θ is sufficient

$L = f(x_1, \theta) \prod_{i=1}^{n} f(x_i, \theta)$

Statistical Modelling

16-03-22

Correlation:

$x \rightarrow$ No. of people
$y \rightarrow$ Amt of covid vaccine
(Variance) mean fortified

↑↑ ↓↓ +ve correlation

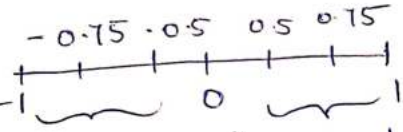
$x \rightarrow$ No. of people vaccinated
$y \rightarrow$ No. of infected

↑↓ ↓↑ -ve correlation.

(correlation coefficient) → r 'r' 'p'

Measures of the strength of the relationship between the 2 variables.

⇒ $r \rightarrow -1$ and 1 x, y



-ve correlated | +vely correlated

No correlation.

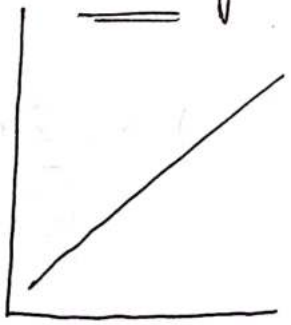
x x_1 y_2 x_3 x_4 x_5
y y_1 y_2 y_3 y_4 y_5

(x_1, y_1) (x_2, y_2) (x_3, y_3) (x_4, y_4) .....

1. Scatter diagram
2. Direct formula
3. Formula correlation
   - Covariance
   - Variance

Scatter diagram:

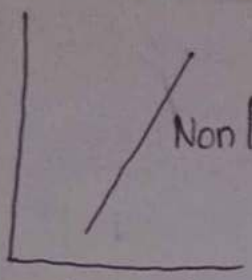
1)



perfect positive correlation (linear)

$r = +1$

2)

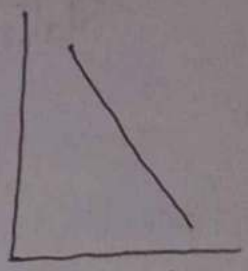


positive non-linear correlation.

0.5 - 0.75

< 0.5

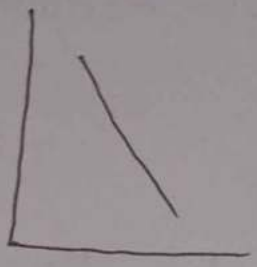
3)



perfect -ve (linear) correlation.

$r = 1$

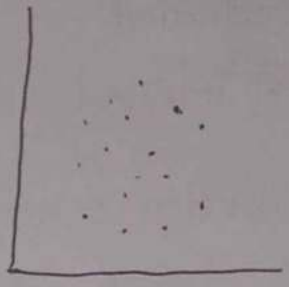
4)



-ve non-linear correlation

0 to -1

5)



No correlation

Formula:-

1) $$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\cdot\sqrt{\Sigma(y-\bar{y})^2}}$$

2) $$r = \frac{N\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{N\Sigma x^2-(\Sigma x)^2}\cdot\sqrt{N\Sigma y^2-(\Sigma y)^2}}$$ $(x = X-\bar{X})$ $(y = Y-\bar{Y})$

3) $$r = \frac{N\Sigma dx\cdot dy - \Sigma dx\cdot\Sigma dy}{\sqrt{N\Sigma dx^2-(\Sigma dx)^2}\cdot\sqrt{N\Sigma dy^2-(\Sigma dy)^2}}$$

[X, Y are large
X, Y are not integers]

$dx = X - A$

$dy = Y - B$

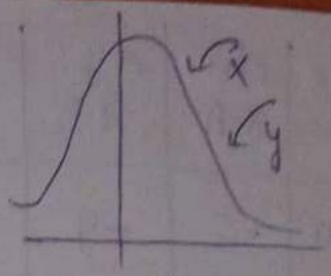
∴ A = Assumed mean in X

B = Assumed mean in Y

$\Sigma(x-\bar{x}) = 0$

4) $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\cdot\sqrt{\Sigma y^2}}$$

Correlation: $r = \frac{Cov(x,y)}{\sqrt{Var\,x} \cdot \sqrt{Var\,y}}$



Problems on finding correlation coefficient:

1) Calculate r blw X & Y

X:	1	3	5	8	9	10
Y:	3	4	8	10	12	11

Sol:

X	Y	dx $X-\bar{X}$	dy $Y-\bar{Y}$	dx^2	dy^2	$dx \cdot dy$
1	3	-5	-5	25	25	25
3	4	-3	-4	9	16	12
5	8	-1	0	1	0	0
8	10	2	2	4	4	4
9	12	3	4	9	16	12
10	11	4	3	16	9	12
		0	0	64	70	65

Mean

$\bar{X} = 6$

$\bar{Y} = 8$

$$r = \frac{N\Sigma dx.dy - \Sigma dx \cdot \Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2} \cdot \sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$$r = \frac{6(65) - 0}{\sqrt{6(64) - 0} \cdot \sqrt{6(70) - 0}} = 0.9711$$

2. Calculate r.

X	10	12	13	16	17	20	25
Y	19	22	26	27	29	33	37

$N = 7$

$\bar{X}, \bar{Y}$ → Not integers

$A = 16$

$B = 28$ } Assumed means.

X	Y	dx X-A	dy Y-B	dx^2	dy^2	$dx \cdot dy$
10	19	-6	-9	36	81	54
12	22	-4	-6	16	36	24
13	26	-3	-2	09	4	06
16	27	0	-1	0	1	0
17	29	1	1	1	1	01
20	33	4	5	16	25	20
25	37	9	9	81	81	81
		1	-3	159	229	186

$$r = \frac{N\Sigma dx\,dy - \Sigma dx \Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2} \cdot \sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{7(186) - (1)(-3)}{\sqrt{7(159) - (1)^2} \cdot \sqrt{7(229) - (-3)^2}}$$

$$= 0.9802$$

X & Y are strongly +vely correlated.

3. 12 pairs of observations: Find r.

$\Sigma X = 30$, $\Sigma Y = 5$, $\Sigma X^2 = 670$, $\Sigma Y^2 = 285$, $\Sigma XY = 334$

Sol: $$r = \frac{N\Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2} \cdot \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} = 0.7835$$

Correction:

($X = 11$, $Y = 4$) was wrongly copied. the correct value being ($X = 10$, $Y = 14$). Find corrected r.

... check if θ is sufficient

$L = (1+\theta)^n \cdot \prod x_i^{\theta}$

$N = 12$

$\Sigma x = 30 - 11 + 10 = 29$

$\Sigma y = 5 - 4 + 14 = 15$

$\Sigma x^2 = 670 - 121 + 100 = 649$

$\Sigma y^2 = 285 - 16 + 196 = 465$

$\Sigma xy = 334 - 44 + 140 = 430$

$r = 0.7746$

Error in r $= 0.7835 - 0.7746$

$= 0.0089$

$E\% = \frac{0.0089}{0.7835} \times 100\% = 1.135\%$

4) Using 44 and 26 respectively as origin of x & y

Sol:

x	y	dx (x-44)	dy (y-26)	dx^2	dy^2	dx·dy
43	29	-1	3	1	9	-3
44	31	0	5	0	23	0
46	19	2	-7	4	49	-14
40	18	-4	-8	16	64	32
44	19	0	-7	0	49	0
42	27	-2	1	4	1	-2
45	27	1	1	1	1	1
42	29	-2	3	4	9	-6
38	41	-6	15	36	225	-90
40	30	-4	4	16	16	-16
42	26	-2	0	4	0	0
57	10	13	-16	169	256	-208
		-5	-6	255	704	-306

A = 44
B = 26

= 334

t

$N = 12$

$\Sigma(x-8)^2 = 150$

$\Sigma x = 120$

$\Sigma(y-10)^2 = 200$

$\Sigma y = 130$

$\Sigma(x-8)(y-10) = 50$

x 18 19 20 ...

y 17 21 32 ...

x 0-10 10-20 20-30 ...

y 10-20 20-30 30-40 ...

Using 20 as the working mean for the price and 70 as the working mean for the demand

price : 14 16 17 18 19 20 21 22 23

Demand : 84 78 70 75 66 67 62 58 60

Ans: 0.954

Pearson's Rank correlation :-

Case 1 :- When values of variables are not repeated:

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

n = no. of values given.

d = difference in ranks.

Case 2 :- When values of variables are repeated.

$$r = 1 - \frac{6\left[\Sigma d^2 + \frac{m(n^2-1)}{12}\right]}{n(n^2-1)}$$

$\frac{m(n^2-1)}{12}$ - Correction factor

m = no. of times the value is repeated.

1) The following are the ranks obtained by 10 students in statistics & maths. Find pearson's Rank correlation coefficient.

S (Rx)	M (R
1	1
2	4
3	
4	
5	
6	
7	
8	
9	
10	

2) Find

x
9
8
8
8
8

S (Rx)	M (Ry)	d^2
1	1	0
2	4	4
3	2	1
4	5	1
5	3	4
6	9	9
7	7	0
8	10	4
9	6	9
10	8	4
		36

$$r = 1 - \frac{6(36)}{10(99)}$$

$$= 1 - \frac{216}{990} = \frac{990 - 216}{990} = 0.7818.$$

2) Find Rank correlation for following data:

x	y	R_x	R_y	d	d^2
92	86	1	2	-1	1
89	83	2	4	-2	4
87	91	3	1	2	4
86	77	4.5	6	-1.5	2.25
86	68	4.5	7	-2.5	6.25
77	83	6.5	4	1.5	2.25
77	52	6.5	9	2.5	6.25
63	83	8	4	4	16
53	37	9	10	-1	1
50	57	10	8	2	4

$$\gamma = 1 - \frac{6\left[\Sigma d^2 + \frac{m(m^2-1)}{12} + \frac{m(m^2-1)}{12} + \frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[51 + \frac{2(3)}{12} + \frac{2(3)}{12} + \frac{3(8)}{12}\right]}{10(99)}$$

$$= 1 - \frac{6[51 + 1/2 + 1/2 + 2]}{990}$$

$$= 1 - \frac{6[54]}{990} = 0.67$$

10 Compititors in beauty contest are ranked by 3 judges. Find which two judges has the nearest approach to common taste in beauty:

J_1	J_2	J_3	d_{xy}	d_{yz}	d_{zx}	d_x^2	d_y^2	d_z^2
1	2	3	-1	-1	2	1	1	4
4	6	7	-2	-1	3	4	1	9
6	5	4	1	1	-2	1	1	4
3	4	5	-1	-1	2	1	1	4
2	7	10	-5	-3	8	25	9	64
9	10	8	-1	2	-1	1	4	1
7	9	9	-2	0	2	4	0	4
8	3	2	5	1	-6	25	1	36
10	8	6	2	2	+4	4	4	16
5	1	1	4	1	-4	16	1	16
						82		158

$$J_1 = 1 - \frac{6(82)}{10(99)}$$

$$= 0.5030$$

$r_2 = 1 - \frac{6(23)}{990} = 0.8667$

$r_3 = 1 - \frac{6(158)}{990} = 0.0424$

2,3 has nearest approach.

$r_{23} > r_{13}, r_{12}$

$r_{12} = 0.5030$; $r_{23} = 0.8667$; $r_{13} = 0.0424$

→ Correlation is a study of the degree of relationship blw 2 variables if the relationship exists.

→ Regression is study relationship blw variables.

Regression equation of 4 on x :-

$$(4 - \bar{4}) = b_{4x}(x - \bar{x})$$

: 4 is dependent variable.

: x is independent variable.

: Used to estimate the value of 4 corresponding to a known 'x' value.

Regression equation of x on 4 :-

: x is dependent variable

: 4 is independent variable

: Used to estimate the value of x correspondi. to known '4' value.

$b_{x4}(4 - \bar{4}) = (x - \bar{x})$

$$(x - \bar{x}) = b_{x4}(4 - \bar{4})$$

b_{x4} = Regression coefficient of x on y.

b_{4x} = Regression " of 4 on x.

$$b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$$

$$\rightarrow r^2 = b_{xy} \cdot b_{yx}$$

$$r = \sqrt{b_{xy} \cdot b_{yx}}$$

→ 'r' is positive, if both b_{xy} & b_{yx} are positive.
→ 'r' is negetive, if both b_{xy} & b_{yx} are negetive
→ In no real life situation, one regression coeff. is +ve and other is negetive.

