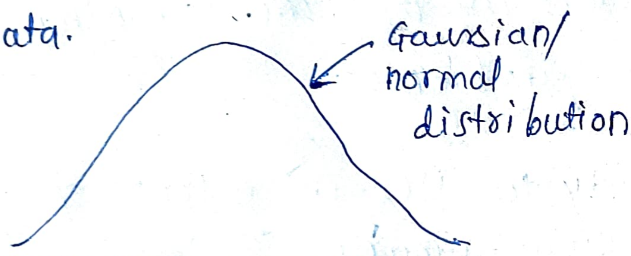
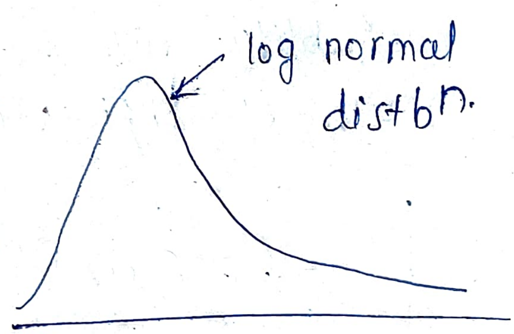
Stats Basic

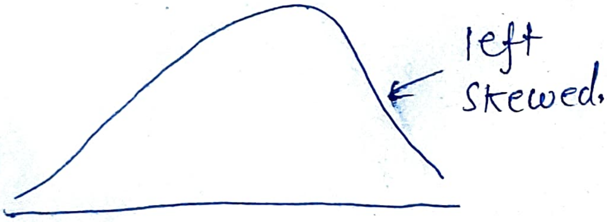
① central limit theoram
② probability
③ permutations and combinations
④ covarience, pearson co-relation,
Spearman co-relation.
⑤ Bernauli's distribution
⑥ Binomial distribution.
⑦ Power law (pareto distribution)

* Central limit theoram

population data.







In any population data (N) with any distribution where sample data (n), consider m such samples as below:

$S_1 = \{x_1, x_2, x_3 \cdots x_n\} \Rightarrow \bar{x}_1$

$S_2 = \{x_1, x_2, x_3 \cdots x_n\} \Rightarrow \bar{x}_2$

$\vdots$

$S_m = \{x_1, x_2, x_3 \cdots x_n\} \Rightarrow \bar{x}_m$

For $n \geq 30$

when the means of m samples are plotted will follow a gaussian distribution (approximately).

* sampling Technique : sampling with replacement

* probablity

It is a measure of likely hood of an event.

Eg: Tossing a fair coin.

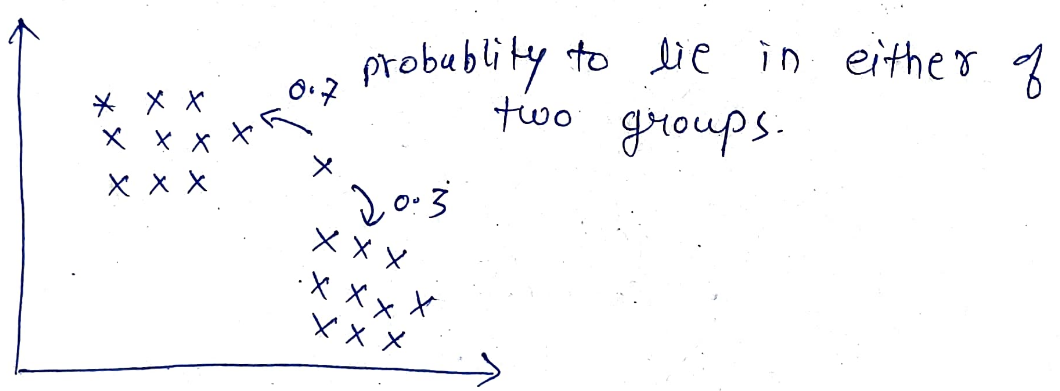
prob (H) = 0.5

P(T) = 0.5

Eg: Rolling a dice

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

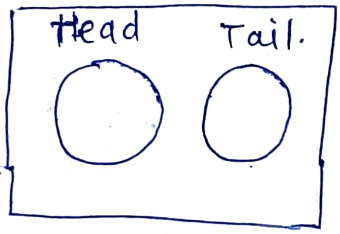
* probablity will be used in machine learning. ie in classification problems



① Mutually Exculsive events

Two events are said to be mutually Exclusive if they cannot occur at same time.

