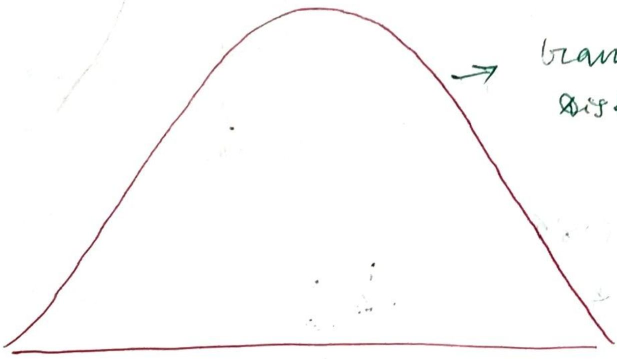
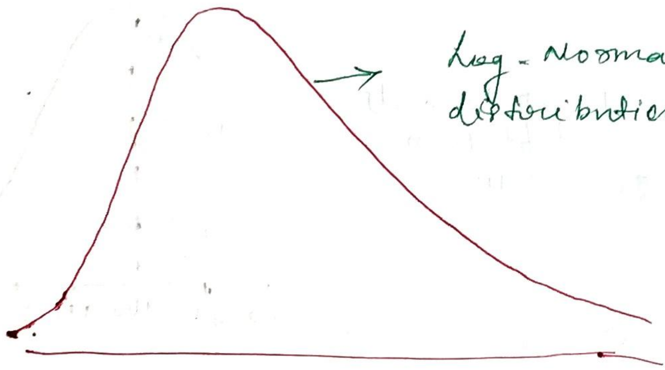
11-Sep-2022

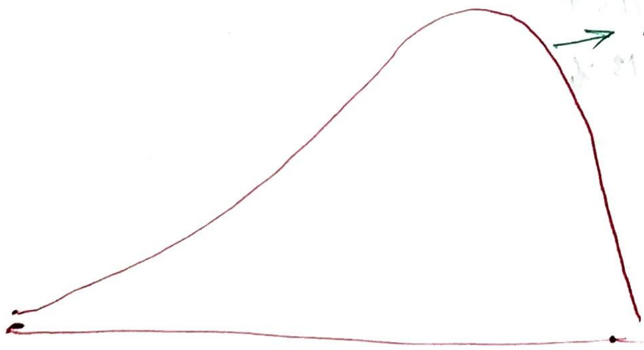
Central Limit Theorem



→ Gaussian / Normal Distribution



→ Log - Normal or Right skewed distribution



→ Left skewed

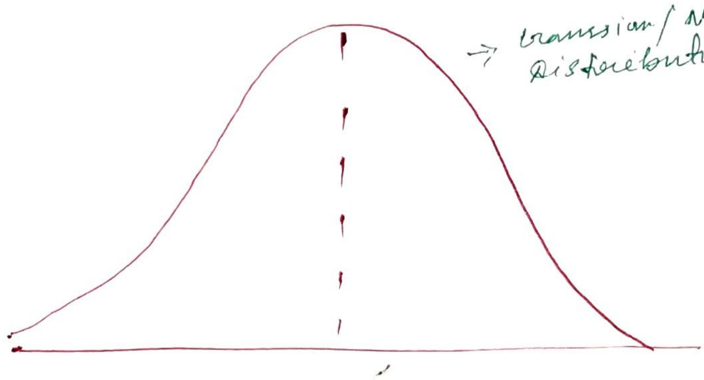
Let's consider ~~of~~ a sample size of $n \geq 30$
Sample data [n].

$S_1 \Rightarrow \{x_1, x_2, x_3, \ldots, x_n\} \rightarrow \bar{x}_1$
$S_2 \Rightarrow \{x_1, x_2, x_3, \ldots, x_n\} \rightarrow \bar{x}_2$
$S_3 \Rightarrow \{x_1, x_2, x_3, \ldots, x_n\} \rightarrow \bar{x}_3$
$\vdots$
$S_m \Rightarrow \{x_1, x_2, x_3, \ldots, x_n\} \rightarrow \bar{x}_m$

Sample mean
$\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots \bar{x}_m$ ⇒ Mean of diff. samples

where,
n = size of sample
m = no. of sample

$X = \{ \bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_m \}$



→ Gaussian / Normal Distribution.

→ Central limit theorem says that whether population data is normally distributed or not normally distributed, but if take a sample data of size greater than or equal to 30, and 'm' no. of sample, then the mean of 'm' sample will follow Gaussian/Normal distribution.