** The two regression lines x on y & y on x intersect each other at $\bar{x}, \bar{y}$ i.e mean values of x and y.

$$b_{yx} = \frac{N\Sigma xy - \Sigma x \cdot \Sigma y}{N\Sigma x^2 - (\Sigma x)^2}$$

$$b_{xy} = \frac{N\Sigma xy - \Sigma x \cdot \Sigma y}{N\Sigma y^2 - (\Sigma y)^2}$$

Homework:

x: 50 55 65 50 55 60 50 65 70 75
y: 110 110 115 125 140 115 130 120 115 160

Sol:

x	y	R_x	R_y	d	d^2
50	110	8	8.5	-1.5	2.25
55	110	6.5	8.5		
65	115	3.5	7	-3.5	12.25
50	125	8	4	4	16
55	140	6.5	2	4.5	20.25
60	115	5	7	-2	4

50	130	8	3	5	25
65	120	3.5	5	-1.5	2.25
70	115	2	7	-5	25
75	160	1	1	0	0
					111

$= 1 - \frac{6(111)}{990}$

$= 1 - 0.672$

$= 0.328$

2) The coefficient of rank correlation of marks obtained by 10 students in 2 subjects was found to be 0.2. It was later discovered that diff in ranks in 2 subjects of a student was wrongly taken as 9 instead of 7. Find corrected 'r'.

Sol:

Given, $r = 0.2$, $n = 10$

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$0.2 = 1 - \frac{6\Sigma d^2}{990}$$

$$-6\Sigma d^2 = 99 \times 2 - 990$$

$$\Sigma d^2 = \frac{990 - 99 \times 2}{6}$$

$$\Sigma d^2 = 132$$

$\therefore$ Corrected $d^2 = 132 - (9)^2 + (7)^2 = 132 - 81 + 49$

$= 100$

Corrected 'r' $= 1 - \frac{6 \times 100}{990}$

$= 1 - \frac{60}{99} = \frac{39}{99} = 0.394$

Problems of linear regression:

Following table gives age x in years of car and annual maintaince (y) in 100's of Rs. Estimate maintainence cost of 4 year old car after finding regression equations:-

sol:-

x	y	x^2	y^2	xy
1	15	1	225	15
3	18	9	324	54
5	21	25	441	105
7	23	49	529	161
9	22	81	484	198

$\bar{x} = \frac{25}{5} = 5 \; ; \; \bar{y} = 19.8$

$b_{yx} = \frac{N\Sigma xy - \Sigma x \cdot \Sigma y}{N\Sigma x^2 - (\Sigma x)^2} \; ; \; b_{xy} = \frac{N\Sigma xy - \Sigma x \cdot \Sigma y}{N\Sigma y^2 - (\Sigma y)^2}$

$= \frac{5(533) - 25(99)}{5(165) - (625)} \; ; \; = \frac{5(533) - 25(99)}{5(2003) - (9801)}$

$= \frac{2665 - 2475}{200} \qquad \frac{= 190}{214}$

$= \frac{190}{200} = 0.95 \qquad = 0.89.$

Regression line of x on y

$(x - \bar{x}) = b_{xy}(y - \bar{y})$

$(x - 5) = 0.89(y - 19.8)$

Y on x

$(y - \bar{y}) = b_{yx}(x - \bar{x})$

$(y - 19.8) = 0.95(x - 5)$

$\therefore$ Regression line of y on x

$y - 19.8 = (0.95)(x-5)$

Cost of 4-year old $\Rightarrow x = 4$

$y - 19.8 = (0.95)(4-5)$

$y - 19.8 = -0.95$

$y = 19.8 - 0.95$

$\boxed{y = 18.85}$

Maintenence = 18.85 in 100's

18.85×100

$\leqq 1885$ Rs.

Regression line of x on y

$x - 5 = 0.89y - 17.62$

$x = 0.89y - 17.62 + 5$

$x = 0.89(y) - 12.62$

A panel of judges A & B graded 7 debators and independently awarded following marks. If judge-A has awarded 36 marks for 8th debator. find marks given by judge B for 8th debators.

Sol:

Debator	M by A	M by B.
1	40	32
2	34	39
3	28	26
4	30	30
5	44	38
6	38	34
7	31	88

X	Y	X-A	Y-B	$(X-A)^2$	$(Y-B)^2$	$(X-A)(Y-B)$
40	32	10	2	100	4	20
34	39	4	9	16	81	36
28	26	-2	-4	4	16	8
30	30	0	0	0	0	0
44	38	14	8	196	64	112
38	34	8	4	64	16	32
31	28	1	-2	1	4	-2

$$\bar{X} = \frac{\Sigma X}{n} \qquad \bar{Y} = \frac{\Sigma Y}{n}$$

$$\bar{X} = \frac{245}{7} = 35 \qquad = \frac{227}{7} = 32.4$$

$$b_{YX} = \frac{N\Sigma(X-A)(Y-B) - \Sigma(X-A)\cdot\Sigma(Y-B)}{N\Sigma(X-A)^2 - (\Sigma(X-A)^2}$$

$$= \frac{7(206) - (35)(17)}{7(381) - 1125}$$

$$= \frac{1442 - 595}{2667 - 1125} = \frac{847}{1442} = 0.58$$

$$(Y-\bar{Y}) = b_{yx}(X-\bar{X})$$

$$(Y - 32.4) = 0.58(X - 35)$$

$$(Y - 32.4) = 0.58(36 - 35)$$

$$(Y - 32.4) = -0.58$$

$$Y = 32.4 - 0.58$$

$$\Rightarrow Y = 31.82$$

$$\therefore \boxed{Y = 31.82}$$

$(x-A)(y-B)$
20
36
8
0
112
32
-2

1) For two variable x & y, the equations of the regression lines are $9y - x - 288 = 0$ and $x - 4y + 38 = 0$. Find the mean values of x & y.
b) coefficient of correlation blw x & y
c) the ratio of standard deviation of y to x.
d) the most probable value of y when $x = 145$
e) the most probable value of x when $y = 35$

sol: Given regression lines.

$$9y - x - 288 = 0$$
$$x - 4y + 38 = 0$$

a) Solving we get,

$$9y - 4y - 250 = 0$$
$$5y - 250 = 0$$
$$\boxed{y = 50}$$

$$x - 4y + 38 = 0$$
$$x - 200 + 38 = 0$$
$$\boxed{x = 162}$$

b)

$\Rightarrow$ $x - 4y + 38 = 0$
$-4y = -x - 38$
$4y = x + 38$
$y = \frac{1}{4}(x + 38)$
$y = \frac{1}{4}(x - (-38))$

$9y - x - 288 = 0$
$-x = -9y + 288$
$x = 9y - 288$
$x = 9(y - 32)$

$b_{xy} = 9 \quad b_{yx} = \frac{1}{4}$

$r^2 = b_{xy} \cdot b_{yx} = \frac{9}{4}$

$r = \sqrt{\frac{9}{4}} = \frac{3}{2} = 1.5$ (X)

$x - 4y + 38 = 0$
$9y - x - 288 = 0$

$X = 4y - 38$
$X = 4(y - \frac{19}{2})$

$9y = x + 288$
$y = \frac{1}{9}(x + 288)$

$r^2 = \frac{4}{9} \Rightarrow r = \frac{2}{3} = 0.6$

Regression line of x on y

$x - 4y + 38 = 0$

Regression line of y on x $\rightarrow$ $9y - x - 288 = 0$.

$\boxed{r = 0.6667}$

c) Standard deviation $= \sigma$. $\Rightarrow b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$.

$$\frac{\sigma_y}{\sigma_x} = \frac{b_{yx}}{r}$$

$$\frac{\sigma_y}{\sigma_x} = \frac{0.667}{9} = \frac{1}{0.667 \times 9} = 0.1665$$

d) When $x = 145$

$x - 4y + 38 = 0$

$145 - 4y + 38 = 0$

$4y = 145 + 38 \Rightarrow y =$

e) When $y = 35$

$9(35) - x - 288 = 0 \Rightarrow x =$

The following data pertains to the marks in subjects A & B. Mean marks in A = 39.5, Mean marks in B = 47.5. Standard deviation in A = 10.8 S.D of B = 16.8. Coeff of cor (r) = 0.42. Given the estimate marks in B for a candidate who scored 51 marks in A.

Sol:- $\bar{x} = 39.5, \bar{y} = 47.5$

$\sigma_x = 10.8$; $\sigma_y = 16.8$

$r = 0.42$.

$$b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y} \quad ; \quad b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$$

$$= 0.42 \cdot \frac{(10.8)}{16.8} \qquad = 0.42 \frac{(16.8)}{(10.8)}$$

$b_{xy} = 0.27$; $b_{yx} = 0.65$

Given $x = 51$

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$
$$(51 - 39.5) = 0.27(y - 47.5)$$
$$11.5 = 0.27y - 12.825$$
$$0.27y = 24.325$$
$$y = (x)$$

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$y - 47.5 = 0.65(51 - 39.5)$$
$$y - 47.5 = 7.475$$
$$\boxed{y = 54.97}$$

In a partially destroyed records, var(x) = 25, regr. equation of x on y = 5x - y = 22, Reg. eq of y on x is 64x - 45y = 24. Find $\bar{x}$ & $\bar{y}$, r, σ_y, ?

Sol:-

$5x - y = 22$

$64x - 45y = 24$

Multi ① x 45

$225x - 45y = 990$

$64x - 45y = 24$

$161x = 966$

$$\boxed{\bar{x} = 6}$$

$30 - y = 22$

$$\boxed{\bar{y} = 8}$$

$Var(x) = 25$

$\sigma_x = \sqrt{25} = 5$

x on y = $5x - y = 22 \Rightarrow x = \frac{1}{5}(y + 22)$

$$\boxed{b_{xy} = 1/5}$$

Y on $X \Rightarrow 64X - 45Y = 24$

$-45Y = 24 - 64X$

$45Y = 64X - 24$

$Y = \frac{1}{45}(64X - 24)$

$Y = \frac{64}{45}(X - 0.375)$

$Y = 1.42\,(X - 0.375)$

$r^2 = b_{xy} \cdot b_{yx}$

$r^2 = \frac{1.42}{5} = 2.86$

Analyse the travel expenses in Rs Y 102 (Y) same ships and the duration in days (X) of these ships. $\Sigma X = 510$, $\Sigma Y = 7140$, $\Sigma X^2 = 4150$, $\Sigma Y^2 = 740200$, $\Sigma XY = 54900$. find a) the 2 regression equations b) the given trip takes 7 days c) how much money should be be allotted such that he will not run short of money.

Sol:- $\Sigma X = 510$; $\Sigma Y = 7140$

$\Sigma X^2 = 4150$; $\Sigma Y^2 = 740200$. $\Sigma XY = 54900$.

$\bar{X} = \frac{\Sigma X}{n} = \frac{510}{102} = 5$

a) $\bar{Y} = \frac{\Sigma Y}{n} = \frac{7140}{102} = 70$

$b_{XY} = \frac{N\Sigma XY - \Sigma X \cdot \Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2} = \frac{1958400}{N2(740200) - (7140)^2}$

$= 0.7986$

$$b_{yx} = \frac{N\Sigma xy - \Sigma x \cdot \Sigma y}{N\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{102(54900) - (510)(7140)}{102(4150) - (510)^2}$$

$$= \frac{1958400}{423,300 - 260100}$$

$$= \frac{1958400}{163200} = 12.$$

$$r^2 = b_{xy} \cdot b_{yx} = 12 \times 0.7986$$

$$r^2 = 0.9576$$

$$\boxed{r = 0.97}$$

$$x - \bar{x} = b_{xy}(y - \bar{y}) \quad ; \quad y - \bar{y} = b_{yx}(x - \bar{x})$$

$$x - 5 = 0.7986(y - 70) \quad ; \quad y - 70 = 12(x - 05).$$

$$x = 0.798y - 5.086 \qquad y = 12x + 10.$$

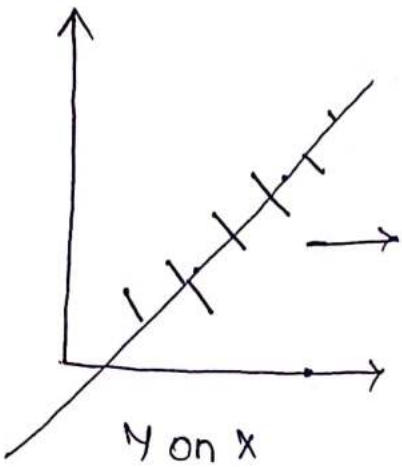
b) Given $x = 7$

y on $x \Rightarrow y = 12x + 10$

$y = 12(7) + 10$

$y = 84 + 10 = 94$

Principle of least squares: (fitting of straight lines)



→ principle of least squares fitting a straight lines.

