Day - 4 Statistics

23/02/23

1. Central Limit Theorem
2. Probability
3. Permutation & Combination
4. Covariance, Pearson Correlation, spearman Rank correlation.
5. Bernoulli's Distribution
6. Binomial Distribution.
7. Power law (Pareto Distribution)

V.V.I

1. Central Limit Theorem :→

$\boxed{n \geq 30}$

Sample size ↓ $\boxed{n}$

no. of samples ↓ $\boxed{m}$ (the larger the value the better result)

Sampling with replacement

$S_1 = \{x_1, x_2, x_3 ---- x_n\} \rightarrow \bar{x}_1$ (1st sample)

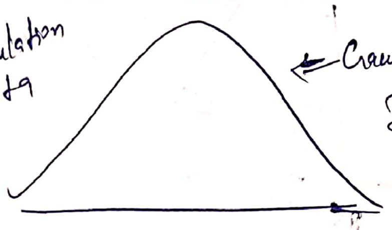
$S_2 = \{x_3, x_4, x_2 ---- x_n\} \rightarrow \bar{x}_2$ (2nd sample)

$S_3 = \{x_4, x_5, x_2 ---- x_n\} \rightarrow \bar{x}_3$

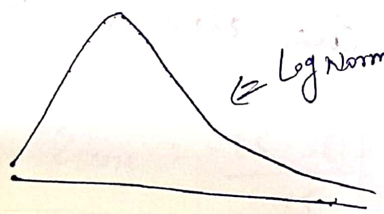
$S_m = \{ ---- \} = \bar{x}_m$

Population Data

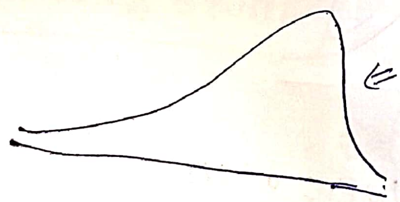
$\boxed{N}$



← Gaussian/Normal Distribution



← Log Normal Distr



← Left Skewed ⇒

Gaussian Distr. ⇒

Sample data can have different no. of elements.

Stmnt:→ Let's suppose take any Population Data (N) ~~Distr~~ either Gaussian or Log normal distr or left skewed. Once we select the sample data from any one of them of size i.e. $\boxed{n \geq 30}$ and $\boxed{\text{no. of samples is } m}$. Suppose we have taken diffn- diffn samples i.e $S_1, S_2 -- S_m$ and their mean is $\bar{x}_1, \bar{x}_2, \bar{x}_3 ---- \bar{x}_m$ then if once we plot the graph of their $\boxed{\text{mean}}$ ~~we get~~ using histogram we get Gaussian Distr.

This theorem only holds for Sa... it is fo...

Defn → The central limit theorem states that if you have a population with mean 'μ' and standard deviation 'σ' and take sufficiently large random samples (i.e. $n \geq 30$) from the population with replacement, then the distribution of the sample means will be approximately normally distributed. Proved.

Q Calculate the size of shark throughout the world?

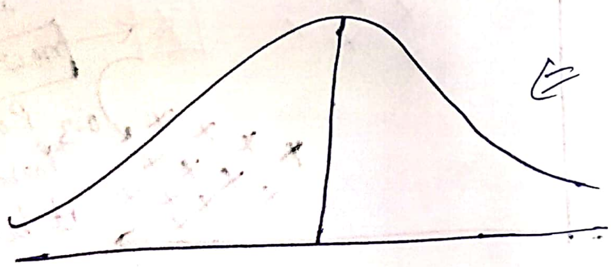
Sol $\boxed{n \geq 30}$ no. of Sample = 10 (the larger the value the better the result)
↑ i.e different regions.

Now find the size of shark from 10 different-different regions.

R_1 (Region 1) = $\{x_1, x_2, \ldots, x_n\}$ = $\bar{R_1}$

R_2 = $\{\ldots\}$ = $\bar{R_2}$

R_{10} = $\{\ldots\}$ = $\bar{R_{10}}$



⇐ Gaussian Distr.

Through this we can conclude many things like mean, median, mode etc.
Based on this lot of assumption can be made.

for $n < 30$ central limit theorem fails.

as $n \uparrow\uparrow\uparrow$ it gets more inclined towards Gaussian Distr.
(no. of samples increases).

→ Here with replacement means some of the values get repeated or mayn't get repeated in sample data.

② **Probablity** :→ Probablity is a measure of the likelihood of an event.