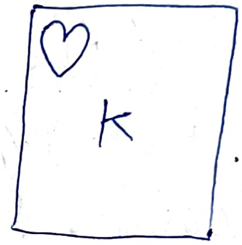
Eg: 1) Tossing a coin   2) Rolling a dice.



② Non-mutual Exclusive Event

Two events can occur at the same time.

Eg: picking randomly a card from a deck of cards, two events, "heart" and "king" can be selected



A) mutually Exclusive event

Que: 1) what is the probablity of a coin landing on heads or tails.

(Addition rule for mutually exclusive event)

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

$$= 0.5 + 0.5$$

$$= 1$$

2) what is the probablity of getting 1 or 6 or 3 while rolling a dice.

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

$$= 1/6 + 1/6 + 1/6$$

$$= 3/6$$

$$= 1/2$$

$$= 0.5.$$

B) Non- mutually Exclusive Event

Que: 1 In Bag of marbles.

10 Red, 6 Green 3 (R & G)

when picking randomly from a bag of marbles what is probability of choosing a marble that is read or green.

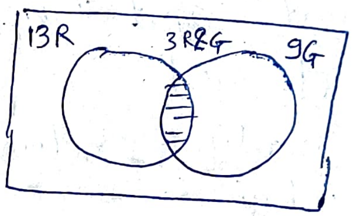
(Addition rule for non-mutual Exclusive event)

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

$\therefore$ $P(\text{Red}) = \cancel{10/19}$ $13/19$

$P(\text{Green}) = \cancel{6/19}$ $9/19$

$P(R \& G) = \cancel{3/19}$ $3/19$



$P(\text{Red or Green}) = \cancel{10/19} + \cancel{6/19} - \cancel{3/19}$

$\cancel{= 13/19}$

$= 13/19 + 9/19 - 3/19 = 19/19 = 1$

Que: 2 Deck of cards

probability of choosing Heart or Queen card.

$P(\heartsuit \text{ or } Q) = P(\heartsuit) + P(Q) - P(\heartsuit \text{ and } Q)$

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

$= \frac{16}{52}$

$= 4/13$

Multiplication rule.

1) **Dependent Event:**

Two events are dependent if they offect one another.

Eg: Bag of marble $\left\{ \begin{array}{l} 4w - 0000 \\ 3y - 000 \end{array} \right\}$.

$p(w) = 4/7 \xrightarrow[\text{that 1 yellow}]{\text{After}} p(y) = 3/6.$

ie $P(y)$ depends on $p(w)$.

Que: what is the probablity of rolling a 5 and then a 3 with a normal 6 sided dice.

(Independent Event)

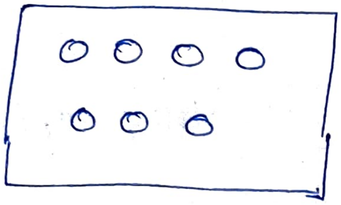
$\Rightarrow$ $P(5) = 1/6$, $P(3) = 1/6$.

multiplication rule for independent event.