Y on X	X on Y
$y = a + bx$	$x = a + by$
$\Sigma y = na + b\Sigma x$	$\Sigma x = na + b\Sigma y$
$\Sigma xy = a\Sigma x + b\Sigma x^2$	$\Sigma xy = a\Sigma y + b\Sigma y^2$

Fit a straight line trend for the given data by method of least squares.

YEAR	1979	1980	1981	1982	1983	1984	1985
Output (in cr of Rs)	672	824	968	1205	1464	1758	2058

Sol:-

YEAR	Output	x'	y'	x'^2	$x'y'$
1979	672	-3	-533	9	1599
1980	824	-2	-381	4	762
1981	968	-1	-237	1	237
A (1982)	(1205) B	0	0	0	0
1983	1464	1	259	1	259
1984	1758	2	553	4	1106
1985	2058	3	853	9	2559
		0	514	28	6522

$$\Sigma x'y' = a\Sigma x' + b\Sigma x'^2$$

$\Sigma y' = na + \Sigma x'$

$516 = 7a$

$a = \frac{516}{7}$

$a = 73.428$

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$6522 = 0 + 28b$

$28b = 6522$

$b = 232.92$

$y' = a + bx'$

$y - 1205 = 73.428 + 232.92(x - 1982)$

A group of 5 students took a test before and after training and obtained following scores. Find by method of least squares find the straight line of best fit.

Before training	After training	xy	x^2
3	4	12	9
4	5	20	16
4	6	24	16
6	8	48	36
8	10	80	64

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$\Sigma y = na + \Sigma x$

$33 = 5a + 25b$

$5a = 8$

$a = 8/5$

$a = 1.6$

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$184 = a(25) + b(141)$

$184 = 40 + b(141)$

$b = 1.02$

$y = a + bx$

$y = 1.6 + 1.02x$

$5(5a+25b=33)$

$25a+141b=184$

$\Rightarrow 25a+125b=165$

$25a+141b=184$

$16b=19$

$b=19/16=1.1875$

$5a+25(1.1875)=33$

$5a=33-29.68$

$a=\frac{3.3125}{5}=0.6625$

$\boxed{Y=0.6625+1.1875X}$

Multiple Correlation & Multiple regression:

Multiple correlation:- When values of one variable are associated with (or) influenced by the other variable, linear correlation exist.

When values of one variable are associated with more than 1 variable is M.C.

Ex:- The yield of crop (X_1) is associated with (or) influenced by climate conditions (X_1) quality of seed (X_2), fertilizers (X_3), irrigation (X_4) weather conditions (X_5) e.t.c....

Yule's Notation:- Let X_1, X_2, X_3 be 3 random variables such that $E(X_1)=E(X_2)=E(X_3)=0$

the equation of x_1 depending on x_2, x_3 is $x_1 = a + b_{12.3} X_2 + b_{13.2} X_3$

For notes, 12.18 page of SC Gupta.

Let us consider,

Then the equation of plane of regression on X_1 on X_2, X_3 is $X_1 = a + b_{12.3} X_2 + b_{13.2} X_3$.

Taking expectation on b.s, the plane of regression of X_1 on X_2, X_3 becomes $X_1 = B_{12.3} \cdot X_2 + b_{13.2} X_3$.

The coeff $b_{12.3}, b_{13.2}$ are known as partial regression coefficients of X_1 on X_2 and X_1 on X_3.

Coefficient of Multiple Correlation:

In tri-varient distribution, in which each of variables X_1, X_2 & X_3 has N observations, the multiple correlation coeff of X_1 on X_2, X_3 denoted by $R_{1.23}$ is a simple corr coeff between X_1 and the joint effect of X_2 & X_3 on X_1.

Formula to find $R_{1.23}$

$$R_{1.23} = \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sqrt{\sigma_1^2(\sigma_1^2 - \sigma_{1.23}^2)}}$$

$$R_{1.23}^2 = \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sigma_1^2} \quad ; \quad R_{1.23}^2 = 1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}$$

$$1 - R_{1.23}^2 = \frac{\sigma_{1.23}^2}{\sigma_1^2}$$

$$1 - R_{1.23}^2 = \frac{w}{w_{11}}$$

$$w = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

$$w_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix}$$

Properties of multiple correlation coefficient:

i) MCC measures the closeness of the association between the observed value and the expected values of a variable obtained from multiple linear regression of that variable on other variable.

ii) MCC between observed & expected values, when the expected values are calculated from the linear relation of variables determined the by method of least squares is always greater that of where expected values ase calculated from any other linear combination of variables.

iii) $R_{1.23}$ must lie b/w -1 and $+1$. But $R_{1.23}$ is a non-negetive quantity. $\therefore 0 \leq R_{1.23} \leq 1$.

iv) If $R_{1.23} = 1$, the predicted value of X_1, the multiple linear regression equation of X_1 on

X_2 and X_3 is said to be perfect prediction formula.

v) If $R_{1.23} = 0$; then the total correlation involving X_1 is 0, which means X_1 is completely uncorrelated with all the other values.

vi) $R_{1.23}$ is not less than any total corr. coeff.

$$R_{1.23} \geq R_{1.2}, R_{1.3}, R_{2.3}$$

Example ①:- $R^2_{1.23} = 1 - \frac{w}{w_{11}}$

$$= \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

i) From the data relating to yield of dry bark (X_1), height (X_2) and girth (X_3) of 18 Cinchona plants, the following cc are obtained. $r_{12} = 0.77$ $r_{13} = 0.72$ $r_{23} = 0.52$ find, MCC $R_{1.23}$, $R_{2.13}$, $R_{3.12}$

Sol:- $R^2_{1.23} = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{2.3}}{1 - r_{23}^2}$

$$= \frac{(0.77)^2 + (0.72)^2 - 2(0.77)(0.72)(0.52)}{1 - (0.52)^2}$$

$$= 0.72602$$

$R_{1.23} = 0.8520$

$R^2_{2.13} = \frac{r_{21}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}$

$$= \frac{(0.77)^2 + (0.52)^2 - 2(0.77)(0.72)(0.52)}{1 - (0.72)^2}$$

$$= 0.5953 \Rightarrow R_{2.13} = 0.7716$$

$$R^2_{3.12} = \frac{R_{31}^2 + r_{32}^2 - 2 r_{12} \cdot r_{23} \cdot r_{13}}{1 - r_{12}^2}$$

$$= \frac{(0.72)^2 + (0.52)^2 - 2(0.77)(0.72)(0.52)}{1 - (0.77)^2}$$

$$= 0.52124$$

$$R_{3.12} = 0.7219$$

In a trivarient distribution $\sigma_1 = 2$, $\sigma_2 = 3$, $\sigma_3 = 3$ $r_{12} = 0.7$, $r_{23} = r_{31} = 0.5$. find $R_{1.23}$, $R_{2.13}$, $R_{3.12}$ $b_{12.3}$, $b_{13.2}$, $b_{23.1}$, $\sigma_{1.23}$.

Sol:- $r_{12} = 0.7$, $r_{23} = r_{31} = 0.5$

$$R^2_{1.23} = \frac{r_{12}^2 + r_{13}^2 - 2 r_{12} r_{23} r_{31}}{1 - r_{23}^2}$$

$$= \frac{(0.7)^2 + (0.5)^2 - 2(0.7)(0.5)(0.5)}{1 - (0.5)^2}$$

$$= 0.5199 \Rightarrow \boxed{R_{1.23} = 0.7211}$$

$$R^2_{2.13} = \frac{(0.7)^2 + (0.5)^2 - 2(0.7)(0.5)(0.5)}{1 - (0.5)^2}$$

$$= \frac{0.39}{0.75} = 0.52 \Rightarrow \boxed{R_{2.13} = 0.7211}$$

$$R^2_{3.12} = \frac{(0.5)^2 + (0.5)^2 - 2(0.7)(0.5)(0.5)}{1 - (0.7)^2}$$

$$= \frac{0.15}{0.51} = 0.2941 \Rightarrow \boxed{R_{3.12} = 0.5423}$$

The regression equation of x_1 on x_2 & x_3

$$(x_1-\bar{x}_1)\frac{w_{11}}{\sigma_1}+(x_2-\bar{x}_2)\frac{w_{12}}{\sigma_2}+(x_3-\bar{x}_3)\frac{w_{13}}{\sigma_3}=0$$

where $\bar{x}_1, \bar{x}_2, \bar{x}_3$ are means of x_1, x_2, x_3.

$\sigma_1, \sigma_2, \sigma_3$ are S.D of x_1, x_2, x_3

$$w=\begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} \qquad w_{11}=\begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix}$$

$$w_{12}=-\begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} \qquad w_{13}=\begin{vmatrix} r_{21} & 1 \\ r_{31} & r_{32} \end{vmatrix}$$

x_2 on x_1 & x_3

$$(x_2-\bar{x}_2)\frac{w_{22}}{\sigma_2}+(x_1-\bar{x}_1)\frac{w_{21}}{\sigma_1}+(x_3-\bar{x}_1)\frac{w_{23}}{\sigma_3}=0$$

x_3 on x_1 & x_2

$$(x_3-\bar{x}_3)\frac{w_{33}}{\sigma_3}+(x_1-\bar{x}_1)\frac{w_{31}}{\sigma_1}+(x_2-\bar{x}_2)\frac{w_{32}}{\sigma_2}=0$$

Find w and w_{ij} for $i=j=1,2,3$. Given $r_{12}=0.80$, $r_{23}=-0.56$, $r_{31}=-0.40$.

Sol:
$$\begin{vmatrix} 1 & 0.80 & -0.40 \\ 0.8 & 1 & -0.56 \\ -0.40 & -0.56 & 1 \end{vmatrix}$$

$$1(0.8\times-0.56)-0.8(0.8-0.56\times0.4)-0.4(0.8\times-0.56+0.40).$$

$$=-0.448-0.8(0.576)-0.4(-0.448+0.40)$$

$$=-0.448-0.4608+0.0192=-0.8896=0.2448$$

$$W_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = \begin{vmatrix} 1 & -0.56 \\ -0.56 & 1 \end{vmatrix} = 1-(0.56)^2 = 0.6864$$

$$W_{12} = -\begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = \begin{vmatrix} 0.8 & -0.56 \\ -0.4 & 1 \end{vmatrix} = 0.576$$

$$W_{13} = \begin{vmatrix} r_{21} & 1 \\ r_{31} & r_{32} \end{vmatrix} = \begin{vmatrix} 0.8 & 1 \\ -0.4 & -0.56 \end{vmatrix} = 0.848$$

$$W_{22} = +\begin{vmatrix} 1 & r_{13} \\ r_{31} & 1 \end{vmatrix} = \begin{vmatrix} 1 & -0.4 \\ -0.4 & 1 \end{vmatrix} = \cancel{0.576}\ 0.84$$

$$W_{23} = \begin{vmatrix} 1 & 0.8 \\ -0.4 & -0.56 \end{vmatrix} = -0.24$$

$$W_{21} = -\begin{vmatrix} 0.8 & -0.4 \\ -0.56 & 1 \end{vmatrix} = 0.576$$

$$W_{31} = \begin{vmatrix} 0.8 & -0.4 \\ 1 & -0.56 \end{vmatrix} = -0.848$$

$$W_{32} = \begin{vmatrix} 1 & -0.4 \\ 0.8 & -0.56 \end{vmatrix} = -0.24$$

$$W_{33} = \begin{vmatrix} 1 & 0.8 \\ 0.8 & 1 \end{vmatrix} = 0.36.$$

Find the regression equation of X_1 on X_2 & X_3 for the above data and

Trait	Mean	S.D
X_1	2802	4.42
X_2	4.91	1.10
X_3	594	85

where X_1 represents seed per acre, X_2 represent rainfall in , X_3 represent accumulated

temperature above 42° F.

