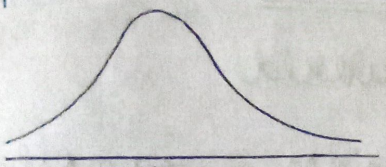
STATISTIC FOR DATA SCIENCE [DAY 4]

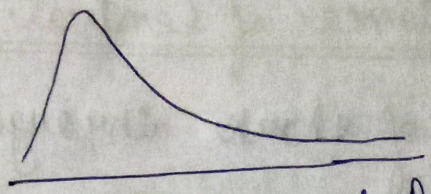
(1) Central Limit Theorem
(2) Probability
(3) Permutation and Combination
(4) Covariance, Pearson Correlation, Spearman Rank correlation
(5) Bernoulli's Distribution
(6) Binomial Distribution
(7) Power Law {Parcto Distribution}

(1) Central Limit Theorem

Population Data (N)



[Normal Distribution]



[Log Normal Distribution]

[Right skewed]

* In the above distribution data, if we consider sample data (n)

Let 1st sample data $\{x_1, x_2, x_3, \ldots x_n\}$ of mean $\bar{x}_1$

2nd sample data $\{x_1, x_2, x_3, \ldots x_n\}$ of mean $\bar{x}_2$

3rd sample data $\{x_1, x_2, x_3, \ldots x_n\}$ of mean $\bar{x}_3$

.

.

m^{th} sample data $\{\ldots\}$ of mean $\bar{x}_m$

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

Let, $n \geq 30$ if we plot all the sample data then the central limit theorem then we will get a Gaussian / Normal distribution

Let us consider a population data which may be normally distributed or not be then the Central Limit Theorem say that if we take sample of size $n \geq 30$ and selected 'm' no. of samples of respective mean and if we plot the data then we will get normal distribution of the sample mean.

Importance of Central Limit Theorem

Size of shark through the world

⇒ We will take 10 different region, sample population of size $n \geq 30$. and make an assumption on the size of shark.

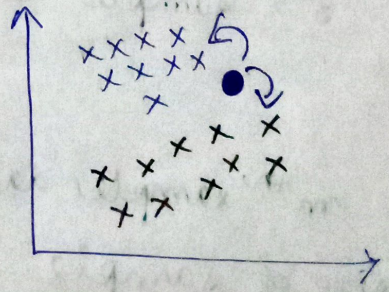
PRACTICAL APPLICATION

PROBABILITY :- Probability is a measure of the likehood of an even

Eg; Tossing a fair coin $P(H) = 0.5$ $P(T) = 0.5$

Rolling a Dice ., $P(1) = \frac{1}{6}$, $P(2) = \frac{1}{6}$, $P(3) = \frac{1}{6}$

To check whether '●' data belongs to '×××' or '××××' ⇐ we will use probability

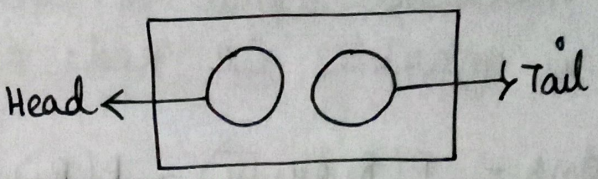


(i) Mutual Exclusive Event

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

Example (i) Tossing a Coin (ii) Rolling a dice

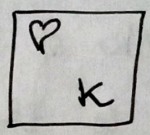
(ii) Non-Mutual Exclusive Event



Two events can occurr at the same time

Example (i) Winning and Loosing game

(ii) Picking Randomly a card from a deck of cards, two events "heart" and "king" can be selected



PROBLEM STATEMENT : MUTUAL EXCLUSIVE EVENT

Q1 what is the probability of coin landing on heads or tails

Ans:- $P(A \text{ or } B) = P(A) + P(B) = \frac{1}{2} + \frac{1}{2} = 1$

Q2 what is the probability of getting 1 or 6 or 3

Ans:- $P(1) = \frac{1}{6}$, $P(6) = \frac{1}{6}$, $P(3) = \frac{1}{6}$

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

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

Problem Statement : Non-Mutual Exclusive Event

Q1 Bag of Marbles : 10 Red, 6 Green, 3 (R & G)

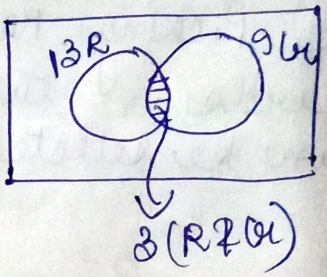
When picking randomly from a bag of marbles what is the probability of choosing a marble i.e. Red or green

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

$= \frac{10}{19} + \frac{6}{19} - \frac{3}{19} = \frac{13}{19}$

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

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



Q2 Deck of cards

What is the probability of choosing heart ~~and~~ or Queen

Ans:- $P(H \text{ or } Q) = P(H) + P(Q) - P(H \text{ and } Q)$

$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{13}{52} + \frac{3}{52}$

$= \frac{16}{52}$

Multiplication Rule

(i) Dependents Events : Two events are dependent if they effect one another.

