold daily at a shop is uniformly distributed with a max. of 40 candies & a min. of 10 candies. Find prob. of daily sales to fall btw 15 & 30?

A)

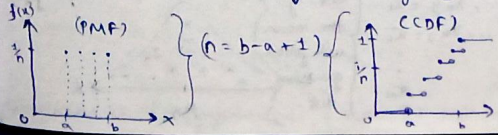


$x_1 = 15$, $x_2 = 30$ → $P(15 \le x \le 30) = (x_2 - x_1) * \frac{1}{b-a}$

$= (30-15) * \frac{1}{30} = 0.5$

2) Discrete Uniform Distribution:

It's a prob. distribution where each possible outcome of a discrete random variable has an equal chance of occuring. ex: rolling a dice → each face has prob. of 1/6.

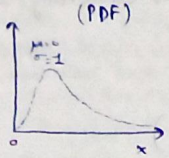


(n = b - a + 1)

- Notation → U(a, b)
- Parameters: a, b where $b \ge a$
- PMF $\Rightarrow \frac{1}{n}$
- Mean = Median = $\left(\frac{a+b}{2}\right)$

→ Log Normal Distribution:

It is a prob. distribution of a random variable whose logarithm is normally distributed. This means if you take the natural logarithm of the values, they will follow a normal distribution.

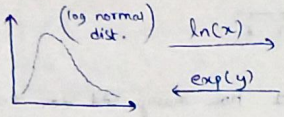


$X \sim$ Log Normal Distribution $(\mu, \sigma) \Rightarrow Y = \ln(X) =$ normal distribution

is a right-skewed distribution.

Natural log: $[\log_e]$

log normally distributed ⇐ $X = \exp(Y)$

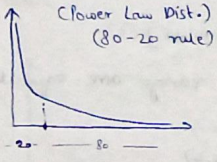


→ QQ Plot: It is used to check whether a distribution is log normal or not

ex: Wealth distribution of the world, salaries of employees in a company.

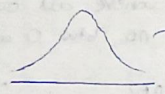
→ Power Law Distribution:

It is a type of statistical distribution where the freq. of an event decreases as the event size increases. In other words, small events are very common, & large events are rare but significant.

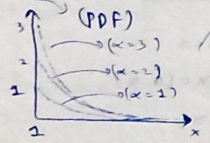


ex: 80% of the wealth are distributed with 20% of the total population.

Box-Cox: transform → (checked via QQ plot)



- Pareto distribution is a specific type of power law distribution. It exemplifies the power law concept by quantifying how a few high-impact events dominate many low-impact ones.



→ using 'Box-Cox' transformation, we can covert it to a normal dist.

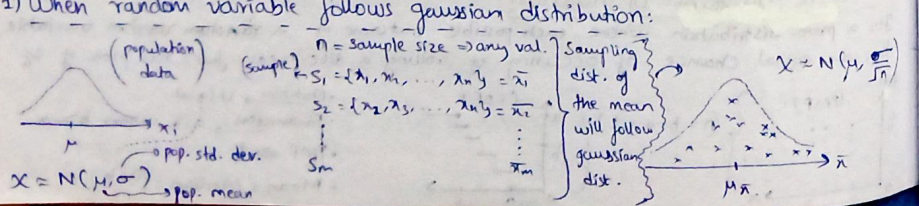
- α = shape parameter. It determines the distribution's 'fatness' (or) conc. of wealth, size, (or) other measured quantities.
- Higher 'α' = few large values & more small values indicating a steeper drop-off.

ex: 80% of the defects can be solved if we solve 20% of the defects.

→ Central Limit Theorem:

It states that the distribution of the sum (or average) of a large no. of independent, identically distributed random variable approaches a normal distribution, regardless of the original distribution of the variables.

(1) When random variable follows gaussian distribution:



(2) When random variable doesn't follows gaussian dist.

$X \neq N(\mu, \sigma)$: → Here, specifically we must have: $n \geq 30$ (sample size)

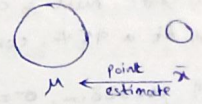
(central limit theo.) ← CLT → $\mu_{\bar{x}}$, $\bar{x}$

Estimates: It's a specified observed numerical value used to estimate an unknown population parameter.

Types:

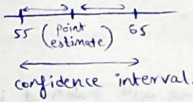
→ Point Estimate: single numerical value used to estimate an unknown population parameter.

ex: Sample mean is a point estimate of a population mean.



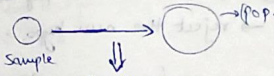
→ Interval Estimate: range of values used to estimate the unknown population parameter

ex:



INFERENTIAL STATS AND HYPOTHESIS TESTING

The main aim of inferential stats is to come up with conclusions (or) inferences from a sample data about an unknown population parameter.



