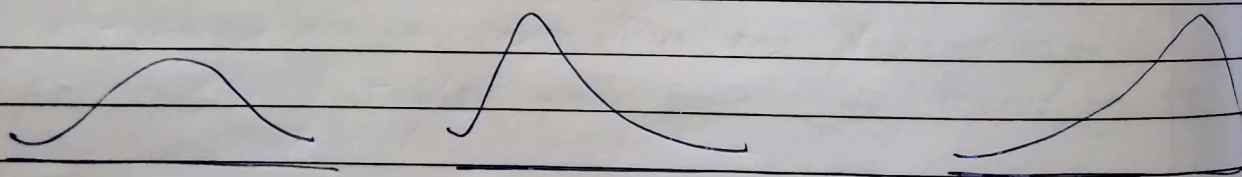
Day 4 stats

(1) Central limit Theorem
(2) Probablity
(3) Permutation and combination
(4) Covariance, Pearson correlation, Spearman Rank correlation
(5) Bernoulli distribution
(6) Binomial distribution
(7) Power law & Pareto Distribution)

V-V-Imp:

Central Limit Theorem:

→ For any kind of distribution
→ Distribution can be



Gaussian / normal Dist.

log normal dist

left + skewed

or some other distribution

Population data – N
, sample data (n)

No.	Sample data		Means
S_1 (first)	$\{x_1, x_2, x_3 \ldots x_n\}$	→	$\bar{x}_1$

1st sample $\{x_1, x_2 \cdots x_4 \cdots x_n\} \longrightarrow \bar{x}_2$

2nd $\{x_2, x_3 \cdots x_n\} \longrightarrow \bar{x}_3$

3rd $\{x_1, x_3 \cdots x_{n-1}\} \longrightarrow \bar{x}_4$

⋮

like this we will take m no. of samples

$S_m \{x_3 \cdots x_7\} \longrightarrow \bar{x}_m$

Here we have to take sample size (n) greater than 30

$$\boxed{n \geq 30}$$

⇒ For any kind of distribution where population data is N,

⇒ If we take m no. of samples where sample size is ≥ 30 and if we take the mean of all the samples ~~mean~~ and plot it as a histogram then it will be gaussian ~~flat~~ / Normally distributed

→ If $n < 30$, then central limit Theorem will not be applicable

$n \geq 30$ is scientifically framed

Page No.:
Date:

Q) Size of shark throughout the world ?

Soln: Take sample of size ≥ 30.
Let us take 10 no. of samples from different regions
Then find the mean of all the samples.
We will get Gaussian distribution and we can do many assumptions

★ More no of m will give more accurate results

Probability

Probability is a measure of the likelihood of an event

Eg - Tossing of a coin (unbiased)
Head - 1/2
Tail - 1/2

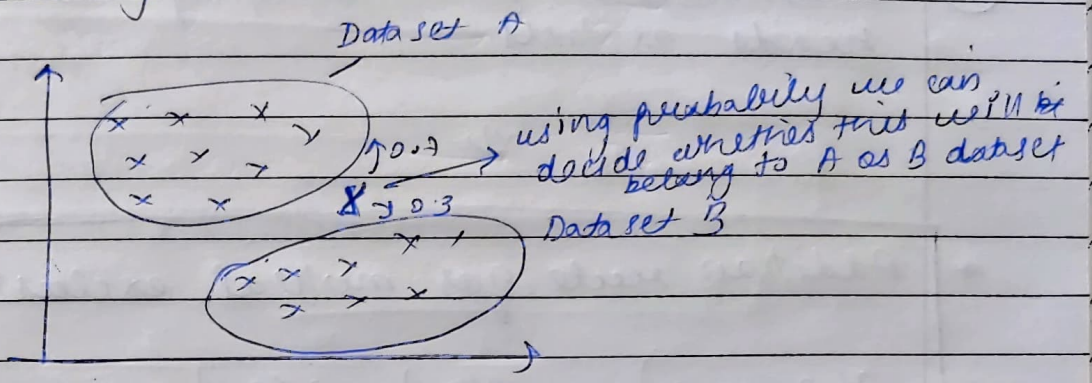
- Rolling of a dice

P(1) = 1/6 (Prob. of getting 1)