Sol: $\frac{(x_1-\bar{x}_1)}{\sigma_1}\omega_{11} + \frac{(x_2-\bar{x}_2)}{\sigma_2}\omega_{12} + \frac{(x_3-\bar{x}_3)}{\sigma_3}\omega_{13} = 0$

$$\frac{(x_1-28.02)(0.6864)}{4.42} + \frac{(x_2-4.91)(0.576)}{1.1} +$$

$$\frac{(x_3-5.94)(0.848)}{85} = 0$$

Find the regression equation of x_2 on x_1 & x_3 for the following data:

x_1	35.8	4.2	$r_{12} = 0.6$
x_2	52.4	5.3	$r_{13} = 0.7$
x_3	48.8	6.1	$r_{23} = 0.8$
	(Mean)	(S.D)	

Sol: x_2 on x_1 & x_3

$$\frac{(x_2-\bar{x}_2)}{\sigma_2}\omega_{22} + \frac{(x_1-\bar{x}_1)}{\sigma_1}\omega_{21} + \frac{(x_3-\bar{x}_3)}{\sigma_3}\omega_{23} = 0$$

$$\omega_{22} = \begin{vmatrix} 1 & r_{13} \\ r_{31} & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0.7 \\ 0.7 & 1 \end{vmatrix} = 0.51$$

$$\omega_{21} = \begin{vmatrix} r_{12} & r_{13} \\ r_{32} & 1 \end{vmatrix} = \begin{vmatrix} 0.6 & 0.7 \\ 0.8 & 1 \end{vmatrix} = -0.04$$

$$\omega_{23} = \begin{vmatrix} 1 & r_{12} \\ r_{31} & r_{32} \end{vmatrix} = \begin{vmatrix} 1 & 0.6 \\ 0.7 & 0.8 \end{vmatrix} = -0.38$$

$$\frac{(x_2-52.4)(0.51)}{5.3} + \frac{(x_1-35.8)(-0.04)}{4.2} + \frac{(x_3-48.8)(-0.38)}{6.1} = 0.$$

Time Series:- Time series analysis means analysing time oriented data & forecasting the future values of time series

Forecasting. Forecast is prediction of some future events. this kind of analysis and forecasting is used in finance, economics. Analysis of political & social policy sessions. managing production operation; investigating the impact of human & the policy decision that they make in environment., etc..

Also, forecasting ex-spands many other fields. business, govt, medicines, environment science, social sience, politics etc..

Forecasting problems are classified as....

a) Short-term [days, weeks, months]
b) Medium term [1 (or) 2 years]
c) Long-term [more-than 2 years].

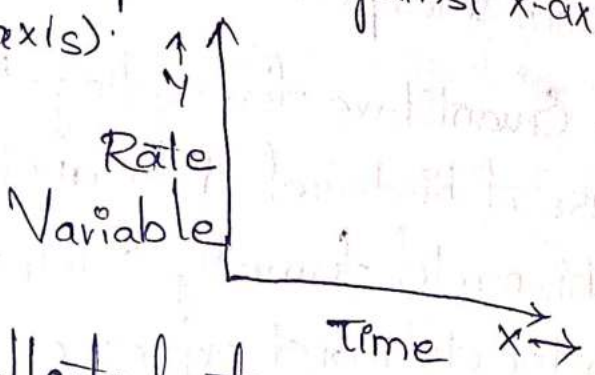
* Short-term and medium-term forecasting situations occur in budgeting selecting new research & development projects e.t.c...

* Long-term forecast is required for strategy planning.

* Time Series is a time oriended (or) chronological sequence of observations on a

variable of interest.

* The graph drawn for this purpose is called time series plot. (Time plotted against x-axis, rate variable on Y-axis).



* The rate variable is collected at equally spaced time periods.
* Time series forecasting utilize daily, weekly monthly, quarterly, (or) annual data.
* Forecasting is important because the prediction of future events is a critical input into many types of planning & decision making process.
* There are 2 broad types of forecasting:
  i) qualitative methods.
  ii) quantitative methods.

* Qualitative methods are subjective in nature & require judgement on part of experts. It is often used in situations where there is little (or) no historical data on which to base the forecast.

Eg: Introduction of new product for which there is no relevant history.

This method makes use of marketing tests, surveys of potential customers and experience with sales performance of other products etc

* Quantitave forecasting techniques make use of historical data and forecasting model. This model formally summarises the pattern in the data and expresses a stastitical relationship blw previous and current values of the variable. Then the model is used to project the patterns in the data in to the future.

* 3 main types of forecasting models:
  i) Regression models (or) Casual forecasting models.
  ii) Smoothing models (fomal models)
  iii) General times of series models (formal

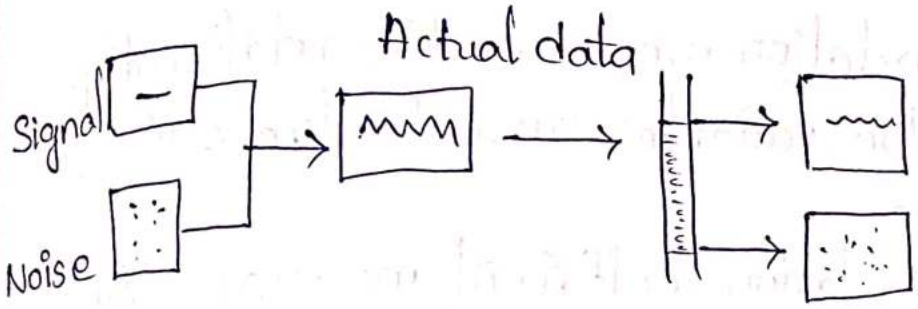
* Regression model make use of relationships between the variable of interest and one more related predicted variable.
Eg: data on house purchases.

* Smoothing models typically employ a simple function of precious observation to provide a forecast of the variable of techniqueto

separate a data set consisting of 2 distinct components - signal and noise.

Signal represents any position pattern caused by the intrinsic dynamics of process from which the data is collected.

A smoother acts as a filter of obtaining an estimate for signal.



General time series: General time series models employ the statistical data properties of the historical data to specify a formal model and then estimate the unknown parameters of this model by least squares.

Continous & Discrete time series:-

Time series is said to be continous when the observations are made continously in time.

Time series is said to be discrete when observations are taken only at specific times usually equally spaced.

Auto-Covariance (And) Auto-correlation :-

A major diagnostic tool which is used

for time series analysis is Acuto-correlation function which helps to describe the evolution of process through time.

* Auto-correlation represents the degree of similarity blw a given time series and a lad version of itself over successive time intervals.
* Auto correlation measures the relationship between the variables current value & its past value.
* Auto correlation coefficient measures the correlation blw the observations at different distances apart. Auto correlation blw successive observations is given by.....

$$r_1 = \frac{\sum_{t=1}^{N-1} (x_t - \bar{x})(x_{t+n} - \bar{x})}{\sum_{t=1}^{N} (x_t - \bar{x})^2}$$

Autocorrelation blw observations from the distances 'k' apart is given by

$$r_k = \frac{\sum_{t=1}^{N-K} (x_t - \bar{x}) \cdot (x_{t+k} - \bar{x})}{\sum_{t=1}^{N} (x_t - \bar{x})^2}$$

r_k = Auto-correlation coefficient at lag k

Auto-covariance:-

Auto-covariance at lag k is given by

$$C_k = \frac{1}{N}\sum_{t=1}^{N-K}(x_t - \bar{x})(x_{t+k} - \bar{x})$$

C_k = Auto co-variance.

$$C_0 = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})^2$$

$$r_k = C_k / C_0, \quad k = 1, 2, 3 \ldots n$$

A useful aid in interpreting a set of autocorrelation coefficient is a graph called a correlogram, in which r_k is plotted against the lag k

Note: ①

Time series plots can reveal patterns such as random, trends, level shifts, periods, cycles, unusual observations like irregular, fluctuating patterns (or) a combination of patterns.

② A very type of time series is a stationary time series. A time series is said to be strictly stationary if its properties are not affected by a change in time origin. This property is called stationarity of time series

1. Identification
2. Estimation
3. forecasting

} Arima Model.

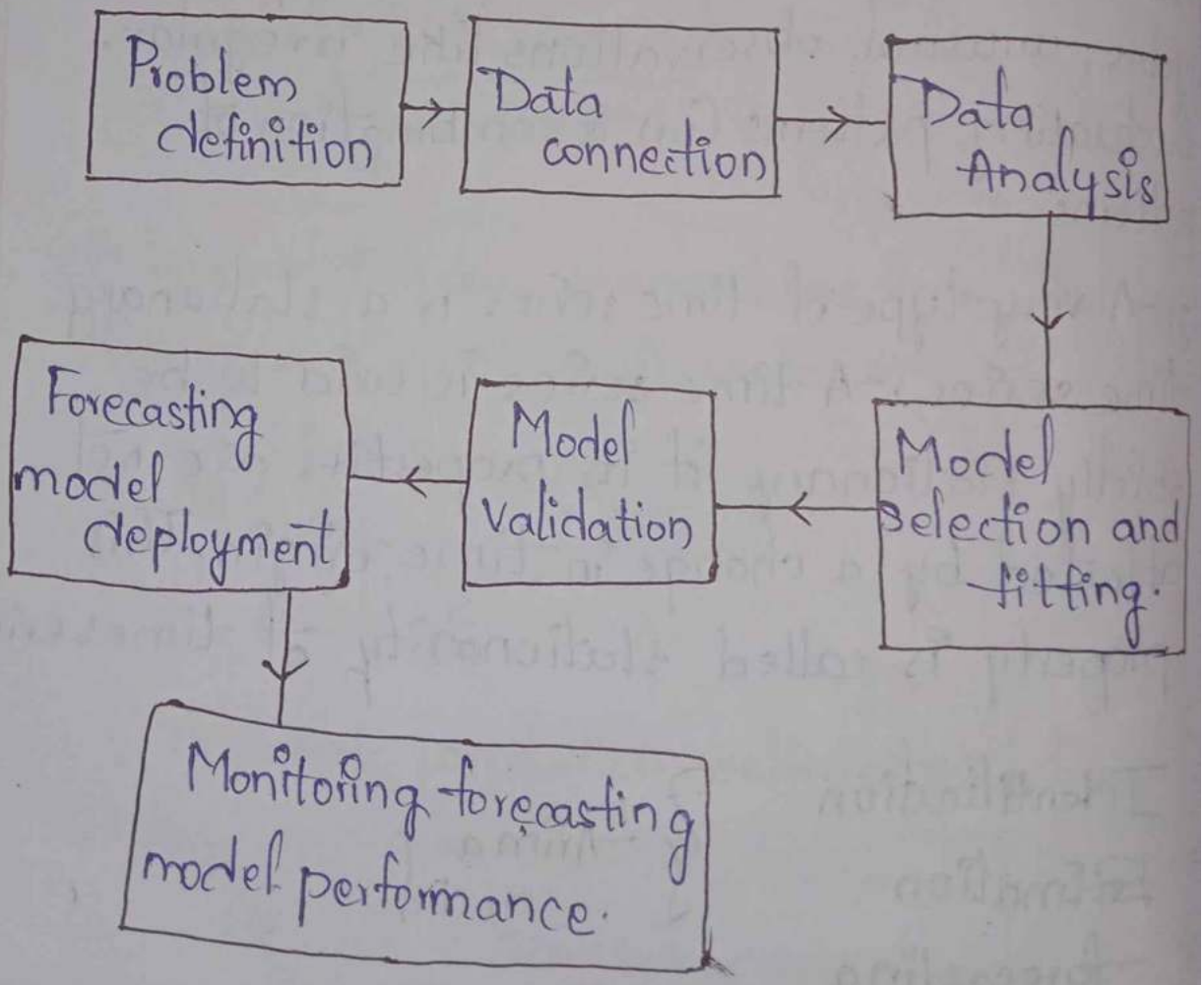
Forecasting process:-

A process is a series of connected activities that transforms one (or) more inputs into one (or) more outputs.

* The activities in forecasting process are:

1. Problem definition
2. Data Connection
3. Data Analysis
4. Model selection and fitting
5. Model validation.
6. Forecasting model deployment.
7. Monitoring forecasting model performance.

" Flowchart for forecasting process "



Arima Models: [ARIMA]

As discussed earlier, there are 3 main types of forecasting techniques. Regression models are casual forecasting models. Forecast Methods based on exponential smoothing may be inefficient & sometimes inappropriate Because they donot take advantage of serial dependence in observations in most efficient way. To formally incorporate this dependent structure, a general class of models called Autoregressive Integrated Moving Average Models [or] ARIMA Models can be considered.

Like other models, this model is used to understand past data (or) credit future data in a series. ARIMA models can capture complex relationships as it takes error terms and observations of lag terms.

Auto Regressive [P]

Past time points of time series data can impact current and future time points. ARIMA models take this concept in to account when forecasting current & future values. ARIMA uses a number of lag observations of time series to forecast

observation. Await is applied is each of past term and the weights can vary based on how recent they are.