Conclusions ↳ using 'Hypothesis Testing'

Hypothesis Testing Mechanism:

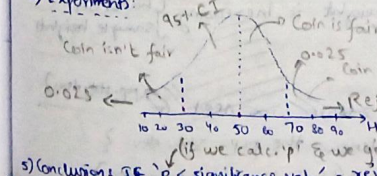
1) Null Hypothesis (H_0): The assumption you are beginning with.
2) Alternate Hypothesis (H_1): opposite of null hypothesis
3) Experiments: statistical analysis (ex: p-value, significance test).
4) Accept or reject the null hypothesis

P-Value:

It's a no. that describes how likely you are to have found a particular set of observations if the null hypothesis were true.

Hypothesis testing: ex: Coin is fair or not {100 times} → P(H) = 0.5, P(T) = 0.5; P(H) = 0.6, P(T) = 0.4; P(H) = 0.7, P(T) = 0.3

1) Null Hypothesis: H_0 = Coin is fair
2) Alternate Hypothesis: H_1 = Coin is not fair
3) Experiments:



4) Significance Value (α): Let $\alpha = 0.05$ ↳ threshold for determining whether a result is statistically significant.

∴ Confidence Interval = $1 - 0.05 = 0.95$

5) Conclusion: If 'p < significance val.' = reject the null hypo., else fail to reject the null hypo.

Hypothesis Testing & Statistical Analysis:

1) Z-test 2) t-test 3) Chi Square (Categorical data) 4) Anova (Variance)

average → z-table, t-table

→ Z-test: We must have: pop. std. dev. & n (sample size) ≥ 30

ex: The avg. heights of all residents in a city = 168cm with a σ = 3.9. A doctor believes the mean to be diff. He measured the height of 36 individuals & found the avg. height to be 169.5cm.

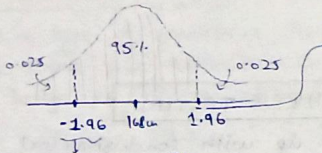
(a) State null & alternate hypothesis

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

Ans) $\mu = 168$cm, $\sigma = 3.9$, $n = 36$, $\bar{x} = 169.5$cm, CI = 0.95 ⇝ $\alpha = 1 - CI = 0.05$

1) Null Hypo.: $H_0 \Rightarrow \mu = 168$cm 2) Alternate Hypo.: $H_1 \Rightarrow \mu \neq 168$cm

3) Based on CI, we will draw "Decision Boundary":



(similar to RHS z-score)

Area (AUC) under curve = 1

to find RHS z-score, we will subtract it from AUC

∴ 1 − 0.025 = 0.9750 → (it's the entire AUC ignoring the part after RHS z-score, denoted by (III))

z-score = 1.96 → (got it from z-table pdf ↗ online)

If 'z' is less than −1.96 (or) greater than +1.96 → reject the null hypo.

* Z-test: $Z_d = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})}$ (for sample data, std. dev. will change) $\left(= Z_{score} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}\right)$

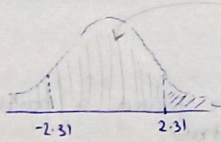
$= \frac{169.5 - 168}{3.9/\sqrt{36}} = 2.31$ } z-score → it tells how many std. dev.'s a data point is from the mean. It's a measure of relative position within a distribution.

∴ Conclusion: 2.31 > 1.96 → reject the null hypo. } ∴ We can confirm that, $p < 0.05$ (significance value.)

Now constructing new graph from above: (2nd way of cross-checking p-value)

z-score of 2.31 = 0.98956 (from z-table pdf)

↳ AUC (IIII)



AUC = 1 − 0.98956 = 0.01044 } ∴ AUC btw [−2.31, 2.31] = 1 − 2×0.01044 = 0.97912

∴ p-value = 0.01044 + 0.01044 = 0.02088 } $p < 0.05$ ⇒ 0.02088 < 0.05 } reject the null hypo.

∴ Final Conclusion: The avg. ≠ 168cm.

Now since z' = 2.31 > 1.96 → means my height is increasing } based on sample height

→ Student t-distribution:

In z-stats when we perform any analysis using z-score, we need σ (pop. std. dev.).

How do we perform our analysis when we don't know the pop. std. dev.?

↳ (Student t distribution)

$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$: (s = sample standard dev.) ; t-table ⇒ t-test

Degree of freedom: refers to the no. of values in the final calc. that are free to vary (wrt t-test).