P(6) = 1/6

→ In some companies aptitude questions on this will be asked (only for freshers)

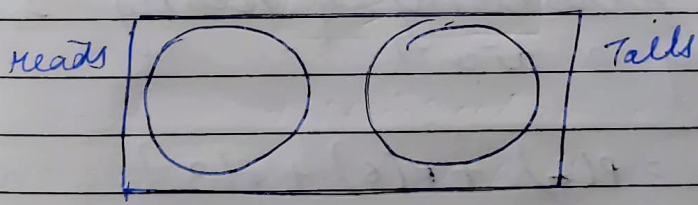
Use of Probability



① Mutually Exclusive

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

Ex - Tossing a coin



→ Rolling of a dice

② Non mutual exclusive Events

Two events can occur at the same time

→ Picking two red card from a deck of card.

Page No.:
Date:

Mutual Exclusive Ques :-

(Q) what is the probability of coin landing on heads or tails

∨

* Addition rule for mutual exclusive events

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

$$P(\text{Heads or Tails}) = P(\text{Heads}) + P(\text{Tail})$$
$$= \frac{1}{2} + \frac{1}{2}$$
$$= 1$$

Q) what is the probablity of getting 1 or 6 or 3 while rolling a dice ?

$\frac{1}{6}\ \frac{1}{1}\ \frac{1}{1}$

$\frac{3}{1}$

$$P(1 \text{ or } 6 \text{ or } 3) = P(1) + P(6) + P(3)$$
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$
$$= \frac{1}{2}$$

Page No.:
Date:
YOUVA

Non mutual Exclusive Event

a) From a bag of marbles
10 - Red
6 - Green
3 - Red & Green

b) When picking randomly from a bag of marbles. what is the probability of choosing a marble that is red or green

$\frac{10+6}{19}$; $\frac{16}{19}$; $\frac{11}{19}$

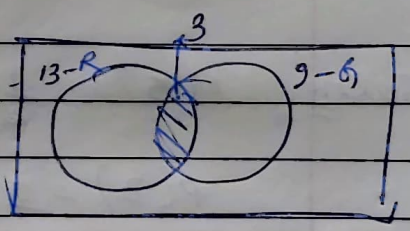
* Addition rule for non mutual ~~exclusive~~ exclusive events

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

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

$$= \frac{10}{19} + \frac{6}{19} - \frac{3}{19}$$

$$= \frac{13}{19} + \frac{2}{19} - \frac{3}{19} = \frac{19}{19} = 1 \text{ (Ans)}$$



Page No.:
Date:

(Q) From a deck of cards what is the probability of choosing Heart or queen

Soln: Heart - 13
Queen - 4

$$P(H \mid Q) = P(H) + P(Q) - P(H \& Q)$$
$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52}$$
$$= \frac{16}{52}$$

* Multiplication Rule

* Dependent Events :- Two events are dependent if they affect one another

Q) What is the probability of rolling a "5" and then a "3" with a "6 sided die"

Soln: $P(5) = \frac{1}{6}$

$P(3) = \frac{1}{6}$

(Independent Event)

$P(\cdot) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

Page No.:
Date:

Multiplication rule for independent events

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

$= \frac{2}{5} * \frac{1}{6} = \frac{1}{3}$

a) From a bag of 4 orange and 3 yellow marbles.

i) Probability of drawing a "orange" and then drawing a "yellow" marble from the bag?

$P(\text{orange}) = \frac{4C_1}{7C_1} = \frac{4}{7}$

~~P(yellow)~~

$P(\text{yellow}) = \frac{3C_1}{6C_1} = \frac{3}{6}$

~~P(Y/O)~~ $= \frac{4}{7} * \frac{3}{6_2} = \frac{2}{7}$ (Ans)

$P(O) * P(Y/O)$

↑

conditional Probability

→ Conditional Probability will be used in ML for Naive Bayes.