E.g. Bag of Marbles { White - 4, Yellow - 3 }

$P(W) = \frac{4}{7}$ → $P(Y) = \frac{3}{6}$

↑ 1 white marble taken out

here white marble effecting yellow marble's probability of occurring.

Q1. What is the probability of rolling a "5" and then a "3" with a normal 6 size sided dice?

Ans:- $P(A \text{ and } B) = P(A) * P(B) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$.

[Independent Events]

* Indepent Events Eg:- Tossing a coin

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

Q2. There are bag of marbles (4orange, 3yellow) what is the probability of drawing a "orange" and then drawing a "yellow" marble from the bag [Dependent Event]

Ans:- Total no. of marble = 4+3 = 7

Probability of orange, $P(O) = \frac{4}{3}$.

After taking out 1 marble orange.

then probability of taking out yellow marble $\boxed{P(Y/O)} = \frac{3}{6} = \frac{1}{2}$

↓

[Conditional Probability]

$\therefore P(O \text{ and } Y) = P(O) * P(Y/O) = \frac{4}{7} * \frac{1}{2} = \frac{4}{14}$

$= \frac{2}{7}$

PERMUTATION

* All the possible arrangements {Dairy Milk, Kit Kat, Milky Bar, Sneakers, 5 star}

5 * 4 * 3

= 60 ways chocolates can kee choosen

* with permutation order matters

$$ {}^{n}P_{r} = \frac{\lfloor n}{\lfloor n-r} $$

n = total no. of objects = 5

r = no. of selection = 3.

$$ = \frac{\lfloor 5}{\lfloor 5-3} = \frac{5 \times 4 \times 3 \times \lfloor 2}{\lfloor 2} = 60 $$

COMBINATION

* Repeation will not occur
* unique combination possible

Formula $$ {}^{n}C_{r} = \frac{\lfloor n}{\lfloor r \ \lfloor n-r} = \frac{\lfloor 5}{\lfloor 3 \ \lfloor 2} $$

$$ = \frac{5 \times 4 \times \lfloor 3}{\lfloor 3 \times 2} $$

$$ = 10. $$

∵ the repeated one is removed

∴ Permutation > Combination.

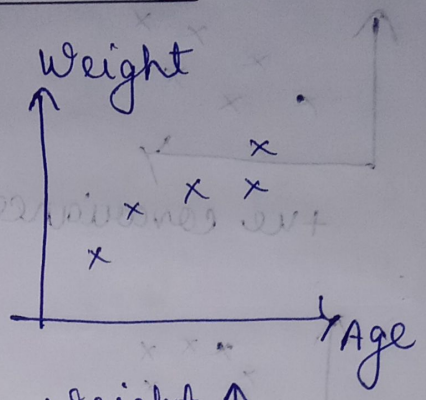
[value] [value]

* Permutation used in **Dream 11**

Combination [Eg; choosing 3 deserts from 10 of menu dishes]

COVARIANCE [Feature Selection]

Age X	Weight Y
12	40
13	45
15	48
17	60
18	62



Age ↑ Weight ↑

Age ↓ Weight ↓

* Quantify the relationship x & y using Mathematical question

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

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

$$Cov(x,x) = Var(x)$$ [Interview Question]

Now, $\bar{x} = 15$, $\bar{y} = 51$

~~$\sum (x_i - \bar{x}) = (12-15) + (13-15) + 0 + (17-15) + (18-15)$~~

~~$= -3 - 2 + 2 +$~~

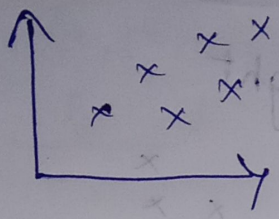
+ve Covariance x↑ y↑

x↓ y↓

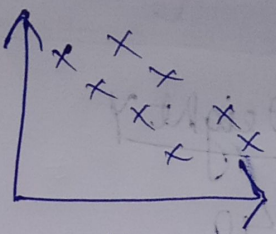
-ve Covariance x↑ y↓

x↓ y↑

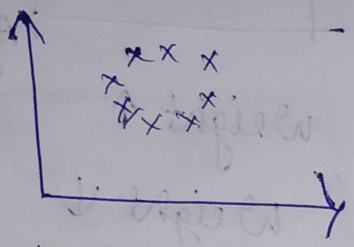
Covariance 0 [No relation with x & y]