↳ = n − 1 ; (n = sample size)

→ T-test: (this prob. aka one sample t-test)

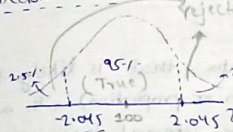
ex: In the pop., the avg. IQ = 100. A team of researchers want to test a new medication to see if it has either a (+)ve (or) (−)ve effect on intelligence, (or) no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a std. dev. of 20. Did the medication affect intelligence? (Let $\alpha = 0.05$ → ∴ CI = 0.95)

Ans) $\mu = 100$, $n = 30$, $\bar{x} = 140$, $s = 20$, CI = 95%, $\alpha = 0.05$

1) Null Hypo.: $H_0 \Rightarrow \mu = 100$ 2) Alt. Hypo.: $H_1 \Rightarrow \mu \neq 100$ (2 tail test) (means it can either be more or less)

2) Significance Level: $\alpha = 0.05$ 3) Dof → n − 1 = 30 − 1 = 29

4) Decision Rule



(due to symm., its (+)ve of RHS)

(from dof & 'α' } we saw the t-table & got the corresponding z-score)

∴ If t-test is less than −2.045 or greater than 2.045, reject the null hypo.

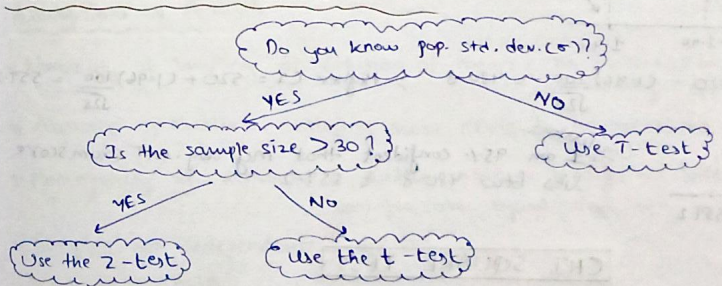
5) Calculate test statistics: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{140 - 100}{20/\sqrt{30}} = 10.96$

Since t = 10.96 > 2.045 ⇒ reject the null hypo.

Conclusion: medication used has affected the intelligence.

↳ & since 't' val. is > 2.045 → ∴ Medication has increased the intelligence.

When to use z-test & t-test:



Type 1 & Type 2 errors:

Reality: Null Hypothesis is true (or) null hypo. is false

Decision: Null Hypo. is true (or) null hypo. is false.

↳ (this is based on our result of hypo. testing)

Outcome 1: We reject the null hypo., when in reality it's false → CORRECT
Outcome 2: We reject the null hypo., when in reality it's true → TYPE 1 ERROR
Outcome 3: We retain the null hypo., when in reality it's false → TYPE 2 ERROR
Outcome 4: We retain the null hypo., when in reality it's true → CORRECT

Bayes Theorem:

Probability → Independent events: ex: rolling a dice
Probability → Dependent events: ex: picking a king & then picking a queen without placing the card in the deck.

- $P(A \text{ and } B) = P(B \text{ and } A) \Rightarrow P(A) * P(B/A) = P(B) * P(A/B)$

(Bayes Theorem) $\Rightarrow P(B/A) = \frac{P(B) * P(A/B)}{P(A)}$ (or) $P(A/B) = \frac{P(A) * P(B/A)}{P(B)}$

(A, B = events ; $P(A/B)$ = prob. of A given B is true ; $P(B/A)$ = prob. of B given A is true ; $P(A), P(B)$ = independent probs. of A & B)

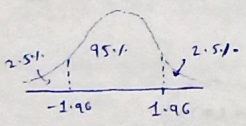
Confidence Interval:

It's a range of values, derived from sample data, that is likely to contain the true population parameter (like a mean or proportion) with a certain level of confidence. for ex., a 95% confidence interval means we're 95% confident that the true values lies within that range.

- CI = Point Estimate ± Margin of Error
- Z test ⇒ $\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$; if we use t-test → replace $(Z_{\alpha/2})$ by $(t_{\alpha/2})$

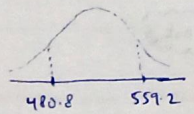
ex: On the verbal section of CAT exam, the std. dev. is known to be 100. A sample of 30 test takers has a mean of 520. Construct 95% CI about the mean.

Sol: $\alpha = 0.05$



$1 - 0.025 = 0.9750$ (2-table) → AUC wrt that = 1.96

Lower CI = $520 - (1.96)\frac{100}{\sqrt{25}} = 480.8$; Higher CI = $520 + (1.96)\frac{100}{\sqrt{25}} = 559.2$



∴ Conclusion: I am 95% confident that my avg. CAT exam score lies btw 480.8 & 559.2

CHI SQUARE TEST

The chi square test for goodness of fit test claims about population proportions. It's a non-parametric test that is performed on categorical [ordinal & nominal] data.

ex: there's a population of male who likes diff. color bikes,

	Theory	Sample
Yellow	1/2	28
Red	1/2	70

(Theory Categorical Distribution) (Observed Categorical Distribution)

Goodness of fit test: wrt sample info. that we have collected, does it supports the theory? This is what basically we are going to do with the help of chi-square test.

ex: In 2010, census of the city → weights of individuals:

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

In 2020, weight of n = 500 individuals were sampled

<50 kg	50-75	>75
140	160	200

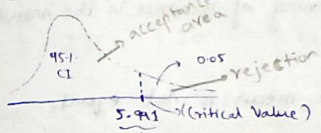
Using $\alpha = 0.05$, would you conclude the population differences of weights has changed in the last 10 years?