AR(x) means x lag terms are going to be use in ARIMA Models. Auto regression is a process of regressing a variable on past values of itself. I

Integrated: [D]

If a trend exist, then timeseries is considered non-stationary and shows seasonality, integrated is a property that reduces seasonality from a timeseries. ARIMA Models have a degree of differencing which eleminates seasonality [Integrated is a property that reduces seasonality from a time series. ARIMA models have a degree of differencing ...] are cycles that repeat regularly over time. Identifying and removing the seasonal component from the time series can result in a clearer relationship blw input and output variables.

Additional information about the seasonal component of time series can provide new information to improve model performance removing seasonality from the time series

is a process and is called seasonal adjustment (or) deseasonalizing. A time series where the seasonal component has been removed is called seasonal (adjustment) stationary.

Moving Average:- (Q)

ARIMA is moving average. Error terms of previous time points are used to predict current and future points observation. Moving average removes non-determinism (or) random movements from a time series. Moving average models have a fixed window and weights are fixed window and weights are relative to that time This implies that the moving average models are more responsive to current event and are more volatile (subject to rapid (or) unexpected change)

P, D, Q are the 3 properties of ARIMA Model.

Time Series model building:-

A three step iterative procedure is used to build an ARIMA Model.

(1.) A tentative model of the ARIMA class identified through analysis of historical data.

(2.) The unknown parameter of the model are estimated.

(3) Through residual analysis, diagnostic checks are performed to determine the adequency of the model, or to indicate the potential improvements.

1. Model identification:-

Forecasting error:- It is understood that forecast is a single number that represents our best estimate of future value of variable of interest. Statistically, this is called as point estimate (or) point forecast. These forecast are almost always wrong. i.e a forecast error is experienced. It is a good practice to accompany a forecast with an estimate of how large a forecast error might be experienced by providing a prediction interval to accompany the point forecast.

The forecast error that results from a forecast of y_t that was made at time period $t-\tau$ is the lead-τ forecast error.

$$e_t(\tau) = y_t - \hat{y}(t-\tau)$$

for $\tau = 1$ $e_t(1) = y_t - \hat{y}_t(t-1)$

The residual is

$$e_t = y_t - \hat{y}_t \text{ where}$$

y_t is the observation and

$\hat{y}_t$ is the fitted value obtained by fitting

a time series model

the ARIMA forecasting equation for a stationary time series is linear (regression type) equation in which the predictions consists of lags of dependent variable and/or lags of forecast error.

Predicted value of y = a constant and/or a weighted sum of one (or) more recent values of y and/or a weighted sum of one or more recent values of errors.

A non-seasonal ARIMA Model is classified as an ARIMA (p,d,q) model where

p is the number of auto-regressive term

d is the number of non-seasonal differences needed for stationarity and

q is the number of lagged forecast error in the prediction equation.

Let y denote the d^{th} difference of y

If $d=0$, $y_t = Y_t$

$d=1$, $y_t = Y_t - Y_{t-1}$

$d=2$, $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$

$= Y_t - 2Y_{t-1} + Y_{t-2}$

In terms of y, the general forecasting equation is

$\hat{y}_t = \mu + \phi y_{t-1} + \cdots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \cdots \theta_q e_{t-q}$

Our aim is to determine the values of p,d,q that are best fit and hence finding $\hat{y}_t$

Note: Identification of the appropriate ARIMA model requires skills obtained by experience.

Some special ARIMA Models:-

ARIMA (1,0,0) - First order auto-regressive model.

ARIMA (0,1,0) - Random walk

ARIMA (1,1,0) - differed I order auto-regressive model.

ARIMA (0,1,1) without constant - Simple exponential smoothing model.

ARIMA (0,1,1) with constant. - Simple exponential smoothing with griwth model.

ARIMA (0,2,1) (or) ARIMA (0,2,2) without constant - linear exponential smoothing model.

ARIMA (1,1,2) without constant. - Damped-trend linear exponential smoothing model.

Parameter estimation:-

There are sereral methods such as method of moments, maximum likelihood

and least squares that can be employed to estimate the parameters in the tentatively identified model. ARIMA models that are non-linear models require the use of non-linear model fitting procedures.

Diagnostic checking :-

After tentative model has been fit to the data, we must examine its adequecy and if necessary suggest potential improvements. This is done through residual analysis.

Hence ARIMA models represent a very powerful and flexible class of models for timeseries & forecasting.

Example 1: From the following data, obtain $R_{1.23}$ and $R_{2.13}$.

X_1	65	72	54	68	55	59	78	58	57	51
X_2	56	58	48	61	50	51	55	48	52	42
X_3	9	11	8	13	10	8	11	10	11	7

Sol:

X_1	X_2	X_3	$(X_1)^2$	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3
65	56	9	4225	3136	81	3640	585	504
72	58	11	5184	3364	121	4176	792	638
54	48	8	2916	2304	64	2592	432	384
68	61	13	4624	3721	169	4148	884	793
55	50	10	3025	2500	100	2750	550	500
59	51	8	3481	2601	64	3009	472	408
78	55	11	6084	3025	121	4290	858	605
58	48	10	3364	2304	100	2784	580	480
57	52	11	3249	2704	121	2964	627	572
51	42	7	2601	1764	49	2142	357	294

$$r_{12} = \frac{N(\Sigma X_1X_2) - (\Sigma X_1)(\Sigma X_2)}{\sqrt{\{N(\Sigma X_1^2) - (\Sigma X_1)^2\}\{N(\Sigma X_2^2) - (\Sigma X_2)^2\}}}$$

$$= \frac{10(32495) - (617)(521)}{\sqrt{(10\times 3873) - (617)(617)} \cdot \sqrt{274230 - (521)^2}}$$

$$= \frac{3493}{\sqrt{(6841)(2789)}} = \frac{3493}{4368.01} = 0.80$$

$$r_{13} = \frac{(10)(6137) - (617)(98)}{\sqrt{(10)(38753) - (617)^2} \cdot \sqrt{(9900) - (98)^2}}$$

a. obtain $R_{1.23}$

57	51
52	42
11	7

X_3	X_2X_3
85	504
92	638
132	384
884	793
550	500
472	408
858	605
580	480
527	572
357	294

$-(521)^2$

To check if $\hat{\theta}$ is sufficient

$$L = (1+\theta)^n \prod_{i=1}^{n} x_i^{\theta}$$

$$= \frac{904}{\sqrt{(6841)(296)}} \cdot \frac{904}{1423} = 0.64.$$

$$r_{23} = \frac{(10)(5178) - (521)(98)}{\sqrt{274230 - (521)^2} \cdot \sqrt{(10)(990) - (98)^2}}$$

$$= \frac{722}{\sqrt{(2789)(296)}} = \frac{722}{908.59} = 0.79.$$

$$R^2_{1.23} = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}$$

$$= \frac{(0.8)^2 + (0.64)^2 - 2 \times 0.8 \times 0.64 \times 0.79}{1 - (0.79)^2}$$

$$= \frac{0.64 + 0.41 + 0.81}{1 - 0.62} = \frac{0.24}{0.38} = 0.65.$$

$$R^2_{1.23} = 0.63$$

$$R_{1.23} = \sqrt{0.63} = 0.79.$$

$$R^2_{2.13} = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{13}^2}$$

$$= \frac{(0.8)^2 + (0.79)^2 - 2(0.8)(0.64)(0.79)}{1 - (0.64)^2}$$

$$= \frac{0.64 + 0.62 - 0.81}{1 - 0.49}$$

$$= \frac{0.45}{0.51} = 0.88$$

$$R^2_{2.13} = 0.88$$

$$\boxed{R_{2.13} = 0.94}$$

Ex:2: From the following data, obtain $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$.

X_1	2	5	7	11
X_2	3	6	10	12
X_3	1	8	6	10

Sol:-

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_2X_3	X_1X_3
2	3	1	4	9	1	6	3	2
5	6	8	25	36	9	30	18	15
7	10	6	49	100	36	70	60	42
11	12	10	121	144	100	132	120	110

$$r_{12} = \frac{(4\times238)-(25)(31)}{\sqrt{(4)(199)-(25)^2}\cdot\sqrt{4(289)-(31)^2}}$$

$$= \frac{174}{\sqrt{(171)(195)}} = \frac{174}{182.61} = 0.94$$

$$r_{13} = \frac{(4)(169)-(25)(20)}{\sqrt{(4\times199)-625}\cdot\sqrt{(4)(146)-(20)^2}}$$

$$= \frac{184\cancel{76}}{\sqrt{\cancel{(195)}(184)}} = \frac{184\cancel{76}}{\cancel{189.42}\ 177.38} = 0.94$$

$$r_{23} = \frac{(4)(201)-(31)(20)}{\sqrt{(4)(289)-(31)^2}\cdot\sqrt{4\times146-(20)(20)}}$$

$$= \frac{184}{\sqrt{(195)(184)}}$$

$$= \frac{184}{189.42} = 0.97$$

$$R_{1.23}^2 = \frac{(0.97)^2 + (0.99)^2 - 2(0.97)(0.99)(0.94)}{1-(0.97)^2}$$

$$= \frac{0.058}{0.059} = 0.98$$

$$\boxed{R_{1.23} = 0.99}$$

$$R_{2.13}^2 = \frac{(0.97)^2 + (0.97)^2 - 2(0.97)(0.99)(0.97)}{1-(0.99)^2}$$

$$= \frac{0.19}{0.20} = 0.95$$

$$R_{3.12}^2 = \frac{(0.99)^2 + (0.97)^2 - 2\times(0.97)(0.99)(0.97)}{1-(0.97)^2}$$

$$= \frac{0.58}{0.591}$$

$$= 0.981$$

$$\therefore R_{3.12}^2 = 0.981$$

$$\boxed{R_{3.12} = 0.99}$$

Ex:③ The following data is given: $R_{1.23}, R_{2.13}, R_{3.12}$

X_1	60	68	50	66	60	55	72	60	62	51
X_2	42	56	45	64	50	55	57	48	56	42
X_3	74	71	78	80	72	62	70	70	76	65

Sol:-

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}\cdot r_{23}\cdot r_{13}}{1-r_{23}^2}$$

$$R_{2.13}^2 = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}\cdot r_{23}\cdot r_{13}}{1-r_{13}^2}$$

$$R_{3.12}^2 = \frac{r_{13}^2 + r_{23}^2 - 2r_{12}\cdot r_{23}\cdot r_{13}}{1-r_{12}^2}$$

Unit-2

Estimation

* There are 2 main classes of stastical inference. One is estimation, second is testing hypothesis.
* Estimation can be point estimation (or) interval estimation.
* A test of hypothesis provides the answers to whether the data support (or) contradict an investigator's claim about value of parameter.

Ex: 136, 143, 147, 151, 158, 160
161, 163, 165, 167, 173, 174
181, 181, 185, 188, 190, 205

[Consider this sample.

Sample mean = $\bar{x}$ = 168.2

Q_1 = 158, Q_2 = median = 166

Q_3 = 181.

Sample S.D = 18.10

For the population just concerning the parameter the μ (pop. mean) we may wish to make following types of inferences.

1) point estimation - estimate μ by a single value.
2) Interval estimation - determine an interval of possible values for μ.

3) Testing hypothesis - determine whether (or) not the mean $\mu = 140$ which is mean value of an alternative material.

Point-estimation:- It concerns with choosing of statistic which is single number calculated from sample data.

Parameter - population mean (μ).

Data - A random sample $x_1, x_2 \cdots x_n$

Estimator - $\bar{x}$

Estimate of standard error $= \frac{s}{\sqrt{n}} = \frac{s}{\sqrt{n}}$

for the previous example, $\bar{x} = 168.2$

Estimated standard error

$$= \frac{s}{\sqrt{n}} = \frac{18.10}{\sqrt{18}} = \frac{18.1}{4.24} = 4.27.$$

Theory of estimation:- let unknown parameter θ take any value on the set (Θ)

Parameter space:- The set Θ which is set of all possible values of θ is called parameter space.

Consider a general family of distributions $f(x, \theta_1, \theta_2 \cdots \theta_k)$. $\theta_i \in \Theta$, $i = 1, 2 \cdots k$. let us consider a random sample $x_1, x_2 \cdots x_n$ of size 'n' from a population with probability

function $f(x, \theta_1 \cdots \theta_n)$ where $\theta_1 \cdots \theta_n$ are unknown population parameters, there will be an infinite no. of functions of sample values called statistics which may be proposed as estimators of one (or) more of parameters.

