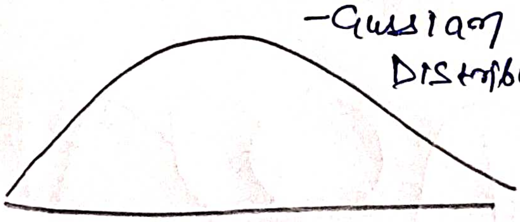
DAY-4 Statistics
Data Science

DAY-4

Agenda

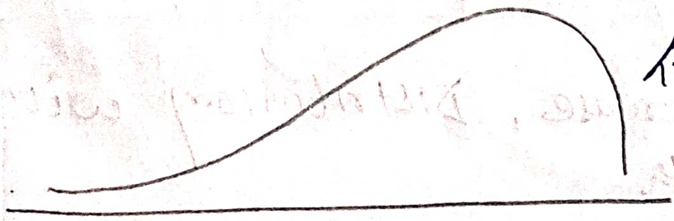
1. Central Limit theoram
2. Probability
3. Permutation & combination
4. Co-variance, Pearson correlation & Spearman Rank correlation

1. Central Limit theoram



Gussian Distribution

Log Normal Distribution



Left Skewed Distribution

→ Lets take m, no. of Samples of n size

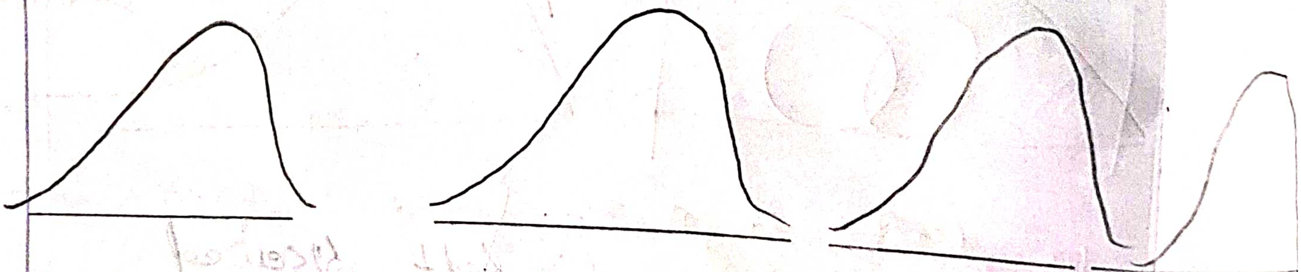
$S_1 = \{x, x_{10}, x_{11}, \ldots, x_n\} = \bar{x}_1 = \bar{S}_1$

$S_2 = \{x_3, x_4, x_5, \ldots, x_n\} - \bar{x}_2 \Rightarrow \bar{S}_2$

Suppose we have population Data which have Gaussian Distribution or Normal Distribution or Any kind of Data :-

Central Limit theoram states that if you have a Population with μ and σ and take sufficiently large random sample from Population where $n \geq 30$ & n no. of sample & if we take all the mean of sample and plot in histogram the we will get Guassian Distribution.

* the more number of n will be there, the better the Result are.



⇒ As n will increase, Distribution will become smooth

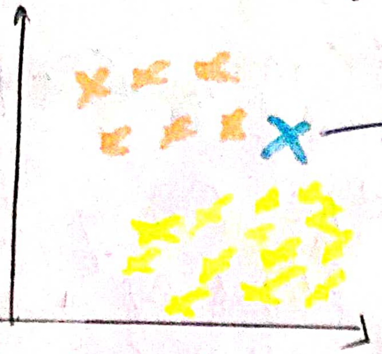
Probability

Probability is the Measure of likelihood of an event.

① Probability of coming head in coinflip $= 1/2$

② Prob of getting 1 in sixface dice $= 1/6$

Practical usage

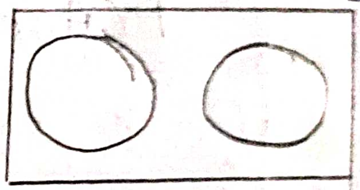


→ with the help of Probability, we will define x belongs to Which collection.

① Mutual Exclusive event

To event are mutually exclusive, if they can't occur at same time.

\# Tossing coin ② Rolling dice



② NON Mutual Exclusive event

Two events can occur same time

\# Picking a random card from deck, here king & heart come at same time.

Mutual exclusive Event

① What is the Probability of coin lending on head or tails?

$\Rightarrow P(A \text{ or } B) = P(A) + P(B) = 1/2 + 1/2 = \underline{1}$

② What is the Probability of getting 1 or 6 or 3 on Rolling a dice?