Sol: 2010 → expected, 2020 → observed

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

(expected)

1) Null hypo.: The data meets the expectation.
Alt. hypo.: The data does not meet the expectation.
2) $\alpha = 0.05$, CI = 95%.
3) Degree of freedom → dof = k − 1 = 3 − 1 = 2 ; {k = no. of categories (here k = 3)}
4) Decision Boundary: {Here in chi square test → we have right skewed graph}



(got from chi square table pdf)

If $\chi^2 > 5.99$ → reject null hypo.
else, we fail to reject null hypo.

5) Calculate Chi Square Test Statistics: $\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$

$\therefore \chi^2 = \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250} = 26.66$ → ∴ 26.66 > 5.99 → ∴ reject null hypo.

∴ Conclusion: The weights of 2020 population are different than those expected in 2010 pop.

ANOVA

It's a statistical method used to compare the means of 2 (or) more groups.

1) factors (variable) 2) Levels } ex: Mode of payment (factor) — GPAY, PAYTM } (levels)

(Analysis of Variance)

Assumptions in ANOVA

1) Normality of Sampling distribution of mean: The distribution of sample mean is normally distributed.
2) Absence of Outliers: outlying score needs to be removed from the dataset.
3) Homogenity of Variance: population variance in different levels of each independent variable are equal. {ex: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$}
4) Samples are independent & random.

Types of ANOVA:

1) One Way ANOVA: one factor with atleast 2 levels & these levels are independent

ex: Doctor wants to test a new medication to decrease headache. They split the participants in 3 conditions [10mg, 20mg, 30mg]. Doctor asks the participants to rate the headache [1-10]

Medication → (factor)

10mg	20mg	30mg
5	7	2
3	4	6

2) **Repeated Measures ANOVA**: one factor with atleast 2 levels & levels are dependent.

eg:

Levels →	Running → {Factor} Day 1	Day 2	Day 3	
	8	5	4	→ 1st person
	7	4	9	→ 2nd person

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

eg:

Gender → {Factor} (independent)	Running → {Factor} D1 (dependent)	D2	D3
Male	8	5	4
	9	4	3
Female	2	4	6
	7	8	3

(levels)

- **Hypothesis Testing in ANOVA**: {Partitioning of Variance in the ANOVA}

Null Hypo.: $H_0 \Rightarrow \mu_0 = \mu_1 = \mu_2 \cdots \mu_k$

Alt. Hypo.: $H_1 \Rightarrow$ atleast one of the sample mean is not equal.

Test Statistics → F Test = $\frac{\text{Variance between Samples}}{\text{Variance within Samples}}$; eg: (var. btw. samples)

	x_1	x_2	x_3
var. within samples	1	6	5
	2	7	6
	4	3	3

eg: Doctors want to test a new medication which reduces headache. They split the participant into 3 cond.'s [15mg, 30mg, 45mg]. Later on, the doctor asks the patient to rate the headache btw [1-10]. Are there any differences btw the 3 conditions using $\alpha = 0.05$?

Ans) ① Define Null & Alt. Hypo.? : $H_0: \mu_{15} = \mu_{30} = \mu_{45}$

H_1: not all μ are equal.

(Given)

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

② Significance: $\alpha = 0.05 \rightarrow$ CI = 0.95

③ Calc. degree of freedom: $N = 21$, $a = 3$, $n = 7$

(N: total sample size; a: total no. of categories (7×3); n: sample size)

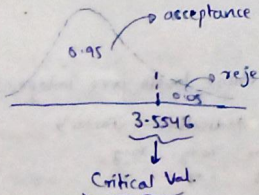
$df_{between} = a - 1 = 3 - 1 = 2$

$df_{within} = N - a = 21 - 3 = 18$ ($2 + 18 = df_{total}$)

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

∴ range: $(df_1, df_2) = (2, 18)$ → used in f-table → f table + α → critical value

④ Decision Boundary: (F-test follows right skewed distribution)



Decision Rule: If F is greater than 3.5546, reject the null hypo.