The best estimate would be one that falls nearest to the true value of parameter to be estimated. The estimating functions are referred to as estimators.

∴ We wish to determine the functions of sample observations

$$T_1 = \hat{\theta}_1(x_1, x_2, x_3 \cdots x_n)$$
$$T_2 = \hat{\theta}_2(x_1, x_2, x_3 \cdots x_n)$$
$$\vdots$$
$$T_n = \hat{\theta}_n(x_1, x_2, x_3 \cdots x_n)$$

such that their distribution is concentrated as closely as possible near the true value of parameter.

Characteristics of good estimaters:-

The following are the criteria that should be satisfied by good estimator.

1. Consistency
2. Unbiasedness
3. Efficiency
4. Sufficiency.

1. **Consistency**: An estimator $T_n = T(x_1, x_2, \ldots x_n)$ based on a random sample of size 'n' is said to be consistent estimator of $\gamma(\theta)$, $\theta \in \Theta$, the parameter space, if T_n converges to $\gamma(\theta)$ in probability i.e if $T_n \xrightarrow{P} \gamma(\theta)$ as $n \to \infty$ (or) T_n is consistent estimator of $\gamma(\theta)$ if $\forall \epsilon > 0, \eta > 0$, there exists a positive integer $n \geq m(\epsilon, \eta) \Rightarrow$

$$P[|T_n - \gamma(\theta)| < \epsilon] \to 1 \text{ as } n \to \infty$$

$$P[|T_n - \gamma(\theta)| < \epsilon] > 1 - \eta \quad \forall\ n \geq m.$$

where m is a very large value of n.

Consistency is a property concerning the behaviour of an estimate for indefinitely large values of sample size 'n', i.e as $n \to \infty$.

Moreover if $\exists$ a consistent estimator (say) T_n of $\gamma(\theta)$, then infinitely many such estimators can be constructed, for eg,

$$T_n' = \left(\frac{n-a}{n-b}\right) T_n = \left(\frac{1 - a/n}{1 - b/n}\right) T_n \to T_n \xrightarrow{P} \gamma(\theta) \text{ as } n \to \infty$$

$\therefore$ for different values of a and b, T_n' is also consistent for $\gamma(\theta)$.

Note: If $X_1, X_2 \ldots X_n$ is a random sample from a population with finite mean.

$E(X_i) = \mu < \infty$, then

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} E(X_i) = \mu$ as $n \to \infty$.

Hence sample mean ($\bar{X}_n$) is always a consistent estimation of population mean μ.

Sufficient conditions for consistency:

Let $\{T_n\}$ be a sequence of estimators such that $\forall\, \theta \in \Theta$,

Expectation of statistic = parameter

(i) $E_\theta(T_n) = \gamma(\theta)$, $n \to \infty$

(ii) $Var_\theta(T_n) \to 0$ as $n \to \infty$, then

T_n is consistent estimator of $\gamma(\theta)$

Unbiasedness:- Unbiasedness is a property associated with finite n. A statistic $T_n = T(x_1, x_2 \ldots x_n)$ is said to be an unbiased estimator of $\gamma(\theta)$ if

$E(T_n) = \gamma(\theta)$ $\forall\, \theta \in \Theta$

If $E(T_n) > \theta$, T_n is said to be positively and if $E(T_n) < \theta$, then it is said to ~~baiased~~ be negetively biased.

The amount of bias $b(\theta) = E(T_n) - \gamma(\theta)$, $\theta \in \Theta$

Examples:

① If $x_1, x_2 \ldots x_n$ is random sample from a normal population $N(\mu, 1)$ show that $t = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ is an unbiased estimate of $\mu^2 + 1$.

Sol: $E(x_i) = \mu$

$Var(x_i) = E(x_i^2) - [E(x_i)]^2$

$1 = E(x_i^2) - \mu^2$

$N(\mu, \sigma)$
$\sigma = 1$
$Var(x) = 1$

$E(x_i^2) = 1+\mu^2$

$E(t) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)$

$= \frac{1}{n}\sum_{i=1}^{n} E(x_i^2)$

$= \frac{1}{n}\sum_{i=1}^{n} (1+\mu^2)$

$= \frac{1}{n}(n)(1+\mu^2)$

$= 1+\mu^2$ $\therefore$ t is an unbiased estimator of $1+\mu^2$.

Eg②: If T is an unbiased estimator of θ, show that T^2 is biased estimator of θ^2.

Sol:- Since T is an unbiased estimation of θ,

$E(T) = \theta$

$Var(T) = E(T^2) - [E(T)]^2$

$= E(T^2) - \theta^2$

$E(T^2) = \theta^2 + Var(T)$, $Var(T) > 0$

Since, $E(T^2) \neq \theta^2$, T^2 is biased estimator for θ^2

Eg:③ If $x_1, x_2 \ldots x_n$ are random observations on a Bernoulli variate x taking the value 1 with probability p and value '0' with probability (1-p), s.t $\frac{\sum x_i}{n}\left(1 - \frac{\sum x_i}{n}\right)$ is consistent estimator of $p(1-p)$.

Sol:- Since $x_1, x_2 \ldots x_n$ are identical, independent distributed Bernoulli Variates, with parameters 'p'

$T = \sum_{i=1}^{n} x_i \sim$ Binomial dist (n, p)

with mean $= np$,

Variance $= npq$

$E(T) = np$

$Var(T) = npq$

Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = T/n$

$E(\bar{x}) = \frac{1}{n}E(T) = \frac{1}{n} \cdot np = p$.

$Var(\bar{x}) = Var(T/n) = \frac{1}{n^2} Var(T) = \frac{1}{n^2}(npq) = \frac{pq}{n}$

$Var(\bar{x}) = \frac{pq}{n} \to 0$ as $n \to \infty$.

$E(\bar{x}) \to p$ as $n \to \infty$
$Var(\bar{x}) \to 0$ as $n \to \infty$
} $\therefore \bar{x}$ is consistent estimators of p.

Also $\frac{\sum x_i}{n}\left(1 - \frac{\sum x_i}{n}\right) = \bar{x}(1-\bar{x})$, being a polynomial in $\bar{x}$ is a continous fn of $\bar{x}$. $\therefore \bar{x}$ is consistent est. of p, $\bar{x}(1-\bar{x})$ is consistent estimator of $p(1-p)$.

3. Efficiency:-

A criterian which is based on the variances of sampling distribution of estimators is known as efficiency.

If for 2 consistent estimators, T_1 T_2 of a certain parameter θ, we have

$V(T_1) < V(T_2) \; \forall \; A$,

then T_1 is more efficient than T_2 for all sample sizes.

Most efficient estimator:- If in a class of consistent estimators for parameter, $\exists$ one whose sampling variance is less than that of any such estimator, it is called the most efficient estimator.

To check if $\hat{\theta}$ is suffici

$$L = (1+\theta)^n \prod_{i=1}^{n} x_i^{\theta}$$

Whenever such an estimator it is called (or) exists, it provides an criterion for measurement of efficiency of the other estimators.

Efficiency: If T_1 is the most efficient estimator with variance V_1 and T_2 is any other estimator with variance V_2, then the efficiency E of T_2 is defined as

$$E = \frac{V_1}{V_2}$$

obviously 'E' cannot exceed 1.

In general if $T, T_1, T_2, \ldots T_n$ are all estimators of $\gamma(\theta)$ and var(T) is minimum, then the efficiency E_i of T_i $(i = 1, 2, \ldots n)$ is defined as

$$E_i = \frac{\text{Var } T}{\text{Var } T_i} \quad i = 1, 2, \ldots n$$

obviously $E_i < 1$, $i = 1, 2 \ldots n$.

Note: In normal samples, since the sample mean $\bar{x}$ is most efficient estimator of μ. the efficiency E of median of such samples (for large n) is..

$$E = \frac{\text{Var}(\bar{x})}{\text{Var}(\text{median})} = \frac{\sigma^2/n}{n\sigma^2/2n} = \frac{2}{\pi} = 0.637.$$

Eg: A random sample $(X_1, X_2 \ldots X_5)$ of size 5 is drawn from a normal population with unknown mean μ. Consider the follg. estimato

to estimate μ.

i) $t_1 = \frac{X_1+X_2+\cdots+X_5}{5}$

ii) $t_2 = \frac{X_1+X_2}{2}+X_3$

iii) $t_3 = \frac{2X_1+X_2+\lambda X_3}{3}$ where λ is $\Rightarrow t_3$ is an unbiased estimator of μ.

a) Find λ b) Are t_1, t_2 unbiased.

c) State giving reasons, the estimator which is best among t_1, t_2 and t_3.

Sol:- Given $E(x_i) = \mu$

$Var(x_i) = \sigma^2$ (say)

$Cov(x_i, x_j) = 0 \ (i \neq j = 1, 2, \cdots n)$

i) $E(t_1) = \frac{1}{5}\sum_{i=1}^{5} E(x_i) = \frac{1}{5}\sum_{i=1}^{5} \mu = \frac{1}{5} \times 5\mu = \mu$

t_1 is an unbiased estimator of μ.

(ii) $E(t_2) = \frac{1}{2}E(X_1+X_2) + E(X_3)$

$= \frac{1}{2}[E(X_1)+E(X_2)] + E(X_3)$

$= \frac{1}{2}(\mu+\mu) + \mu = 2\mu$

t_2 is not an unbiased estimator of μ (Positively biased)

iii) $E(t_3) = \mu$

$\frac{1}{3}E(2X_1+X_2+\lambda X_3) = \mu$

$2E(X_1)+E(X_2)+E(X_3) = \mu X_3$

$2\mu+\mu+\lambda\mu = 3\mu$

$(3+\lambda)\mu = 3\mu$

$\boxed{\lambda = 0}$

$V(G_1) = \frac{1}{25}[V(X_1)+V(X_2)+\cdots+V(X_5)]$

$= 1/25 (5\sigma^2) = 1/5$

$V(t_3) = 1/9 [4V(X_1) + V(X_2)] = 1/9 (4\sigma^2 + \sigma^2) = 5/9\sigma^2 \ (\because \lambda = 0)$

$V(t_1)$ is the least, t_1 is the best estimator of μ.

Eg: X_1, X_2 and X_3 is a random sample of size 3 from a pop with mean value 'H' and var σ^2, $T_1 T_2 T_3$ are the estimators used to estimate the mean value 'H', where

$T_1 = X_1 + X_2 - X_3$

$T_2 = 2X_1 + 3X_3 - 4X_2$

$T_3 = (\lambda X_1 + X_2 + X_3)/3$

{Soln scGupta}

i) Are T_1, T_2 unbiased estimators?

ii) Find λ s.t T_3 is unbiased.

iii) With this value of λ, is T_3 a consistent estimator?

iv) Which is the best estimator.

Minimum variance unbiased estimator (MVUE)

If a statistic $T = T(x_1, x_2 \cdots x_n)$ based on sample of size of 'n' is s

(i) T is unbiased for $\gamma(\theta)$ $\forall \theta \in \Theta$ and

(ii) It has the smallest variance among the class of all unbiased estimates of $\gamma(\theta)$, then T is called the MVUE of $\gamma(\theta)$

In short

T is MVUE of $\gamma(\theta)$ if

$E_\theta(T) = \gamma(\theta) \ \forall \ \theta \in \Theta$

$Var_\theta(T) \leq Var_\theta(T') \ \forall \ \theta \in \Theta$ where T' is

any other unbiased estimator of $\gamma(\theta)$

Note: 1. An MVUE is unique

2. The correlation coefficient between among efficient estimator and any other estimator is the efficiency e is $\sqrt{e}$.

Sufficiency:-

An estimator is said to be sufficient for the parameter, if it contains all the information in the sample regarding the parameter.

If $T = t(x_1, x_2 \ldots x_n)$ is an estimator of a parameter θ, based on a sample $x_1, x_2 \ldots x_n$ of size 'n' from the population in the density $f(x, \theta)$, & if the conditional distribution of $x_1, x_2, \ldots x_n$ given T, is independent of θ, then T is sufficient estimator of θ

Factorisation therom (Neymon's therom):-

The necessary & suff. condition for a distribution to admit sufficient statistic.

$T = t(x)$ is sufficient for 'θ' iff the joint density function L of the sample values can be expressed in the form.

$L = g(t(x))h(x)$ where $g(t(x))$ depends on 'θ' and x only through the value of $t(x)$ and $h(x)$ is independent of θ.