Permutation

→ With permutation, order matters
→ Possible arrangements

$$nP_n = \frac{n!}{(n-n)!}$$

n = Total no of objects
n = no of selections

Ex
Chocolate = { Dairy Milk, Kit Kat, Sneakers, 5 stars, Milkybar }

If we have do do selection of 3 chocolates from the set then no of possibilities will be

$5 \times 4 \times 3 = 60$

{ DM, KK, 5S }
{ DM, 5S, KK }

$$5P_3 = \frac{5!}{(5-3)!} = \frac{5!}{2!} = \frac{5 \times 4 \times 3 \times \cancel{2!}}{\cancel{2!}} = 60$$

Combination

→ Unique combination
→ No repeated values.

{DM, KK, SS}
{DM, SS, KK} X

$$nC_r = \frac{n!}{r!(n-r)!}$$

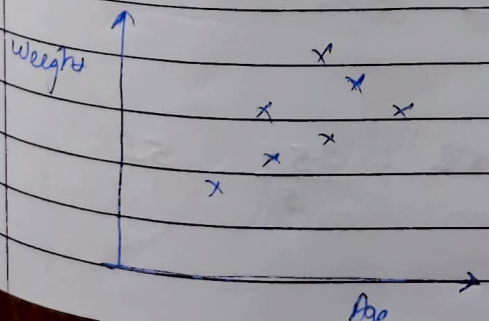
$$5C_3 = \frac{5!}{3!\,2!} = \frac{5 \times 4 \times 3!}{3! \times 2 \times 1} = 10$$

Covariance

→ Feature selection is done using this

Age	weight
12	40
13	45
15	48
17	60
18	62

conclusions Age ↑ weight ↑
Age ↓ weight ↓



Page No.:
Date:
YOUVA

Formula:

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

Variance of x

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

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

$\Downarrow$

$\sigma(x,x)$

*** Interview Ques:

$\therefore$ Cov(X, X) = Variance of X

Q) From the previous dataset of ~~Dog~~ Age and weight. Quantify the relationship X & Y using mathematical question

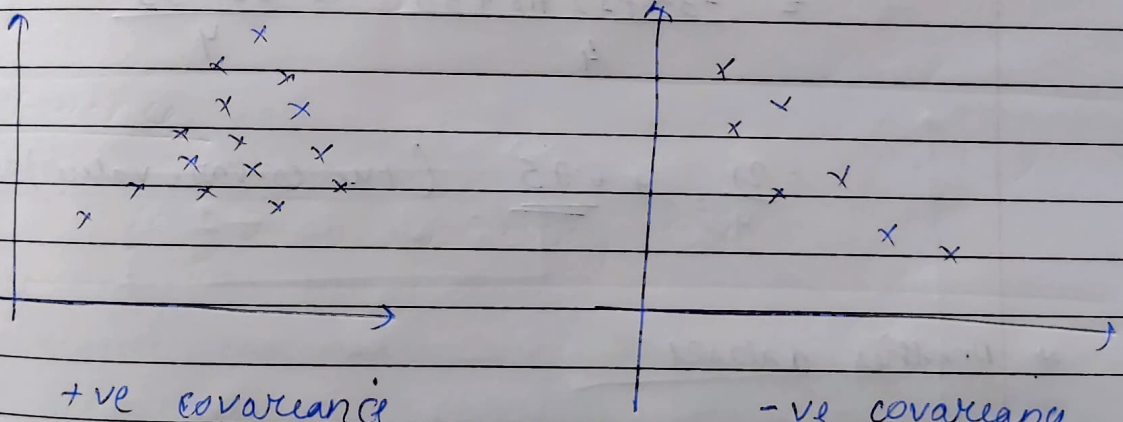
$$\bar{x} = \frac{12+13+15+17+18}{5} = 15$$