⑤ Calc. F test Statistics:

We need to fill the following table: (Sum of squares; mean squares)

	SS	df	MS	F
Between				
Within				
Total				

- $SS_{between} = \frac{\sum(\sum a_i)^2}{n} - \frac{T^2}{N}$

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

$= 98.67$

{15mg = 9+8+7+8+8+9+8 = 57
30mg = 7+6+6+7+8+7+6 = 47
45mg = 4+3+2+3+4+3+2 = 21

- $SS_{within} = \sum y^2 - \frac{\sum(\sum a_i)^2}{n}$; $\sum y^2$ = sum of all (values squared)

$= 9^2+8^2+7^2+8^2+8^2+9^2+8^2 + 7^2+6^2+6^2+7^2 + 8^2+7^2+6^2+4^2+3^2+2^2+3^2+4^2+3^2+2^2$

$= 853$

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

$= 10.29$

	SS	df	MS = (SS/df)
Between	98.67	2	49.34
within	10.29	18	0.54
Total	108.96	20	49.88

F → Var. btw samp. / Var. within samp. = $\frac{MS_{btw}}{MS_{within}}$

∴ F-test = $\frac{MS_{btw}}{MS_{within}} = \frac{49.34}{0.54} = 86.56$

∴ Since $86.56 > 3.5546$ → we reject H_0. ∴ Conclusion: There are differences btw those 3 conditions.

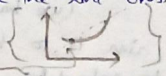
DIFFERENTIAL CALCULUS

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

Interpretation of Slope:

- (+)ve slope: if slope > 0, the line rises from L→R as it moves. Larger slope = steeper line
- (−)ve slope: if slope < 0, lines falls as it moves L→R. More (−)ve the slope = steeper line in downward dir".
- Zero slope: if slope = 0, line is horizontal → no vertical change as line moves from L→R.
- Undefined slope: line is vertical & slope is undefined.

Slope of eqn. of line: $y = mx + c$ (m: slope; c: intercept → point where the line crosses Y-axis.)



Slope: $\frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{f(x_0+h) - f(x_0)}{x_0 + h - x_0}$ → ∴ Slope of secant line = $\frac{f(x_0+h) - f(x_0)}{h}$

∴ Derivative of $f(x) = f'(x) = \lim_{h \to 0} \frac{f(x_0+h) - f(x_0)}{h} = \frac{d(f(x))}{dx}$

Power Rule: If $f(x) = x^n$ where $n \neq 0$ → $f'(x) = n \cdot x^{n-1}$

- $\frac{d}{dx}(c) = 0$ (constant)
- $\frac{d}{dx}(c \cdot f(x)) = c \cdot \frac{d}{dx}(f(x)) = c \cdot f'(x)$
- $\frac{d}{dx}(f(x) + g(x)) = \frac{df(x)}{dx} + \frac{d(g(x))}{dx}$

If $f(x) = \sin(x)$ → $f'(x) = \cos(x)$ & If $f(x) = \cos(x)$ → $f'(x) = -\sin(x)$

$f(x) = \ln(x)$ → $f'(x) = \frac{1}{x}$; $f(x) = e^x$ → $f'(x) = e^x$; $\frac{d}{dx}(h(x).f(x)) = h'(x)f(x) + h(x)f'(x)$ (Product Rule)

Chain Rule:

If $y = f(g(x))$ where $y = f(u)$ & $u = g(x)$, then $\frac{dy}{dx} = \frac{df}{du} \times \frac{du}{dx} = f'(g(x)).g'(x)$

ex: $y = \sqrt{\sin(3x)}$

Step 1: Identify the functions: Outer func. : $f(u) = \sqrt{u} = u^{1/2}$; middle func. : $g(v) = \sin v$; inner func. : $v = h(x) = 3x$

Step 2: $f'(u) = \frac{1}{2}(u)^{-1/2} = \frac{1}{2\sqrt{u}}$, $g'(v) = \cos v$, $h'(x) = 3$

Step 3: Apply the chain rule: $\frac{du}{dx} = \frac{df}{du} \times \frac{du}{dv} \times \frac{dv}{dx} = f'(g(h(x))) . g'(h(x)) . h'(x)$

$$\therefore \frac{du}{dx} = \frac{3\cos(3x)}{2\sqrt{\sin(3x)}}$$

- **Applications:**
  - → Backpropagation in neural networks (training DL Models)
  - → Gradient Descent Optimization
  - → Feature Engineering
  - → Regularization Techniques : Overfitting & Underfitting.

APPLICATION OF LINEAR ALGEBRA, STATS, & DIFFERENTIAL CALCULUS

i/p → (train) → Model → o/p

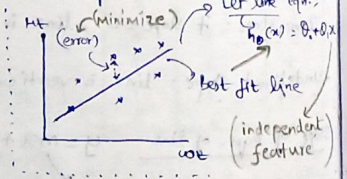
(simple linear regression) → (since we have only 1 i/p we call it 'simple'. If multiple i/p's → Multiple LR)

ex:

i/p: Wt	o/p → dependent feature: Ht.
74	170cm
80	180cm
75	175.5cm
⋮	⋮

θ_0 = Intercept ; θ_1 = slope

Let line eqn.: $h_\theta(x) = \theta_0 + \theta_1 x$



Cost function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

($h_\theta(x^{(i)})$ = predicted ; $y^{(i)}$ = true o/p ; their difference = error)