→ the larger the value of n, better will be ~~res~~ result.

Que Calculate the size of shark throughout the world?

Ans Assume 10 different region of sample size ≥ 30. then mean of the sample will follow normal distribution.

Probability

- It is a measure of the likelihood of an event.

e.g: Tossing a fair coin

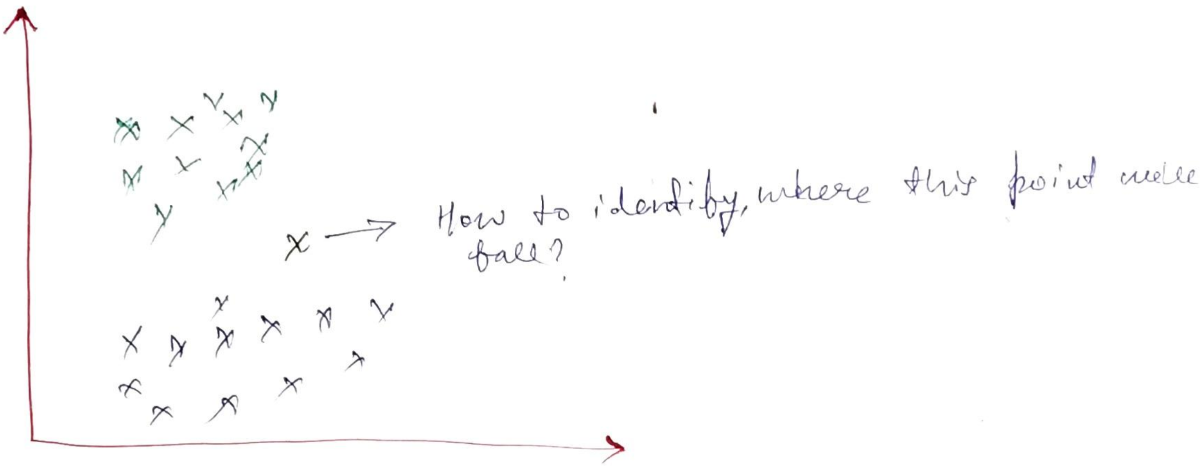
$P(H) = 0.5 \quad P(T) = 0.5$

SHOLAY Movie → win $P(H) = 1$ (unfair coin)

e.g.: Rolling a Dice

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

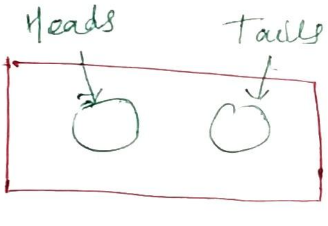
Application



Mutual Exclusive Events

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

e.g: ① Tossing a coin
② Rolling a dice



No intersection.

Non- Mutual Exclusive Events

- Two events can occur at the same time.

e.g: ① Picking randomly a card from a deck of cards, two events "heart" and "King" can be selected.

② Bag of marbels (G, Y, R marbels)

Addition Rule for Mutual Exclusive Events

Que what is the probability of coin landing on heads or tails?

Ans
$$\boxed{P(A \text{ or } B) = P(A) + P(B)} \rightarrow \text{Addition Rule}$$

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

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

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

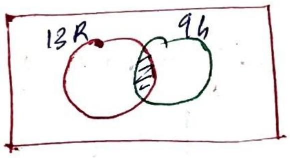
Addition Rule for Non-Mutual Exclusive Events

Que Bag of marbels : 10 Red, 6 Green, 3 (Red and Green)
when picking randomly from a bag of marbels, what is the probability of ~~choosss~~ choosing a marble that is red or green?

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

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

$$\cancel{P(R \text{ or } G) = \frac{10}{19} + \frac{6}{19} - \frac{3}{19}}$$



$$\cancel{P(R \text{ or } G) = 13/19}$$

$P(R) = 13/19$

$P(G) = 9/19$

$P(R \text{ and } G) = 3/19$

Total Red Marbels = 10 + 3 = 13

Total Green Marbels = 6 + 3 = 9

$$P(R \text{ or } G) = \frac{13}{19} + \frac{9}{19} - \frac{3}{19} = 1$$

Que. deck of cards → what is the probability of choosing heart and queen? → Non-Mutual Exclusive Events

Ans

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

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

$$= \frac{16}{52}$$

Multiplication Rule

① Dependent Events

- Two events are dependent, if they affect one another.

e.g. Bag of marbels { o o o o / o o o }

$P(R) = 4/7$

↓

1 ~~Green~~ Red Marble is taken

↓

$P(R) = 3/6$ (changes has affected the probability)

② Independent Events

: Two events doesn't affect each other.