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

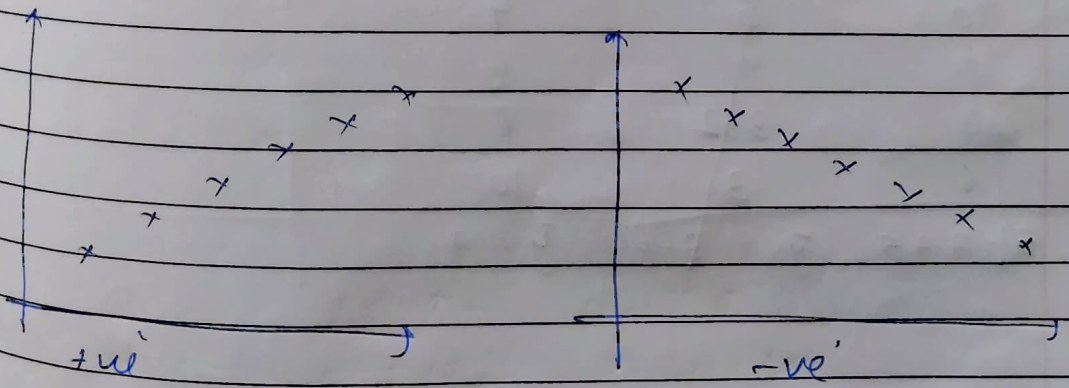
+ve covariance X ↑ Y ↑
X ↓ Y ↓

-ve covariance X ↑ Y ↓
X ↓ Y ↑

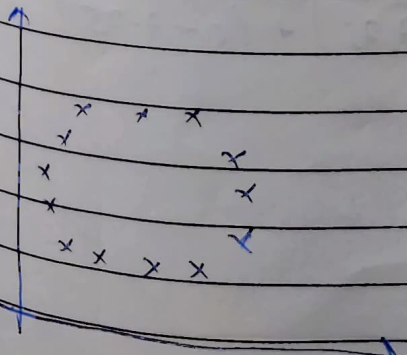
0 covariance → No relation b/w X & Y



+ve covariance -ve covariance



+ve -ve



Zero covariance

~~Cov(x, y) = (12-15)+(13-15)~~

$$Cov(x,y) = \frac{(12-15)(40-51)+(13-15)(45-51)+(17-15)(60-51)+(18-15)(61-51)}{5-1}$$

$$Cov(x,y) = \frac{(-3)(-11)+(-2)(-6)+(2)(9)+(3)(10)}{4}$$

$$= \frac{-33+12+18+30}{4} = \frac{60-33}{4}$$

$$= \frac{27}{4} = +6.75 \quad (+ve \text{ co-var. value})$$

* Dnatries dataset

x	y	y
10	5	4
8	4	6
7	3.2	8
6	3	10

$\bar{x} = 7.75$ $\bar{y} = 3.8$ $\bar{y} = 7$

$$(10-7.75)+(8-7.75)+(7-7.75)+(6-7.75)$$

~~(10-75~~

$$\frac{(10-7.75)(4-7) + (8-7.75)(6-7) + (7-7.75)(8-7)}{(6-7.75)(10-7)}$$

$$\frac{3}{}$$

$$-3.25$$

value will be negative. can be observed from eqn above

Pearson correlation coefficient (-1 to 1)

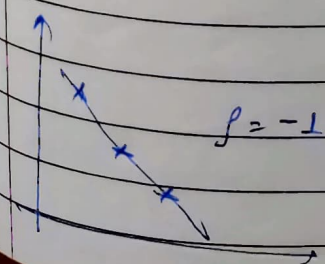
→ Holds only for linear data

→ Covariance can have any +ve or -ve values but by using this Pearson we can define the range to -1 to +1.

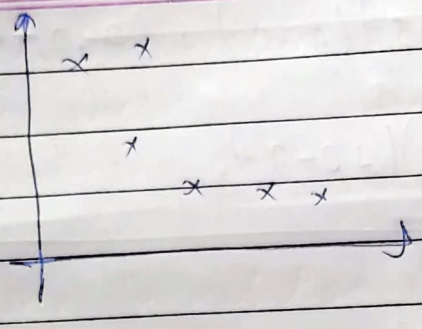
$$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

positive

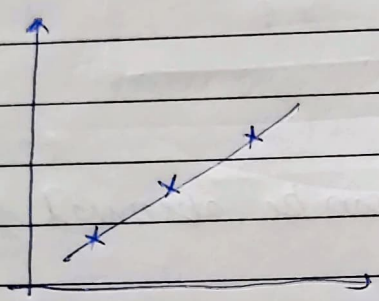
* more value towards +1 ⇒ more, correlated it is
" " " -1 ⇒ more negative correlated it is