Note:- Joint density fn. of the sample values is the probability density of them.

x \ y	y_1	y_2	...	y_n
x_1	$p(x_1,y_1)$		...	$p(x_1,y_n)$
x_2				
⋮				
x_n	$p(x_n,y_1)$		...	$p(x_n,y_n)$

Points to remember:

1. $f(x) = \frac{1}{2a}$ $\quad a-\theta < x < a+\theta$ depends on θ.

2. The original sample $x = (x_1, x_2 \cdots x_n)$ is always a sufficient statistic.

Invariance property of sufficient Estimation:-

If T is a suff estimation for the parameter θ and if $\psi(T)$ is a 1-1 function of T, then $\psi(T)$ is sufficient for $\psi(\theta)$.

Sol:- Given $f_\theta(x_i) = \begin{cases} \frac{1}{\theta} & 0 \le x_i \le \theta \\ 0 & \text{otherwise} \end{cases}$

Uniform distribution
$f(x) = \frac{1}{b-a}$
$[a,b]$
$0 \to$ otherwise

let $k(a,b) = \begin{cases} 1 & \text{if } a \le b \\ 0 & \text{if } a > b \end{cases}$

then $f_\theta(x_i) = \frac{k(0,x_i) \cdot k(x_i,\theta)}{\theta}$

$\max x_i = t(x)$

$L \; (\text{jdf}) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \left[\frac{k(0,x_i)\, k(x_i,\theta)}{\theta} \right]$

$\frac{k(t(x),\theta)}{\theta^n} = \frac{k(0,\min x_i) \cdot k(\max x_i,\theta)}{\theta^n}$ $\quad 1 \le i \le n$, $1 \le x_i \le n$ $\to 0 \le x_i \le \theta$

$k(0,\min x)$

$f_\theta = \frac{k(0,x_i) \cdot k(x_i,\theta)}{\theta} = \frac{1 \cdot 1}{\theta} = \frac{1}{\theta}$

$\hookrightarrow = g_\theta(t(x)) \cdot h(x)$

where $g_\theta(t(x)) = \frac{k(t(x),\theta)}{\theta^n}$, $f(x) = \max x_i$ $\quad 1 \le i \le n$.

and $h(x) = k\left(0, \min_{1\le i\le n} x_i\right)$

$\therefore$ by factorisation therom.

$T = \max_{1\le i\le n} x_i$ is suff. statistic for θ.

Eg:2 let $x_1, x_2 \cdots x_n$ be a r.s from $N(\mu, \sigma^2)$ pop. find the suff. estimators for μ and σ^2.

Sol: let $\theta = (\mu, \sigma^2)$ $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$.

Then, $L = \prod_{i=1}^{n} f_\theta(x_i) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$

$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[\frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i^2) - 2\mu\Sigma(x_i) + n\mu^2\right)\right]$

$= g_\theta(t(x)) \cdot h(x)$.

where $g(t(x)) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[\frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu t_1(x) + n\mu^2\right)\right.$ ($t_2(x)$)

$t(x) = (t_1(x), t_2(x)) = (\Sigma x_i, \Sigma x_i^2)$ and $h(x) = 1$

Then,

$t_1(x) = \Sigma x_i$ is a sufficient estimator for μ.

$\&\ t_2(x) = \Sigma x_i^2$ is a sufficient estimator for σ^2.

Eg:③ let $x_1, x_2 \cdots x_n$ be a r.s from a population with pdf. $f(x,\theta) = \theta x^{\theta-1}$, $0 < x < 1$, $\theta > 0$

S.T $t_1 = \prod_{i=1}^{n} x_i$ is sufficient for θ.

Sol: $L = \prod_{i=1}^{n} f(x_i, \theta) = \theta^n \prod_{i=1}^{n} x_i^{\theta-1}$

$= \theta^n \prod_{i=1}^{n} x_i^{\theta}$

$\prod_{i=1}^{n} x_i$

$= g(t_1, \theta), h_i(x_1, x_2 \cdots x_n)$ (say)

$\left(\theta^n \prod_{i=1}^{n} x_i^{\theta}\right)\left(\frac{1}{\prod_{i=1}^{n} x_i}\right)$

By factorisation therom;

$\therefore t = \prod_{i=1}^{n} x_i$ is sufficient estimator for θ.

Eg:④ let $x_1, x_2 \cdots x_n$ be a r.s from Cauchy population. $f(x,\theta) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}$ $-\infty < x < \infty$ $-\infty < \theta < \infty$

Examine if T a sufficient statistic for θ.

Sol:- $L(x,\theta) = \prod_{i=1}^{n} f(x_i, \theta)$

$= \frac{1}{\pi^n} \prod_{i=1}^{n} \left[\frac{1}{1+(x_i-\theta)^2}\right]$

$\neq g(t_1, \theta)\, h(x_1, x_2 \cdots x_n)$

$\therefore$ by factorisation then, there is no single statistic, which alone is sufficient est. of θ.

Methods of estimation:-

The some of important methods for good estimation:-

1) Method of Maximum likelihood Estimation (MLE)
2. Method of Minimum Variance
3. Method of Moments
4. Method of least squares
5. Method of Minimum chi-square
6. " " Inverse probability

Method of Maximum likelyhood estimation :

likelyhood function :

Let $x_1, x_2 \cdots x_n$ be a random sample of size 'n' from a population with density function $f(x, \theta)$ then the likelyhood function of sample values $x_1, x_2 \cdots x_n$ denoted by $L = L(\theta)$ is their joint density function given by :

$$L = f(x_1, \theta) \cdot f(x_2, \theta) \cdots f(x_n, \theta)$$

$$= \prod_{i=1}^{n} f(x_i, \theta)$$

* L gives the relative likelyhood that random variables assume the particular set of values $x_1, x_2 \cdots x_n$.
* For a given sample $x_1, x_2 \cdots x_n$, L becomes a function of variable θ, where 'θ' is parameter. The principle of maximum likelyhood consists in finding an estimator for unknown parameters $\theta = (\theta_1, \theta_2 \cdots \theta_n)$ which maximise the likelyhood functions for variation in parameter i.e to find $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \cdots \hat{\theta}_n)$ so that $L = \hat{\theta} > L = \theta$, for all '$\theta$' belonging to Θ

$$L(\hat{\theta}) = \sup L(\theta) \; \forall \; \theta \in \Theta$$

* Thus, if there exists function $\hat{\theta} = \hat{\theta}(x_1, x_2 \cdots x_n)$ of sample values, which maximises 'L' for

variation in θ, then θ is taken as an estimator of θ.

* $\hat{\theta}$ is called the maximum likelyhood estimator
* Since $L>0$ and $\log L$ is non-decreasing function of L, L, $\log L$ attain the extreme values (maximum, minimum) of same value of $\hat{\theta}$.

$$\therefore \frac{1}{L}\frac{dL}{d\theta} = 0$$

$$\frac{d\log l}{d\theta} = 0$$

If θ is a vector valued parameter, then $\hat{\theta}$ is equal to $= \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3 \cdots \hat{\theta}_n$ is given by the solution of simultaneous equations $\frac{\partial \log L}{\partial \theta_i} = 0$

$$= \frac{\partial \log l(\theta_1, \theta_2 \cdots \theta_n)}{\partial \theta_i} = 0 \quad (i = 1 \text{ to } K)$$

①, ② are likelyhood equations for estimating the parameters.

Properties of MLE:

i) The first and second order derivatives $\frac{\partial \log l}{\partial \theta}$ and $\frac{\partial^2 \log L}{\partial \theta^2}$ exist and a continous functions of θ in a range (R) are, for almost all x.

for every θ in R, $\left|\frac{\partial \log L}{\partial \theta}\right| < f_1(x)$

$\left|\partial^2 \log L / \partial \theta^2\right| < f_2(x)$

where, $f_1(x)$, $f_2(x)$ are integrable functions over $-\infty$ to ∞.

ii) The third order derivative $\frac{\partial^3 \log L}{\partial \theta^3}$ exist such that modulus of $\left|\frac{\partial^3 \log L}{\partial \theta^3}\right| < M(x)$ where $E[M(x)] < K$, a positive quantity.

iii) For every θ in R $[\theta \in R]$, $E\left(-\frac{\partial^3 \log L}{\partial \theta^2}\right) = \int_{-\infty}^{\infty} -\frac{\partial^2 \log L}{\partial \theta^2} \cdot L \cdot dx = I(\theta)$ is finite and non-zero.

iv) The range of integration is independent of θ.

v) Variance of maximum likelyhood estimation

$$\hat{\theta} = \frac{1}{I(\theta)} = \frac{1}{E\left[-\frac{\partial^2 \log L}{\partial \theta^2}\right]}$$

Problems on MLE:

1) In a random sample from normal population, $N(\mu, \sigma^2)$. find MLE for i) μ when σ^2 is known ii) σ^2 when μ is known iii) simultaneous estimation of μ and σ^2.

Sol:- $X \sim N(\mu, \sigma^2)$

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$= \frac{1}{\sigma^n (\sqrt{2\pi})^n} \exp\left[-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{1}{(\sigma^2)^{n/2} (2\pi)^{n/2}} \quad //$$

$$\log L = n/2 \log \sigma^2 - n/2 \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2$$

i) when σ^2 is known, the likelyhood eq for estimating μ.

* $\frac{d \log L}{d\mu} = 0$

$$\frac{d \log L}{d\mu} = -0 - 0 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \mu)(-1)$$

$$\frac{d \log L}{d\mu} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \mu)$$

$$= \frac{2}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

$$= \frac{1}{\sigma^2} \left[\sum_{i=1}^{n} (x_i) - n\mu \right]$$

$$\frac{d \log L}{d\mu} = 0$$

$$\frac{1}{\sigma^2} \left[\sum_{i=1}^{n} (x_i) - n\mu \right] = 0 \quad [\sigma^2 \neq 0]$$

$$\sum_{i=1}^{n} x_i - n\mu = 0$$

$$\sum_{i=1}^{n} x_i = n\mu$$

$$\boxed{\hat{\mu} = \sum_{i=1}^{n} \frac{x_i}{n}}$$

ii) $\frac{d \log L}{d\sigma^2} = 0$

$$\Rightarrow \frac{d \log L}{d\sigma^2} = \frac{-n}{2} \cdot \frac{1}{\sigma^2} - 0 + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

$$\frac{n}{2\sigma^2} = \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu^2)$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

MLE $\boxed{\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 =}$

(iii) The likelyhood eq for simultaneous estimation of μ, σ^2 is

$$\frac{\partial \log L}{\partial \mu} = 0 \quad ; \quad \frac{\partial \log L}{\partial \sigma^2} = 0$$

Thus gives

$\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = S^2$

= Sample variance.

= Sample (mean)

Note: $E(\hat{\mu}) = E[\bar{x}] = \mu$

** $E[\hat{\sigma}^2] = E[S^2] \neq \sigma^2$

* The maximum likelyhood estimator need not necessarily be unbiased but MLE is most efficient.

2) Prove that the MLE of the parameter α of a population having density function $\frac{2}{\alpha^2}x$, $0 < x < \alpha$ for sample of unit size is $2x$, x being the sample value. Show also that the estimator is baised.

Sol: $L = \frac{2}{\alpha^2}(\alpha - x)$

$\log L = \log 2 - 2\log \alpha + \log(\alpha - x)$

MLE: $\frac{\partial \log L}{\partial \alpha} = 0$

$= 0 - 2/\alpha + \frac{1}{\alpha - x}(1)$

$= -2(\alpha - x) + \alpha = -2\alpha + 2x + \alpha = -\alpha + 2x = 0$

$\therefore \boxed{\alpha = 2x}$

Hence MLE of α is given by $\hat{\alpha} = 2\bar{x}$

$E(\hat{\alpha}) = E(2\bar{x}) = 2E(\bar{x})$ $\quad E(x) = \int_0^\infty x \cdot f(x) \cdot dx$

$= 2\int_0^\alpha x \cdot f(x,\alpha) \cdot dx$

$= 2\int_0^\alpha x \cdot \frac{2}{\alpha^2}(x-x) \cdot dx$

$= \frac{4}{\alpha^2}\int_0^\alpha x(-x+\alpha) \cdot dx$

$= \frac{4}{\alpha^2}\left[\frac{\alpha x^2}{2} - \frac{x^3}{3}\right]_0^\alpha$

$= \frac{4}{\alpha^2}\left[\frac{\alpha^3}{2} - \frac{\alpha^3}{3}\right] = \frac{4}{\alpha^2} \times \frac{\alpha^3}{6} = 2\alpha/3$

If $E(\hat{\alpha}) = \alpha$

then α is unbiased

Here $E(\hat{\alpha}) = 2\alpha/3$

$\therefore \hat{\alpha}$ is biased.