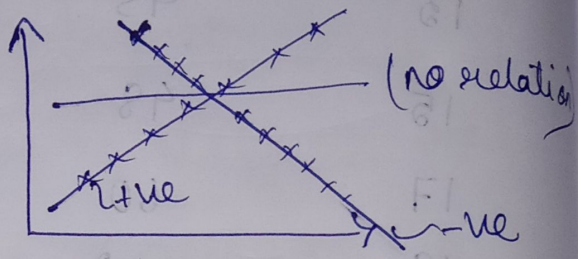
+ve covariance



-ve covariance



No relation



x	y
10	4
8	6
7	8
6	10

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

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

$$= \frac{-13}{3} = -4.33$$

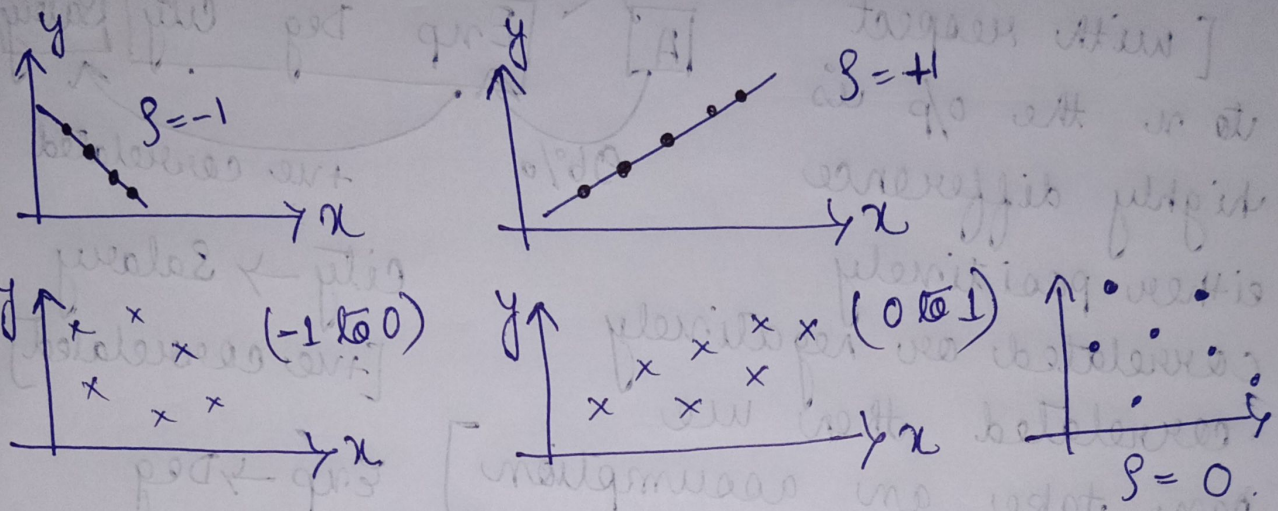
Pearson Correlation Coefficient [-1 ≤ 1]

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

* with the help of this we are trying to restrict the value bet^n -1 ≤ 1

More the value towards +1
⇒ More +ve correlated it is

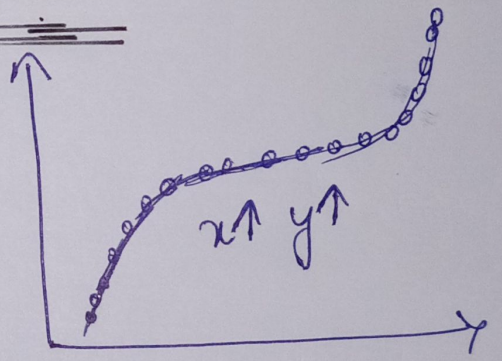
And, More the value towards −1
⇒ More −ve correlated it is



Spearman Rank Correlation

For non-linear data we use Spearman Rand Correlation

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



Spearman Correlation = 1
Pearson correlation = 0.88

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

* Rank = [Assiging value by no. in ascending order]