Que what is the probability of rolling a "5" and then a "3" with the normal 6 sided dice? → Independent Events

Ans

Multiplicative Rule for Independent Events

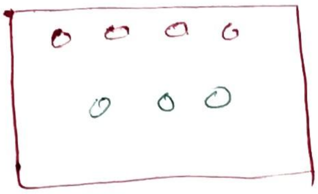
$$\boxed{P(A \text{ and } B) = P(A) * P(B)}$$

$$P(5 \text{ and } 3) = P(5) * P(3)$$

$$= 1/6 * 1/6 = 1/36$$

Multiplicative Rule for Dependent Events

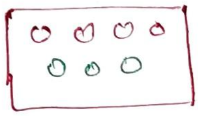
Que

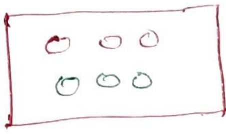


what is the probability of drawing a "Red" and then drawing a "Green" marble from the bag? → Dependent Events.

Ans

$P(R) = 4/7$ → 1 Red taken out → ~~P(R) = 3/6~~





→ $P(G/R) = 3/6 = 1/2$ → Conditional Probability

$$P(R \text{ and } G) = P(R) * P(G/R)$$
$$= 4/7 * 3/6$$
$$= 2/7$$

Permutation

eg: School of children
chocolate Factory
{Dairy Milk, Kit Kat, Milky Bar, Sneakers, 5 Star}

5 __ 4 __ 3 __

$= 5 * 4 * 3$

$= 60$ options or ways (Permutation)

→ with permutation order matters.

{DM KK MB}
{KK DM MB}
{KK MB DM}
} All the possible arrangements.

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

r = no. of selection
n = total no. of objects

→ $\frac{5!}{(5-3)!} = 60$ ways.

Combination

- Repetition will not occur.

{DM KK MB}
{MB KK DM} } not possible, only unique combination is possible.

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

Previous one,

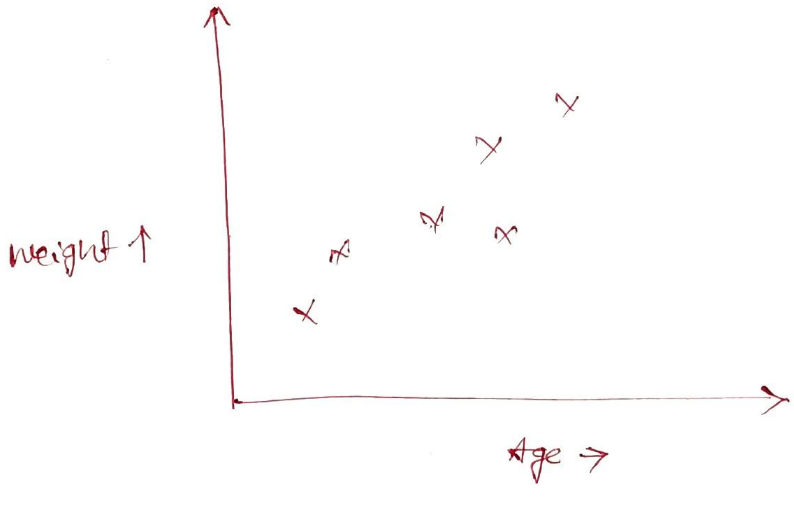
eg: $^{5}C_{3} = \frac{5!}{3!(5-3)!} = \frac{5 \times 4 \times 3!}{3! \times 2!} = 10$ combinations

Covariance

(Helpful in feature selection)

x Age	y Weight
12	40
13	45
15	48
17	60
18	62

Age ↑, Weight ↑
Age ↓, Weight ↓



Que: Can we quantify the relationship b/w x and y using mathematical formula?

Ans: Using covariance, we can measure.

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

Variance, $S^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$

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

$$\boxed{cov(X,X) = S^2}$$

$\rightarrow$ $\boxed{cov(X,X) = Var(X)}$

e.g

$\bar{X} = 15$

$\bar{Y} = 51$

$$cov(X,Y) = \frac{(12-15)(40-51) + (13-15)(45-51) + (15-15)(48-51) + (17-15)(60-51) + (18-15)(62-51)}{5-1}$$

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

$= 14.75\cancel{24}$ (+ve)