$\Rightarrow P(A \text{ or } B \text{ or } C) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$

~~Non Mutual exclusive event~~

Bag of Marbels, 10 Red, 6 Green, 3 Red & Green

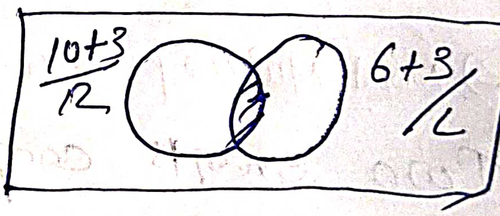
1) When picking Randomly from bag of marbles what is the prob of choosing marble that is Red or Green?

Addition Rule of Non Mutual

~~$13/19 + 9$~~

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

$$\Rightarrow \frac{13}{19} + \frac{9}{19} - \frac{3}{19} = 1$$



2) From deck of cards, Probability of choosing heart or queen at same time?

$$\Rightarrow P(\heartsuit \text{ or } Q) = P(\heartsuit) + P(Q) - P(\heartsuit \,\&\, Q)$$

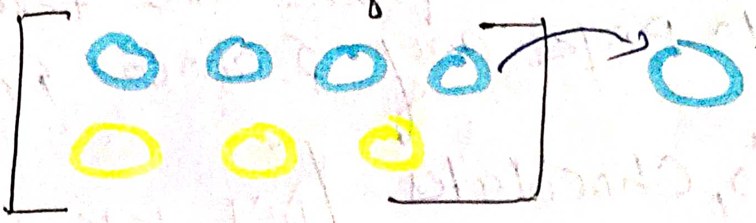
$$\Rightarrow \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}$$

Multiplication Rule

1] Dependent event :→

Two events are dependent if they effect one another

Conditional Probability

colour = 

→ Probability of taking out 1 blue ball = $4/7$

* Now we took one ball out & that Ball is blue & now Probility of taking one Green ball out is = $\boxed{3/6}$

Indipendent Event

⇒ Those events which doesnot effect another one

Q) Probability of rolling a '5' & the '3' in dice

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

$P(A \& B) = 1/6 \times 1/6 = 1/36$

↓

Mean Multiplication

$P(A \& B) = P(A) \times P(B)$

Permutation

Suppose, we are having 5 different kinds of chocolates & we are telling some child to pick any 3 choco from that

So, Chocolates = [DM, KK, MB, SNK, 5 Star]

3 chocolates can be picked

_____ x _____ x _____

⇒ they can choose any like

[KiK, MB, SNK] or [SNK, 5Star, DM]

So, many ways to arrange

Now.

$\underline{5}$ $\underline{4}$ $\underline{3}$

1st Position - Any 5 can be choosen

2nd Position = One got deduct, 4 out of 5 can be choosen

3rd Position = two got deducted, 3 out of 5 can be choosen

there will be = $5 \times 4 \times 3$ = 60 ways

It is a Mathematical technique that determines the numbers of possible arrangement in a set when order of the arrangement Matters

Formula:-

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

n = Total No. of object
r = no. of selection

$$= \frac{5!}{(5-3)!} = \frac{120}{2} = 60$$

COMBINATION

⇒ Combination is the way of selecting object in such a way that order of object is not repetative

[DM, 5star, KK] ⟶ [KK, 5star, DM] — X

element are Repatitive not consider this

$$^{n}C_{r} = \frac{n!}{r!\,(n-r)!} = \frac{5!}{2!\,3!} = 2$$

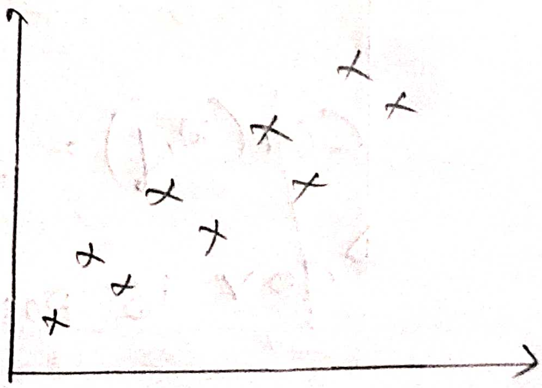
Only unique combination should be there

Co-variance

[Feature Selection]

Age	weight
12	40
13	45
15	48
17	50
18	52

A↑ W↑
A↓ W↓



Co-variance Measure the direction of relationship of two variables

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

$$\boxed{cov(x,x) = var(x)}$$

$$\text{variance}(\sigma^2) = \Sigma\left(\frac{x_i - \bar{x}}{n-1}\right)^2 = \frac{\Sigma (x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

$\hookrightarrow cov(x,x)$

Question

Age	weight
12	40
13	45
15	48
17	60
18	62
75	255

$\bar{x} = 75/5 = 15$

$\bar{y} = 255/5 = 51$

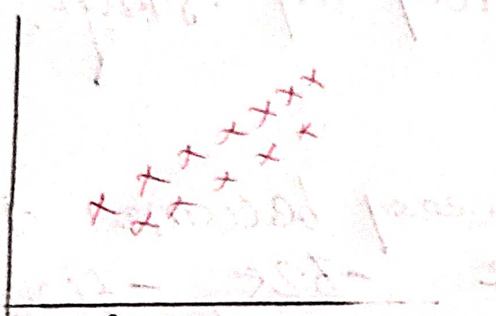
$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
-3	-11	+33
-2	-6	+12
0	-3	0
2	9	18
3	11	33
		96