Eg:→ Tossing a fair Coin. $P(H) = 0.5$ $P(T) = 0.5$

⇓

SHOLAY movie → COIN $P(H) = 1$

⇓

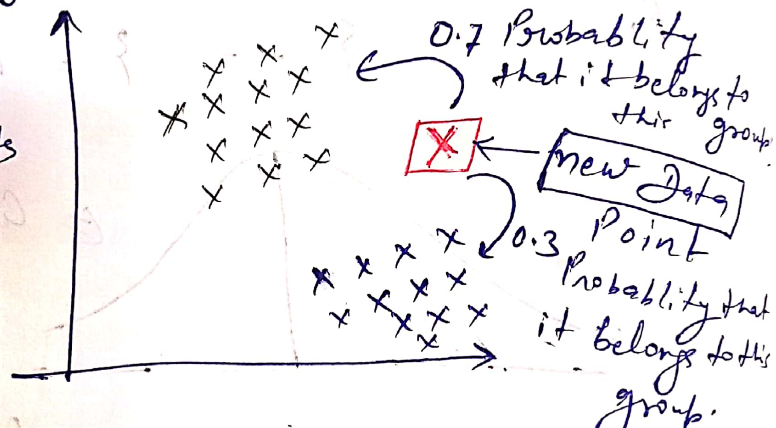
Unfair coin

Rolling a Dice :→

$P(1) = 1/6$, $P(2) = 1/6$, $P(3) = 1/6$

⇒ **Suppose Solving a classification Problem.**

How to decide that the new data Points belongs to which Category?

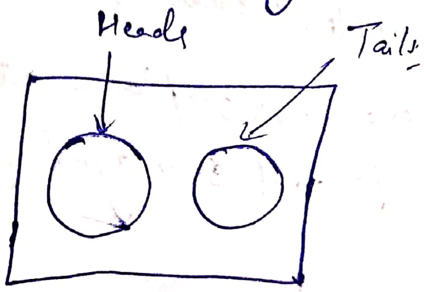


→ Whenever we try to Categorize a new data point that time Probablity Comes into picture.

⑤ **Mutual Exclusive Event** :→

Two events are mutually exclusive if they ~~don't~~ Can't occur at the Same time.

Eg:→ (i) Tossing a coin
(ii) winter & summer
(iii) Rolling a Dice

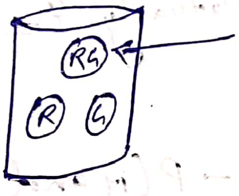


(2) Non-Mutual Exclusive Events :→

Two events can occur at the same time.

eg. (*) Picking randomly a card from a deck of card, two events "heart" and "king" can be selected.

(*)



chance of selecting Red & Green marble at the same time

Mutual Exclusive Problem Statement :→

Q. What is the probablity of coin landing heads or tails.

⇓

Sol^n

Addition rule of mutually exclusive events

$$\boxed{P(A \text{ or } B) = P(A) + P(B)}$$

$$= \frac{1}{2} + \frac{1}{2} = 1 \text{ Ans}$$

Q. What is the probability of getting a 1 or 6 or 3 while rolling a dice?

Sol^n

$$P(1 \text{ or } 3 \text{ or } 6) = P(1) + P(3) + P(6)$$

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} \text{ Ans}$$

Non Mutually Exclusive Events Problem Statement

Q. Bag of Marbles : 10 red, 6 green, 3 Red & Green

Q. When Picking randomly from a bag of marbles what is the probablity of choosing a marble that is red or green?

⇓

Sol:

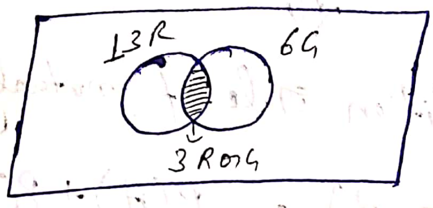
$\boxed{\text{Non Mutually Exclusive Event}}$

Addition rule for non-mutually Exclusive event

$$\boxed{P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)}$$

$$P(R \text{ or } G) = P(R) + P(G) - P(R \text{ or } G)$$

$$= \frac{10+3}{19} + \frac{6+3}{19} - \frac{3}{19} = \frac{19}{19} = \boxed{1} \text{ Ans}$$



Q. From the Deck of Cards : What is the probablity of choosing a $\boxed{♡ \text{ and Queen}}$ ⇒ $\boxed{\text{Non mutually Exclusive}}$

Sol:

$$P(♡ \text{ or Queen}) = P(♡) + P(\text{Queen}) - P(♡ \text{ and Queen})$$

$$= 13/52 + 4/52 - 1/52$$

$$= \boxed{16/52} \text{ Ans}$$

Multiplication Rule

(*) Dependents Events :→ Two events are dependent if they affect one another.