(We call this cost func. as Mean Squared Error)

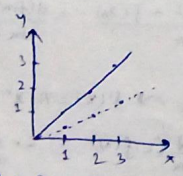
(we have to minimize this)

Our main aim is to basically come up with a best fit line wherein when I try to calculate cost func., it should be minimal.

ex: $h_\theta(x) = \theta_0 + \theta_1(x)$; let $\theta_0 = 0$ ∴ $h_\theta(x) = \theta_1(x)$: (passes via (0,0))

Dataset:

x	y
1	1
2	2
3	3



Let $\theta_1 = 1$ (line)
- $h_\theta(x) = 1$; $x = 1$
- $h_\theta(x) = 2$; $x = 2$
- $h_\theta(x) = 3$; $x = 3$

Let $\theta_1 = 0.5$ (dotted)
- $h_\theta(x) = 0.5$; $x = 1$
- $h_\theta(x) = 1$; $x = 2$
- $h_\theta(x) = 1.5$; $x = 3$

Now: $\theta_1 = 1$

$J(\theta_1) = \frac{1}{2\times3}[(1-1)^2 + (2-2)^2 + (3-3)^2]$

$= 0$

$\theta_1 = 0.5$

$J(\theta_1) = \frac{1}{2\times3}[(0.5-1)^2 + (1-2)^2 + (1.5-2)^2]$

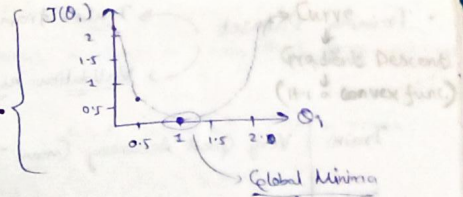
≈ 0.58

$\theta_1 = 0$

$J(\theta_1) = \frac{1}{2\times3}[(0-1)^2 + (0-2)^2 + (0-3)^2]$

≈ 2.3

(Graph when we plot $J(\theta_1)$ & θ_1 for diff. val.'s)



→ Global Minima

(Error has been minimized)

Convergence Algorithm: → optimize the changes of θ_1 value

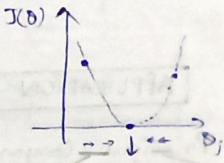
- Repeat until Convergence : { (we are going to change θ_j value)

$$\theta_j = \theta_j - \alpha\left[\frac{\partial}{\partial\theta_j} J(\theta_j)\right]$$

α = Learning rate

(we are calculating slope at that point)

It controls the speed at which the convergence should happen. If it's a very very small val., it will take more to converge. If it's a very big val., then it may continuously jump & may never converge. So select a smaller val. → ex: $\alpha = 0.001$



if we are on LHS we must go to RHS & vice versa

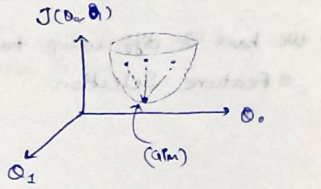
{ Refer Next Chap. → Linear Regression }

Multiple Linear Regression

ex: No. of rooms (independent features) | Size of House | Price (o/p feature)

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots$ { Multiple LR

where, θ_0 = Intercept ; $\theta_1, \theta_2, \theta_3, \dots \theta_k$ = co-efficient (represents)

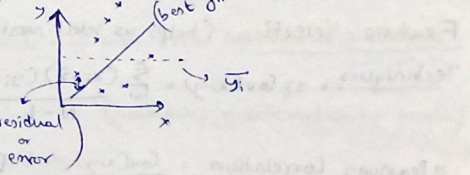


Performance Metrics in LR: (Linear Regression)

{ Refer Next chapter }

- **R-Squared** = $1 - \frac{SS_{res}}{SS_{total}}$ (sum of square)

$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$



≈ gives us accuracy.

- **Adjusted R-Squared:**

ex: Size of House | Price

$R^2 = 73\% \Rightarrow 0.73$

+ No. of bedrooms | Price

$R^2 = 85\% \Rightarrow 0.85$

+ Gender | Price

$R^2 = 87\% \Rightarrow 0.87$

∴ Even though we don't have a feature (~gender) that is directly correlated with the o/p feature, then also accuracy ↑'s. To prevent this we use adjusted R-square.

→ Adjusted R Square = $1 - \frac{(1-R^2)(N-1)}{N-p-1}$

where N = no. of data points, p = no. of independent features

- Training Dataset → Train: train the model
  → Validation: hyperparameter tuning of model.