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) \times (y_i - \bar{y})}{n-1} = 96/4 = 24$$

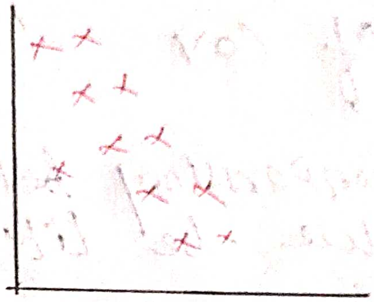
$\Rightarrow$ cov is positive $\boxed{x\uparrow\ y\uparrow \\ x\downarrow\ y\downarrow}$

When cor is negative $\boxed{\begin{matrix} x\uparrow & y\downarrow \\ x\downarrow & y\uparrow \end{matrix}}$

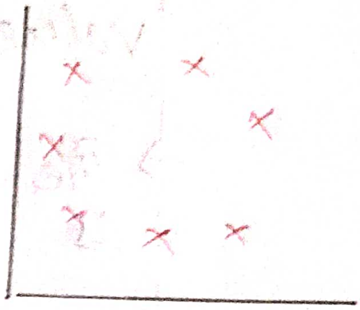
When cor is 0 No Relation between x & y



Positive Co-variance



Negative Co-variance



No-Relation

(2)

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
10	4	2.25	-3	6.75
8	6	1.75	-1	-1.75
7	8	0.75	1	0.75
6	10	-1.75	3	-4.25

$\boxed{\bar{x} = 7.75 \quad \bar{y} = 7}$

cov = -3.25

Negative Relation

Pearson Correlation Coefficient

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}} =$$

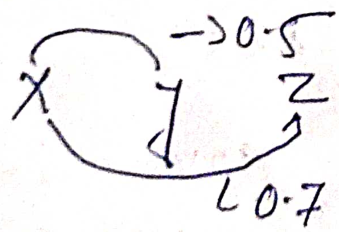
$$\rho(x,y) = \frac{\text{cov}(x,y)}{\sqrt{x}\,\sqrt{y}}$$

→ When we find co-variance?
There is no restriction in range of value of cov

→ The comparision ~~between~~ become difficult as value be like −3.25, −300.25, 100.25 etc.

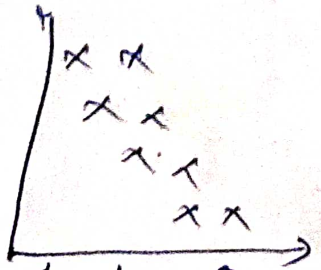
→ So, to restrict the number between range −1 to 1, we divide cov. by $\sqrt{x} \cdot \sqrt{y}$

Suppose

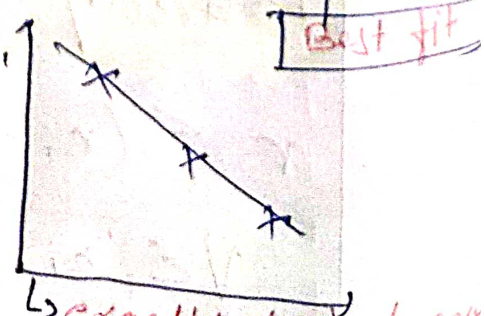


, 0.7 more co-related

→ More the value towards +1, more positive correlated it will be & vice-versa.



−1 to 0
Because Data Points are not on same line



↳ exactly (−1) because Data Points are on same line

+1

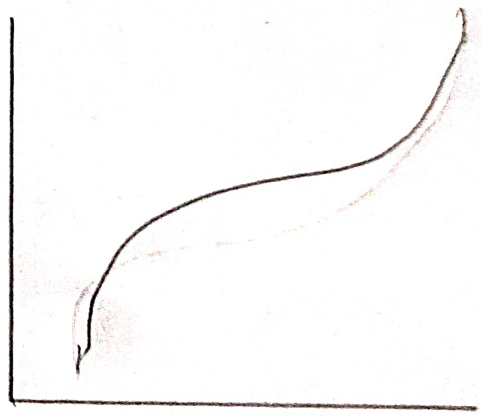
Between 0 to 1

Spearman Rank Correlation

⇒ For linear data only, Person correlation works well

⇒ So, for non linear Data, Spearman Rank correlation work well.

$$r_s = \frac{cov(R(x), R(y))}{\sigma(R(x)) \times \sigma(R(y))}$$



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

↳ Rank assigns in assending order.