Eg:- Bag of Marble . { O O O O ; O O O } → Red Marble, → Black Marble.

Choose 1 Red Marble then 1 Black Marble.

With Replacement

$P(R) = 4/7$ ⟶ $P(B) = 3/6$

Q// What is the probability of rolling a "5" and then a "3" with a normal 6 sided dice. ⇒ {Independent Event, Dependent}

Sol// $P(1) = 1/6$, $P(2) = 1/6$, $P(3) = 1/6$, $P(4) = 1/6$

Independent Event → with Replacement

Multiplication Rule for Independent Event

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

$$= 1/6 * 1/6 = \boxed{1/36} \text{ Ans}$$

P(A or B) → Mutual Exclusive
→ Non Mutual Exclusive

$$\boxed{P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)}$$ Non Mutually Exclusive

$$\boxed{P(A \text{ or } B) = P(A) + P(B)}$$ Mutually Exclusive

Dependent Event :- → without Replacement

$\boxed{\begin{matrix} O\ O\ O\ O \\ O\ O\ O \end{matrix}}$ Dependent Events.

Probability of Drawing a "Red" and then drawing a "Black" marble from the Bag?

Sol

Probablity of Black Marble given Red already choosen

OOOO
OOO → $P(R) = \boxed{4/7}$ → OOO OOO → $P(B/R)$ → $\boxed{3/6}$

Red marble taken out

Conditional Probability

Conditional Probability

$$\boxed{P(R \text{ and } B) = P(R) * P(B/R)}$$

$$= \frac{\cancel{4}^{2}}{7} * \frac{3}{\cancel{6}_{3}}$$

$$= \boxed{2/7} \text{ Ans}$$

Naive Bayes

Permutation :→

School of Children

chocolates available : {Dairy milk, kit kat, Milky Bar, Sneakers, 5 star}

Q. How many ways can we ~~distribute~~ arrange the chocolate ~~among~~ at 3 ~~children~~ places?

Sol

5 * 4 * 3

{DM, KK, MB}, {KK DM, MB}, {} — All the Possible Arrangement

= 60 ways ⟹ Permutation

with Permutation Order matters.

Formula

$$\boxed{{}^{n}P_{r} = \frac{n!}{(n-r)!}} = \frac{5!}{(5-3)!} = \frac{5 \times 4 \times 3 \times \cancel{2!}}{\cancel{2!}}$$

= 60 Ways.

Where n = Total no. of Objects
r = no. of selection.

Hence we can arrange 5 choc at the 3 Places in 60 ways

Combination

~~Repeatation~~

$\boxed{\text{Repeation will not occur.}}$

{DM, KK, MB}
× {MB, KK, DM} ×

Both are same

Only the unique combination will be taken in consideration

$\boxed{\text{Unique Combination}}$

$$\boxed{{}^{n}C_{r} = \frac{n!}{r!\,(n-r)!}} = \frac{5!}{3!\,(5-3)!} = \frac{5 \times \overset{2}{\cancel{4}} \times \cancel{3!}}{\cancel{3!}\,\cancel{2} \times 1}$$

$= 10$ Ans

* Covariance

we use it in {Feature Selection}

X Age	Y Weight
12	40
	45
13	48
15	
17	60
18	62
$\bar{x} = 15$	$\bar{y} = 51$

$\boxed{\text{Age} \uparrow \ \text{weight} \uparrow \quad \text{Age} \downarrow \ \text{weight} \downarrow}$

Quantify the relationship only using mathematical question.

$$\boxed{Cov(x, y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}}$$

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$Cov(x, x) \Leftarrow \boxed{s^2 = \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{n-1}}$$

Relationship b/w Covariance and Variance.