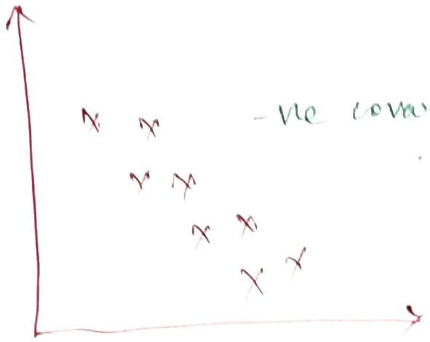
$\downarrow$

$$\boxed{\begin{matrix} X\uparrow & Y\uparrow \\ X\downarrow & Y\downarrow \end{matrix}}$$

+ve covariance $\left.\begin{matrix} X\uparrow & Y\uparrow \\ X\downarrow & Y\downarrow \end{matrix}\right.$

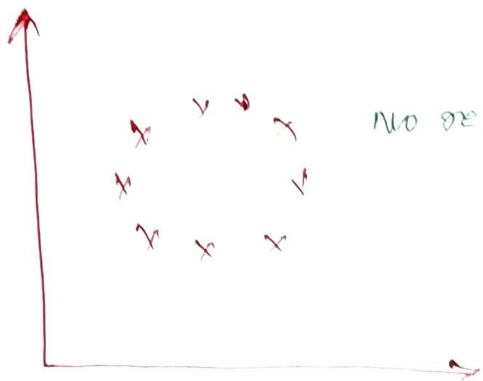
-ve covariance $\left.\begin{matrix} X\uparrow & Y\downarrow \\ X\downarrow & Y\uparrow \end{matrix}\right.$

covariance = 0 } No relationship b/w X and Y

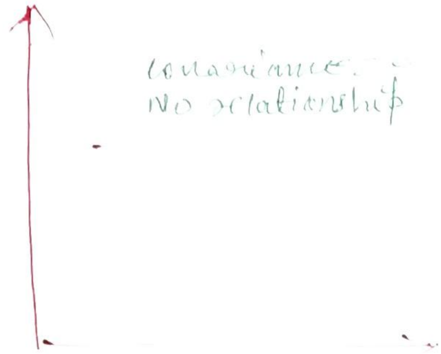


-ve covariance

+ve covariance



No relationship



covariance:
No relationship

e.g

X	Y
10	~~5~~ 4
8	~~4~~ 6
7	~~3.2~~ 8
6	~~3~~ 10

$\bar{X} = 7.75$

$\cancel{\bar{Y} = 3.8}$ $\bar{Y} = 7$

$$\cancel{cov(X,Y) = (10-7.75)(5-3.8) + (8-7.75)(4-3.8) + (7-7.75)(3.2-3.8) + (6-7.75)}$$

$$cov(X,Y) = \frac{(10-7.75)(4-7) + (8-7.75)(6-7) + (7-7.75)(8-7) + (6-7.75)(10-7)}{4-1}$$

$= -3.25$ (-ve covariance)

$$\left\{ \begin{matrix} X\uparrow, Y\downarrow \\ X\downarrow\ Y\uparrow \end{matrix} \right\}$$

Pearson correlation coefficient (−1 to +1)

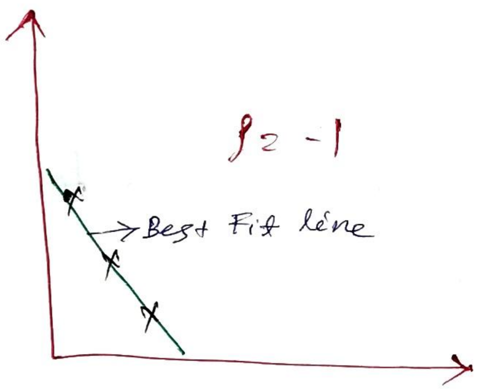
$$\boxed{\rho(x,y) = \frac{\mathrm{Cov}(x,y)}{\sigma_x \cdot \sigma_y}} \qquad \boxed{-1 \leq \rho(x,y) \leq +1}$$

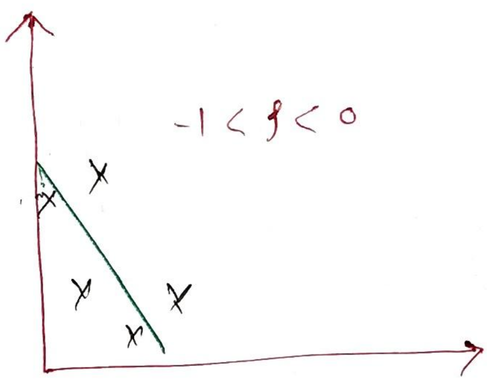
→ there is no such limit in covariance for positive and negative value.