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

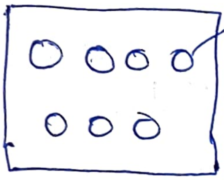
$$\begin{aligned} P(Req) &= P(5) \times P(3) \\ &= 1/6 \times 1/6 \\ &= 1/36. \end{aligned}$$

Que: 2 Bag of marbles



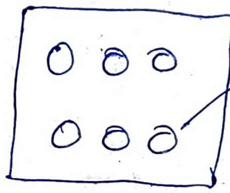
4 orange
3 yellow

what is probablity of drawing a "orange" and then a "yellow" marble from the bag:
(Dependent Event)



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

$\longrightarrow$



$P(\text{yellow}/\text{orange}) = 3/6$

ie P(yellow when orange has happened) $= P(Y/O) = 3/6$

conditional probability

$$P(O \text{ and } Y) = P(O) \times P(Y/O)$$

$$= \frac{4}{7} \times \frac{3}{6} = \frac{2}{7}$$

* Permutations

Choclate
{Dairy milk, kit kat, milkybar, sneakers, 5 stars}

* $\underline{5} \times \underline{4} \times \underline{3} = 60$ ways to pick 3 choclates.

Note: with permutations order matters..
ie All possible Arrangements will be counted.

permutation = $({}^nP_r)$

$n \equiv$ total No. of objects

$r \equiv$ No. of selections

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

$\therefore$ for above example

$n=5 \quad r=3$

permutations = $^5P_r = \frac{5!}{(5-3)!} = \frac{5\times4\times3\times\cancel{2}\times\cancel{1}}{\cancel{2}\times\cancel{1}}$

$= 60$

* <u>combinations</u> $({}^nC_r)$

(Repetations will not occur)

ie unique arrangements are only allowed.

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

$\therefore$ for above choclate example

$$^5C_3 = \frac{5!}{3!\,(5-3)!} = \frac{5!}{3!\times2!} = \frac{5\times\overset{2}{\cancel{4}}\times\cancel{3}\times\cancel{2}\times1}{\cancel{3}\times\cancel{2}\times1\times\cancel{2}\times1}$$

$=$ 10 combinations.

* Co-varience (Feature selection)

Age (X)	weight (Y)
12	40
13	45
15	48
17	60
18	62

Age ↑ weight ↑

Age ↓ weight ↓

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

Note: varience (x), $\sigma^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1} = \dfrac{\sum (x_i - \bar{x})(x_i - \bar{x})}{n-1}$

$$\boxed{\sigma^2 = Cov(X, X)}$$

covarience of (X, X) is varience of X.

for above example

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

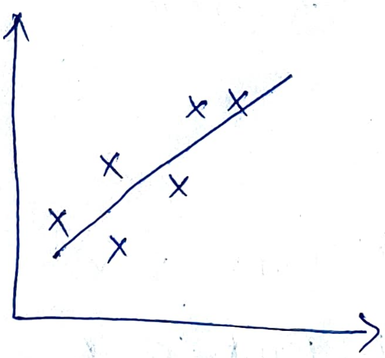
$$\bar{y} = 51$$

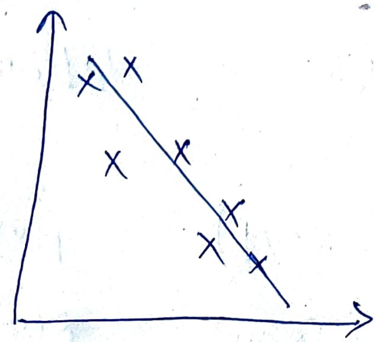
$$Cov(Age, weight) = \frac{(12-15)(40-51) + (13-15)(45-51) + (15-15)(48-51) + (17-15)(60-51) + (18-15)(62-51)}{6-1}$$

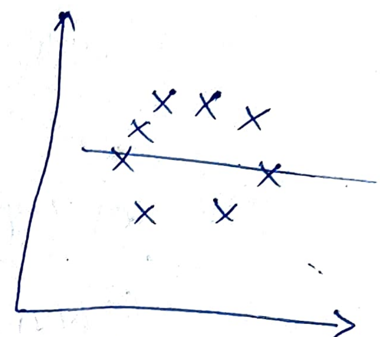
$$= \frac{(-3 \times -11) + (-2 \times -6) + 0 + 2 \times 9 + 3 \times 11}{5}$$

$$Cov(Age, weight) = 19.2$$

* +ve covarience ⇒ X↑ Y↑ or X↓ Y↓
* −ve covarience ⇒ X↓ Y↑ or X↑ Y↓
* Zero covarience ⇒ No relation b/w X and Y.