Train	Very Good Accuracy (means "Low Bias")	Very Good Accuracy (Low Bias)
Test	Very Good Accuracy (means "Low Variance")	Bad Accuracy (High Variance)
	⇓	⇓
	Generalized Model	Model is overfitting (or) Model is underfitting

(Model is not trained well with training data) ← Model is underfitting

APPLICATION OF LINEAR ALGEBRA IN DIMENSIONALITY REDUCTION

Principal Component Analysis (PCA):

1) Curse of Dimensionality: Suppose: 500 features → (~ dimensions)

I have many ML models: M1 (3 features), Mn (6 features), M3 (15 feat.), M4 (50 feat.), M5 (200 feat.)

Acc.1 < Acc.2 < Acc.3 > Acc.4 > Acc.5

→ Model is overfeeded.
∴ Model performance degrades.

We have 2 diff. ways to remove CoD:
1) Feature Selection
2) Dimensionality Reduction (PCA) ⇓ Feature Extraction

- Why Dimensionality Reduction? → Prevent CoD.
  → Improve the performance of the model.
  → visualize & understand the data.

Feature Selection: (helps us select most imp. feat.)

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

+ → $x\uparrow = y\uparrow$ & $x\downarrow = y\downarrow$
0 → No relationship
− → $x\uparrow = y\downarrow$ & $x\downarrow = y\uparrow$

2) Pearson's Correlation $= \frac{Cov(x,y)}{\sigma_x * \sigma_y} = [-1,1]$ → The more towards +1, the more (+)ve correlated 'x' & 'y' are. & similar for (−)ve.

Feature Extraction:

ex: Room Size | No. of rooms | Price
(Transformation to extract new feature)
⇓
House Size | Price

2 feature → 1 feature
Dimensionality Reduction

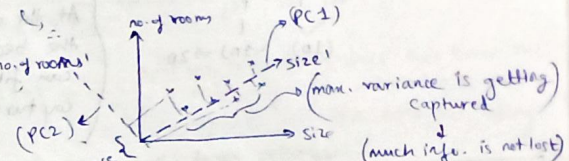
PCA Geometric Intuition:

↳ Goal of PCA: find out principal comp. in such a way that max. variance will be captured.

ex: No. of rooms; spread? ↑ variance ↑; Size

(Area b/w 1st & last datapoint = spread. If spread is huge, var. also increases.)

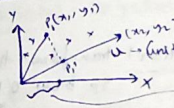
Converted 2D → 1D → but: loss of info. = no. of rooms



Since 2 dimension → so 2 PCs: PC1, PC2
↳ Var(PC1) > Var(PC2)

If 3 dimension: PC1, PC2, PC3 : Var(PC1) > Var(PC2) > Var(PC3)

Maths Intuition:



$Proj_{P_i} u = \frac{P_i \cdot u}{|u|} = P_i \cdot u$ (Scalar val.) → (distance from (0,0) to proj. of P1 on 'u')

∴ Max. Variance $= \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n}$ (Cost func.)

(Goal: find the best unit vector which captures max. variance)

Eigen Vectors & Eigen Values:

1) We need to find out covariance matrix b/w features.
2) Eigen vectors & Eigen Values will be found out from this covariance matrix.
3) Eigen Vector → eigen value (magnitude of eigen vector) → capture the max. var.

$[\ ] * [v] = \lambda * v$

([ ]: suppose we have a matrix; v: eigen vector) gives → max. magnitude → max eigen vector → best principal component

(already mathematically proven)

Steps to calculate eigen vectors & eigen values:

1) Covariance of features: x, y (2 independent features)

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

	x	y
x	Var(x)	Cov(x,y)
y	Cov(y,x)	Var(y)

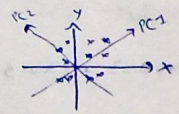
2) $A \cdot v = \lambda \cdot v$ → λ_1, λ_2 → eigen values ⇒ PC1, PC2

(Cov(x,x) = Var(x); Cov(y,y) = Var(y))

Revising:

Given:  ... 2D → 1D.

1) Standardize the data:



λ_1 → magn. of eigen vector

→ Covariance Matrix of X & Y : A =

	x	y
x	Var(x)	Cov(x,y)
y	Cov(y,x)	Var(y)

2×2

→ Find out Eigen vectors & value; $Av = \lambda v$ → λ_1, λ_2 → eigen values.

PC1 PC2

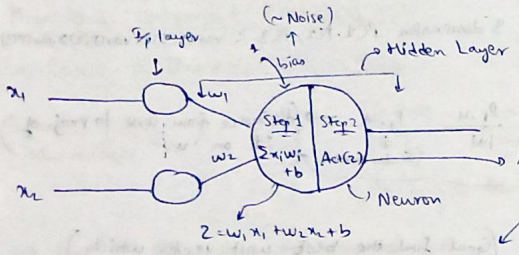
If 3P → 2D : $\lambda_1, \lambda_2, \lambda_3$ → PC1, PC2, PC3 → (1D) (1D) ⇒ 2D

At the end, we're trying to find out the best principal component that can fit on this data such that it captures max. variance.