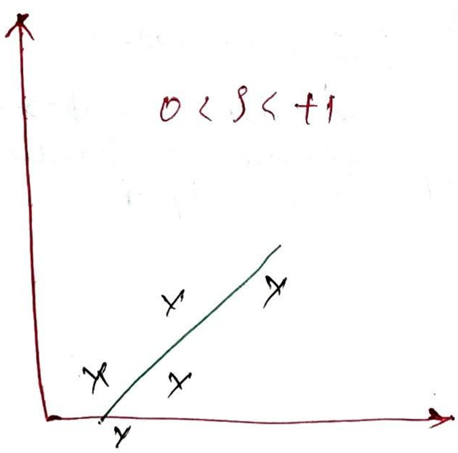
eg $X \xrightarrow{0.5} Y$, $X \xrightarrow{0.7} Z$

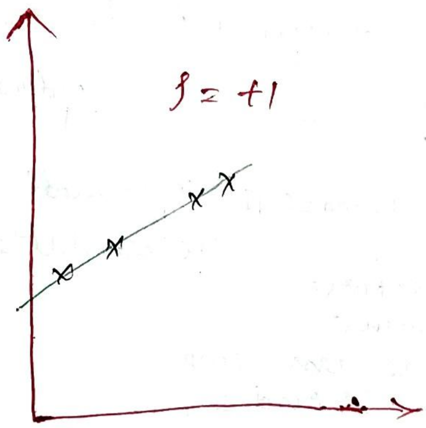
X and Z are highly correlated as compared to X and Y.

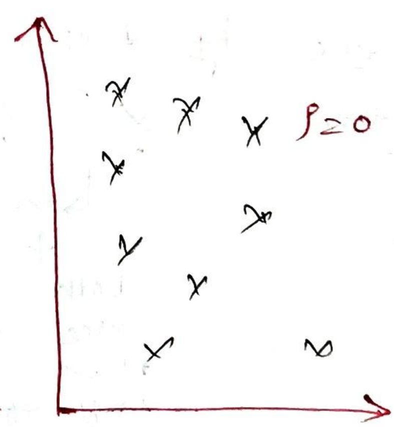
- More +ve value towards +1, more '+ve' correlation
- More −ve value towards −1, more '−ve' correlation











Spearman's Rank Correlation

→ Pearson correlation hold good for linear data only.
→ To overcome this problem, we use Spearman's Rank Correlation.

$$\boxed{r_s = \frac{cov(R(X), R(Y))}{\sigma(R(X)) * \sigma(R(Y))}}$$

where, R(X) = Rank of X
R(Y) = Rank of Y

e.g.:

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

R(X): In ascending order
R(Y): In ascending order

→ If two value are same, we will assign same rank.

→ Find mean of R(X), R(Y) and S.D of R(X), R(Y) to find out Spearman's Rank correlation?

Mean(R(X)) = 2.5
Mean(R(Y)) = 2.5

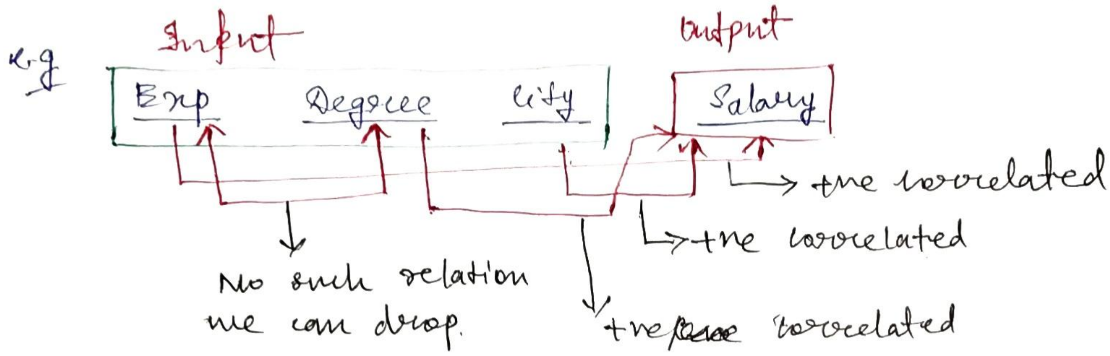
Que why this correlation will be used?

Ans. X Y Z output

→ +ve / -ve correlation.
X will be important feature to predict the output.

correlation = 0.95
Both feature are same, then we can drop any of X and Z.

correlation = 0.01
we can delete Y



→ If two input feature are highly correlated, then we can delete one of these two input feature.