$\boxed{Cov(x,x) = Var(x)} \Leftarrow$

~~Covariance~~ = 0

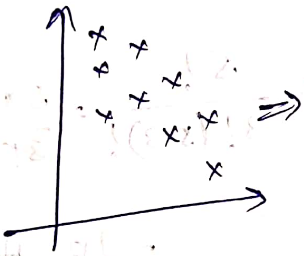
+ve Covariance

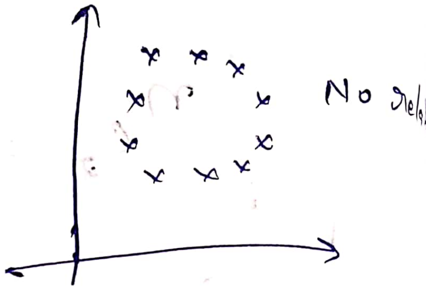
$\boxed{\begin{matrix} x\uparrow & y\uparrow \\ x\downarrow & y\downarrow \end{matrix}}$

-ve Covariance

$\boxed{\begin{matrix} x\uparrow & y\downarrow \\ x\downarrow & y\uparrow \end{matrix}}$

Covariance = 0





$\cancel{Cov(x,y) = (12-15)(40-51) + (13-15)(45-51)}$

$$\bar{x} = \frac{12+13+15+17+18}{5} = \frac{75}{5} = \boxed{15}$$

$$\bar{y} = \frac{40+45+48+60+62}{5} = \frac{255}{5} = \boxed{51}$$

$$Cov(x,y) = \frac{(12-15)(40-51)+(13-15)(45-51)+(15-15)(48-51)+(17-15)*(60-51)+(18-15)(62-51)}{(5-1)}$$

$$= \frac{(-3*-11)+(-2*-6)+(0*3)+(2*9)+(3*11)}{4}$$

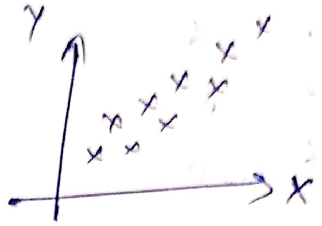
$$= \cancel{\frac{-33-12+0+18+33}{4}} = \cancel{\frac{6}{4}} = \cancel{\frac{3}{2}}$$

$$= \frac{33+12+0+18+33}{4} = \frac{96}{4} = \boxed{24}$$

24 is a +value. Since value of Cov(x,y) is +ve so from there we can conclude that

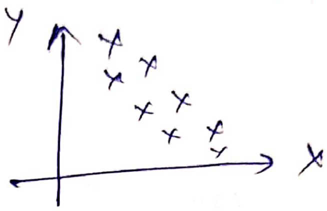
+ve Covariance

X↑ Y↑
X↓ Y↓



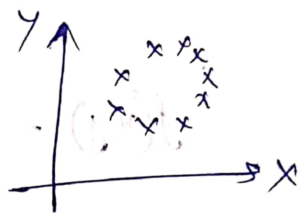
-ve covariance

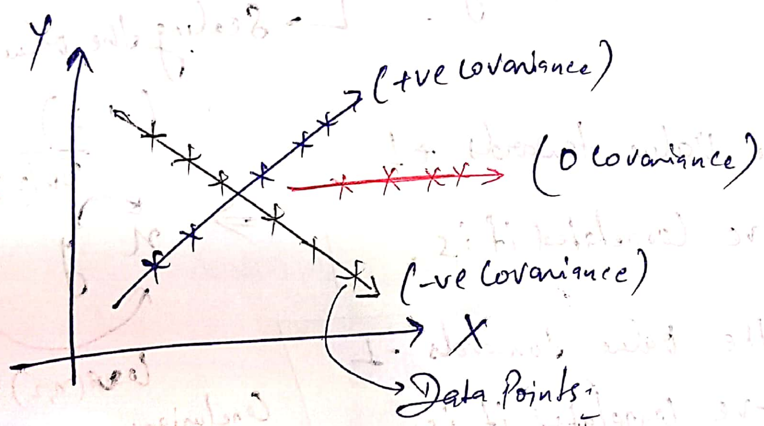
X↑ Y↓
X↓ Y↓



Covariance = 0

No relation b/w x & y





Example for -ve Covariance:→

(1)

X	Y
10	4
8	7
7	8
6	10

$\bar{X} = 7.75 \quad \bar{Y} = 7.75$

$$Cov(x,y) = \frac{(10-7.75)(4-7.75)+(8-7.75)(6-7.75)+(7-7.75)(8-7.75)+(6-7.75)(10-7.75)}{(4-1)}$$

$$= \frac{(2.25 * -3.75)+(0.25 * -1.75)+(-0.75 * 0.25)+(-1.75 * 2.25)}{4}$$

$$\bar{X} = \frac{10+8+7+6}{4} = \frac{31}{4} = 7.75$$

$$\bar{Y} = 7.75$$

$= -2.25 + (-0.4375) + (-0.1875) + (-3.9375)$

$= -6.8125$ (-ve covariance)

(ii)

X	Y
10	4
8	6
7	8
6	10
$\bar{x} = 7.75$	$\bar{y} = 7$

$$Cov(x,y) = \frac{(10-7.75)(4-7)+(8-7.75)(6-7)+(7-7.75)(8-7)+(6-7.75)(10-7)}{3}$$

$$= \boxed{-3.25} \rightarrow \boxed{-ve \text{ Covariance}}$$

Pearson Correlation Coefficient :→ They try to restrict the value in b/w (-1, 1).