+ve covarience


−ve co-varience


no relation

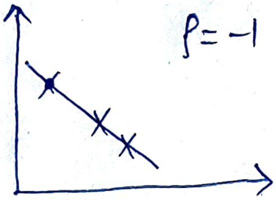
* <u>Pearson</u> <u>co-relation</u> <u>coefficient</u> (−1 to 1).

$$\rho(x,y) = \frac{cov(X,Y)}{\sigma_X \times \sigma_Y}$$

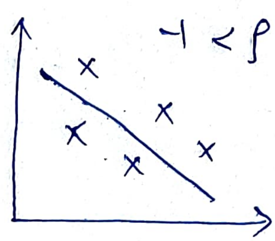
* covarience has no limit to its value so to limit it, we use PCC (ρ).

* more +ve value towards +1 more +ve co-related variables X and Y.

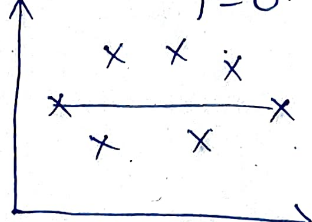
$\rho = -1$



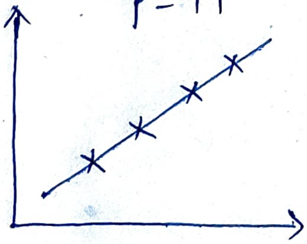
$-1 < \rho < 0$

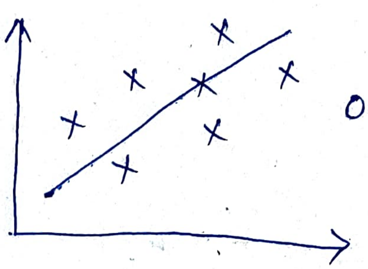


$\rho = 0$



$\rho = +1$




$0 < \rho < +1$

* Spearman's Rank Co-relation.

* pearson co-relation holds good for only linear data

* For non-linear data, we have to use Spearman's Rank co-relation.

$$r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} \times \sigma_{R(y)}}$$

$R(x) \Rightarrow$ Rank of x.

$R(y) \Rightarrow$ Rank of y.

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

mean $(R(y))$ =

mean $(R(x)) = \frac{4+3+2+1}{4}$

$= 2.5$

$\sigma_{R(x)}^2 = \sigma_{R(y)}^2 =$

$= \frac{(4-2.5)^2 + (3-2.5)^2 + (2-2.5)^2 + (1-2.5)^2}{4-1}$

$\sigma^2 = 1.67$

$\sigma_{R(y)} = \sigma_{R(x)} = 1.29$

$Cov(R(x), R(y)) =$

$$\frac{(4-2.5)(1-2.5) + (3-2.5)(2-2.5) + (2-2.5)(3-2.5) + (1-2.5)(4-2.5)}{4-1}$$

$Cov(R(x), R(y)) = -1.67$

$$r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} \times \sigma_{R(y)}} = \frac{-1.67}{1.29 \times 1.29} = -1$$

$\boxed{\therefore r_s = -1}$

* where and why these co-relations will be used??

I/P

x	y	z

⟹ O/P

1) x and O/p High co-related. ⟹ x is important feature
2) y and o/p very low co-related ⟹ y can be dropped.
3) x and z are 95% co-related ⟹ either of x and z can be dropped.

Eg:

I/P

Experience	Degree	City

O/P

Salary

1) Exp and salary ⟹ +ve co-relation
2) City and salary ⟹ +ve co-relation.
3) Exp and degree ⟹ no relation
4) ~~exp~~ deg and salary ⟹ +ve co-relation.