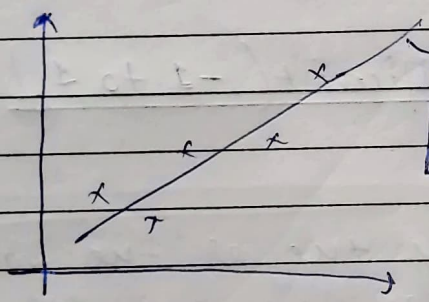
$\rho = -1$ (since all points are in a st. line)



ρ $\boxed{-1 \text{ to } 0}$

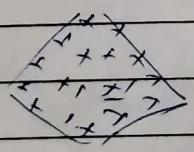


$\boxed{\rho = +1}$

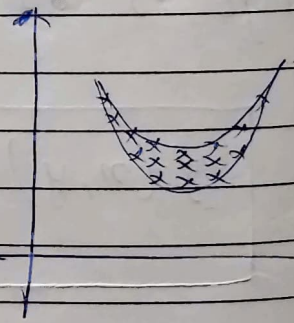


Best fit line

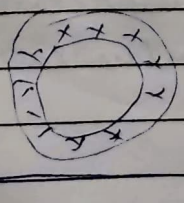
$\boxed{\rho : 0 \text{ to } +1}$



$\boxed{\rho = 0}$



$\rho = 0$

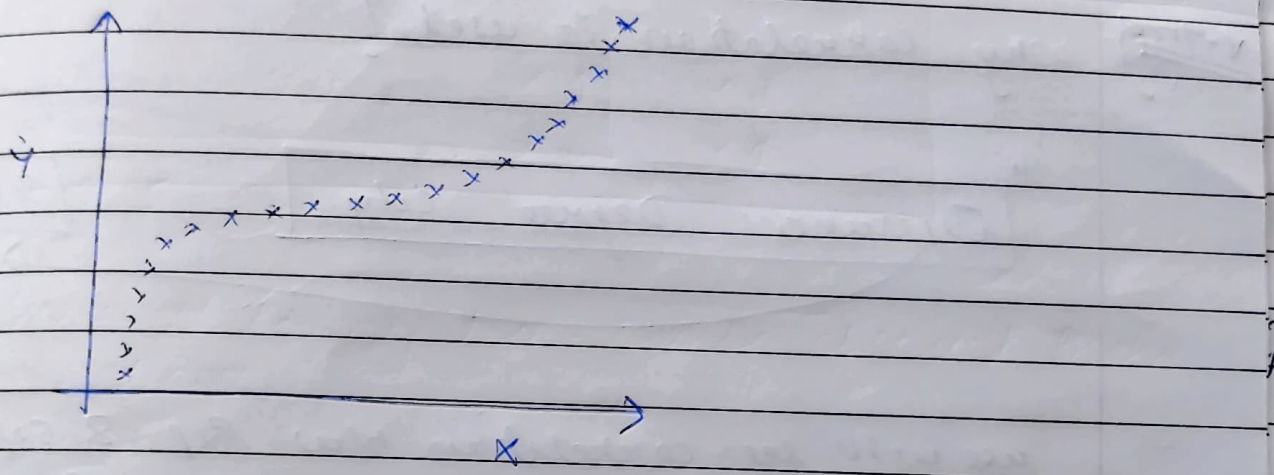


$\rho = 0$

Page No.:
Date:

Spearman Rank correlation

For non linear data



~~$\rho_s = \text{Cov}$~~

$$\rho_s = \frac{\text{Cov}(R(X), R(Y))}{\sigma_{R(X)}\, \sigma_{R(Y)}}$$

X	Y	R(X)	R(Y)
10	4	4	1
8	6	3	2
7	8	2	3
6	16	1	4

Rank is found in ascending order

After finding out the rank we will use only R(X) & R(Y)

→ If two values are same then same rank will be given

(2) why correlation is used?

i/p

Experience	Degree	city

o/p

Salary

→

we will see correlation b/w Exp & salary
↓
this can be +ve or -ve

Degree & salary

city & salary

and lets say we have one more input parameter A which gives same correlation as Experience, then we can drop A parameter

X

A

Exp. Degree city salary

Drop this

np.cov(df['total_bill'], df['tip'])

o/p:- array([[79.25 8.32]
[8.32 3.3]])

covariance of total bill & tip

variance of total bill

variance of tip