{// Refer Next Chap → PCA}

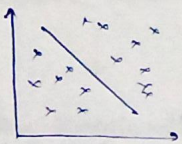
APPLICATION OF DERIVATIVES IN DEEP LEARNING NEURAL NETWORKS

Perceptron : (Artificial Neuron or Neural Network Unit)



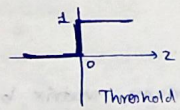
(* weights & bias are randomly initialized.)

Activation function : transforms the o/p btw 0 to 1 (or) -1 to 1.



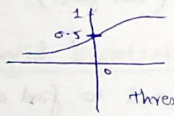
(Perceptron is a linear classifier)

Step func.



$\begin{cases} 0 ; z \le 0 \\ 1 ; z > 0 \end{cases}$

Sigmoid func.

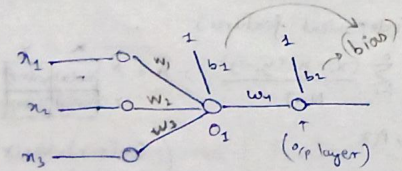


$\begin{cases} 1 ; z > 0.5 \\ 0 ; z \le 0.5 \end{cases}$

Perceptron Models

- Simple Layer Perceptron Model
- Multi Layered Perceptron Model (more than 1 neurons in hidden layers)
  - Forward & Backward Propagation
  - Loss func., Activation func., Optimizers

ex:



Backward Propagation

IQ	Study hrs.	Play hrs.	pass/fail o/p
95	4	4	1
100	5	2	1
95	2	7	0

Let : [0.01, 0.02, 0.03], [0.001]
(w_1, w_2, w_3, b)

Steps:

1) F/w prop.: Act(z) (In hidden Layer 1)

$z = 95 \times 0.01 + 4 \times 0.02 + 4 \times 0.03 + 1 \times 0.001$

$= 1.151$

2) Activation (z) : Sigmoid $= \frac{1}{1+e^{-z}}$ → $f(z) = \frac{1}{1+e^{-1.151}} = 0.759 = o_1$

Hidden Layer 2: (Let $w_4 = 0.02$, $b_2 = 0.03$)

↳ Step 1: $z = o_1 * w_4 + b_2 = 0.02 \times 0.759 + 0.03 = 0.045\,18$

↳ Step 2: Activation (z) $= \frac{1}{1+e^{(0.045\,18)}} = 0.511\,29 = o_2 = \hat{y}$

(actual o/p) $(1 - 0.511\,29) \approx 0.49$ → we have to reduce this error. The only way to minimize is to update all the weights using back propagation.

∴ Loss func. / Error =

- Regression: MSE, MAE, Huber Loss
- Classification: Binary Cross Entropy, Categorical Cross Entropy.

Weight Updation Formula: $w_{new} = w_{old} - \eta \left(\frac{\partial h}{\partial w_{old}}\right)$ * (slope)

(Learning rate) → decides the step size towards global minima i.e. how fast we want to reach global minima

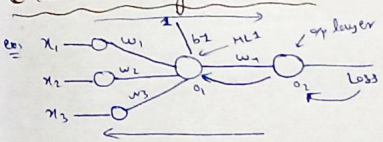
(helps) **Optimizers**: to reduce the loss value.

η ~ generally → 0.001 (initialized by small val.)

$w_{new} = w_{old}$ → 'w' reaches global minima.

Chain Rule of Derivatives:

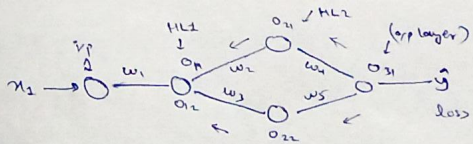
ex:



$w_{4new} = w_{4old} - \eta \left(\frac{\partial h}{\partial w_{4old}}\right)$

$\frac{\partial h}{\partial w_{4old}} = \frac{\partial h}{\partial o_2} * \frac{\partial o_2}{\partial w_{4old}} = \frac{\partial h}{\partial o_2} * \frac{\partial o_2}{\partial o_3} * \frac{\partial o_3}{\partial w_{1old}}$

ex:



$w_{1new} = w_{1old} - \eta \left(\frac{\partial h}{\partial w_{1old}}\right)$

$\frac{\partial h}{\partial w_{1old}} = \left[\frac{\partial h}{\partial o_{31}} * \frac{\partial o_{31}}{\partial o_{21}} * \frac{\partial o_{21}}{\partial o_{11}} * \frac{\partial o_{11}}{\partial w_{1old}}\right]$

$+$

$\left[\frac{\partial h}{\partial o_{31}} * \frac{\partial o_{31}}{\partial o_{22}} * \frac{\partial o_{22}}{\partial o_{12}} * \frac{\partial o_{12}}{\partial w_{1old}}\right]$

{// Refer Next Chap → Deep Learning}