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

→ Scaling the value in b/w (-1, 1)

More the value towards +1 More +ve Correlated it is.

More the value towards -1 More -ve Correlated it is.

⇒ $x \quad y \quad z$

$Cov(x,y) = 0.5$

$Cov(x,z) = 0.7$

Conclusion:- x is more Correlated towards z.

Difference b/w Covariance & Correlation.

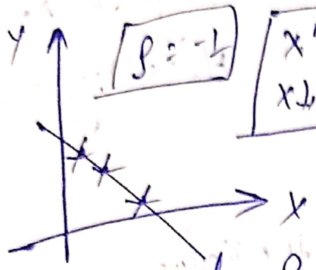
In Case of Covariance there is no range. We Can't Say how much +ve or -ve it will be. But in Case of Correlation the value will be in the range of (-1, 1) so that we can say how much +ve or -ve it is.

By using Pearson Correlation Coefficient we are restricting the value b/w (-1, 1) by dividing it [illegible]

Examples

$\rho = -1$

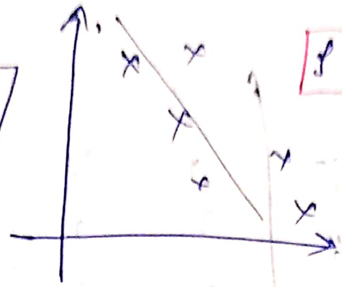
$x\uparrow\ y\downarrow$
$x\downarrow\ y\uparrow$



Since all the data Points are falling on the same line & it is in -ve dirn hence the Value of Pearson Correlation Coefficient $\rho = -1$.

$\rho = -1$ to 0

$-1 \geq \rho > 0$



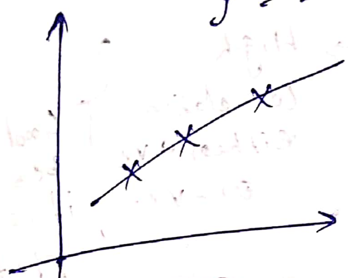
$\rho = (-1$ to $0)$

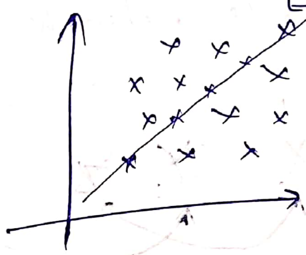
Since all the data Points Can't fall on Same line.

$\rho = 0$ to 1

$0 < \rho \leq +1$

$\rho = 1$





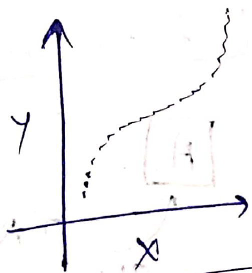
$\rho = 0$

No relationship.

Pearson Correlation holds good only for linear data.

Spearman Rank Correlation :→ We Use Spearman Rank Correlation mainly for non-Linear Data.

Basically here we Calculate the Rank



Formula

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

Pearson Correlation = 0.88 ← for linear Data
Spearman Correlation = 1 ← for non-linear Data.

How to Calculate Rank?

X	Y	R(X)	R(Y)
10	4	4	1
8	6	3	2
7	8	2	3
6	10	1	4

We Assign the Rank in Ascending order.

Higher the Value Higher the Rank.

lower the value lower the Rank.

Rank table Use this for our Calculation

Why this Correlation will be Used?

Sol:

X ~~X~~ Z O/P

High Correlation either +ve or -ve → Good bcz it is important feature to Predict the O/P

Corr(X,Z) = 0.95 it means both the features are almost same and we can delete one of them. Since they are 95% correlated Using One of them will be enough to get the 'O/P'

low Correlation. suppose 0.01 then ignore this feature

Practicle Example: →

~~A~~ : Exp Degree City Salary

+ve Correlated

96%. Since A and Exp is 96% correlated So we Can drop 'A'. we are not losing data by Dropping 'A'.

no Correlation

+ve correlated

+ve or -ve Correlation