2) Find the maximum L.F for parameter λ of a poisson distribution on basis of sample of size n.

Sol: p.d.f of poisson distribution with parameter λ is $p[f(x)]$

$\Rightarrow p[X=x] = f(x,\lambda)$

$= \frac{e^{-\lambda} \cdot \lambda^x}{x!}, \quad x = 0,1,2\cdots$

The likelyhood function $(L) = \prod_{i=1}^n f(x_i,\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \cdot \lambda^{x_i}}{x_i!}$

$= \frac{e^{-n\lambda} \cdot \lambda^{\sum_{i=1}^n x_i}}{x_1! \cdot x_2! \cdots x_n!}$

$\log L = \log_e\left[\frac{e^{-n\lambda} \cdot \lambda^{\sum_{i=1}^n x_i}}{x_1! x_2! x_3! \cdots x_n!}\right]$

$= [-n\lambda]\left[\sum_{i=1}^n x_i \log\lambda\right] - [\log x_1! + \log x_2! + \cdots + \log x_n!$

$= (-n\lambda)\left(\sum_{i=1}^n x_i \log\lambda\right) - \sum_{i=1}^n x_i!$

$$\frac{\partial \log L}{\partial \lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0 \qquad \left[\because \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}\right]$$

$$\frac{\partial \log L}{\partial \lambda} = -n + \frac{n\bar{x}}{\lambda}$$

$$\therefore \frac{\partial \log L}{\partial \lambda} = 0$$

$$-n + \frac{n\bar{x}}{\lambda} = 0$$

$$-n\lambda = -n\bar{x}$$

$$\lambda = \bar{x} \Rightarrow \boxed{\hat{\lambda} = \bar{x}}$$

$$* \frac{1}{\text{Variance}(\hat{\lambda})} = E\left[\frac{-\partial^2 \log L}{\partial \lambda^2}\right]$$

$$= E\left[-\frac{\partial}{\partial \lambda}\left(-n + \frac{n\bar{x}}{\lambda}\right)\right]$$

$$= E\left[-\left(-0 - \frac{n\bar{x}}{\lambda^2}\right)\right]$$

$$= E\left[\frac{n\bar{x}}{\lambda^2}\right] = \frac{n}{\lambda^2}\left[E(\bar{x})\right]$$

$$= \frac{n}{\lambda^2}(\lambda) \qquad [\because E(\bar{x}) = \lambda]$$

$$= n/\lambda$$

$$\frac{1}{\text{Var}(\hat{\lambda})} = \frac{n}{\lambda}$$

$$\therefore \boxed{\text{Var}(\hat{\lambda}) = \frac{\lambda}{n}}$$

3) Show that the sample mean $\bar{x}$ is sufficient for estimating the parameter λ of poisson distribution.

$$L = \prod_{i=1}^{n} \frac{e^{-\lambda} \cdot \lambda^{x_i}}{x_i!}$$

$$= \underbrace{\prod_{i=1}^{n} e^{-\lambda} \lambda^{x_i}}_{} \cdot \underbrace{\prod_{i=1}^{n} \frac{1}{x_i!}}_{}$$

$$= g_\lambda(x_i) \cdot h(x_i) \quad \therefore \text{Its a sufficient estimator}$$

Alternative method:- Check its sufficient estimator / not

Check $\frac{\partial \log L}{\partial \lambda}$ whether it's a function of $\bar{x}, \bar{y}$

$$\frac{\partial \log L}{\partial \lambda} = -n + \frac{n\bar{x}}{\lambda}$$

Obtain the M.L.E of θ in $f(x,\theta) = (1+\theta)x^{\theta}$, $0<x<1$ based on an independent sample of size n. Also estimate whether this estimate is sufficient for θ.

Sol:

$$L = \prod_{i=1}^{n} f(x_i, \theta)$$

$$L = \prod_{i=1}^{n} (1+\theta) x_i^{\theta} \quad [\because 0<x_i<1]$$

$$L = (1+\theta)^n \prod_{i=1}^{n} x_i^{\theta}$$

$$\log L = n \log(1+\theta) + \theta \sum_{i=1}^{n} \log x_i$$

$$\frac{\partial \log L}{\partial \theta} = n\left(\frac{1}{1+\theta}\right) + \sum_{i=1}^{n} \log x_i$$

$$\therefore \frac{\partial \log L}{\partial \theta} = 0$$

$$\frac{n}{1+\theta} + \sum_{i=1}^{n} \log x_i = 0$$

$$\frac{n + (1+\theta)\sum_{i=1}^{n} \log x_i = 0}{(1+\theta)}$$

$$\Rightarrow n + (1+\theta)\sum_{i=1}^{n} \log x_i = 0$$

$$n + \sum_{i=1}^{n} \log x_i + \theta \sum \log x_i = 0$$

$$\theta = \frac{-n - \sum_{i=1}^{n} \log x_i}{\sum_{i=1}^{n} \log x_i}$$

$\therefore$ By M.L.E

$$\hat{\theta} = \frac{-n}{\sum_{i=1}^{n} \log x_i} - 1 = \frac{-n}{\log \prod_{i=1}^{n} x_i} - 1$$

To check if θ is sufficient

$$L = (1+\theta)^n \prod_{i=1}^{n} x_i^{\theta}$$

$$L = \underbrace{(1+\theta)^n \prod_{i=1}^{n} x_i^{\theta-1}}_{g_\theta(x_i)} \cdot \underbrace{\prod_{i=1}^{n} x_i}_{h(x_i)}$$

Test of hypothesis:

Stastical inference can be divided to 2 types:

1) Estimation
2) Test of statistical hypothesis

* A statistical hypothesis is some statement (or) assertion about a population (or) equivalently about the probability distribution characterising a population which we want to verify on basis of information available from sample.

* If statistical hypothesis specifies a population completely then it is termed as simple statistical hypothesis, otherwise a complete statistical hypothesis.

Ex: If $x_1, x_2 \cdots x_n$ is random sample of size 'n' from a normal population of mean μ and variance σ^2 then $H_0 : \mu = \mu_0, \sigma^2 = \sigma_0^2$ is a simple hypothesis.

whereas i) $\mu = \mu_0$
ii) $\sigma^2 = \sigma_0^2$
iii) $\mu < \mu_0, \sigma^2 = \sigma_0^2$
iv) $\mu > \mu_0, \sigma^2 = \sigma_0^2$
v) $\mu = \mu_0, \sigma^2 < \sigma_0^2$
vi) $\mu < \mu_0, \sigma^2 > \sigma_0^2$

are composite hypothesis.

To check if θ is sufficient

$$L = (1+\theta)^n \prod_{i=1}^{n} x_i^{\theta}$$

$$L = \underbrace{(1+\theta)^n \prod_{i=1}^{n} x_i^{\theta-1}}_{g_\theta(x_i)} \cdot \underbrace{\prod_{i=1}^{n} x_i}_{h(x_i)}$$

Test of hypothesis:

Stastical inference can be divided to 2 types:

1) Estimation
2) Test of statistical hypothesis

* A statistical hypothesis is some statement (or) assertion about a population (or) equivalently about the probability distribution characterising a population which we want to verify on basis of information available from sample.

* If statistical hypothesis specifies a population completely then it is termed as simple statistical hypothesis, otherwise a complete statistical hypothesis.

Ex: If $x_1, x_2 \cdots x_n$ is random sample of size 'n' from a normal population of mean μ and variance σ^2 then $H_0 : \mu = \mu_0, \sigma^2 = \sigma_0^2$ is a simple hypothesis.

whereas i) $\mu = \mu_0$
ii) $\sigma^2 = \sigma_0^2$
iii) $\mu < \mu_0, \sigma^2 = \sigma_0^2$
iv) $\mu > \mu_0, \sigma^2 = \sigma_0^2$
v) $\mu = \mu_0, \sigma^2 < \sigma_0^2$
vi) $\mu < \mu_0, \sigma^2 > \sigma_0^2$

are composite hypothesis.

Critical Region:-

Let $x_1, x_2 \ldots x_n$ be the sample observations denoted by 'o'. All the values of 'o' will be aggregate of sample and they constitute a space, called the sample space, denoted by 's'.

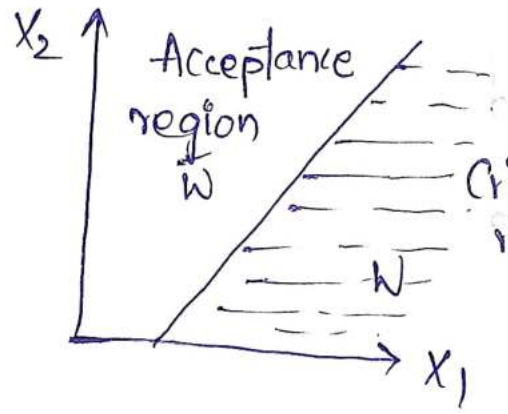
Since the sample values $x_1, x_2 \ldots x_n$ can be taken as point in n-dim space, we specify some region of n-dim space and see whether this point lies with in this region (or) outside the region. We divide the s.s 's' into & disjoint W, W'. $(W' = S - W)$

If the sample points $\in W$, Reject H_0 (accept H_1).

If sample points $\in W'$, Reject H_1 (accept H_0).

The region of rejection of H_0 when H_0 is true is that region of outcome set where H_0 is rejected if sample point falls in that region, and is called as critical region.

Eg:- 2-D



Sample of size (2)

If sample point falls in subset W, H_0 is rejected Otherwise H_0 accept

2-types of errors:-

1. Type I - error.
2. Type II - error.

	Accept H_0	Reject H_0
H_0		
True	✓	Type-I error ✗
False	✗ Type II error	✓

The error of rejecting H_0 (accepting H_1) when H_0 is true is called type-I error.

The error of accepting H_0, when H_0 is false is called type-II error (rejecting H_0).

α = probability of type-I error (producer's risk).

β = probability of type II error (consumer's risk).

'α', the probability of type-I error is known as level of significance of test (or) size of the critical region.

'$1-\beta$' is called the power function of test hypothesis, H_0 against alternative hypothesis H_1

The value of power function at a parameter point is called the power of test at that point.

* Note: An ideal test keeps both the errors under control.

Steps in solving testing of hypothesis problem:-

1. Explicit knowledge about pop, distribution and its parameters. (population)
2. H_0, H_1 setting up.
3. The choice of suitable statistic $t = (x_1, x_2 \cdots x_n)$ called the test statistic.
4. $W, \bar{W}$ (rej reg, Acc. region respectively).
   (i) Rej H_0 (acc H_1) if the value of 't' falls in W.
   (ii) Acc H_0 (Rej H_1) if the value of 't' falls in $\bar{W}$.
5. Compute experimental sample observation, action compute the appropriate test statistic and take.

Optimum test under different situations :-

2 steps:-

1. Choice of 't' 2. choice of W.

Control α, Minimum type II error (β).

α - p of type-I error

β - p of type-II error.

Symbolically $p(x \in W/H_0) = \alpha$ where

$x = (x_1, x_2 \cdots x_n)$

$\Rightarrow \int_W L_0 \, dx = \alpha$ where L_0 is the likelyhood function of sample observation under H_0 and $\rightarrow (1)$

$\int dx$ is the n-fold integral $\iiint \cdots \int dx_1 dx_2 \cdots dx_n$

Also, $P(x \in \bar{W} / H_1) = \beta$

$\int_{\bar{W}} L_1 \cdot dx = \beta \rightarrow (2)$ where L_1 is the likelyhood function of sample observation under H_1.

$\int_W L_1 \cdot dx + \int_{\bar{W}} L_1 \cdot dx = 1$

$\int_W L_1 \cdot dx = 1 - \int_{\bar{W}} L_1 \cdot dx$

$P(x \in W / H_1) = 1 - \beta \rightarrow (3)$

Most Powerful test (MP) of level α:

Let $H_0: \theta = \theta_0$

$H_1: \theta = \theta_1$

The critical region W is the most powerful (MP) critical region of size 'α' for testing $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$ if

$P(x \in W / H_0) = \int_W L_0 \cdot dx = \alpha$

and $P(x \in W / H_1) \geq P(x \in W_1 / H_1)$ $\forall$ critical region W_1.

Uniformly most powerful test (UMP test) of level α:

The region W is called UMP critical region of size 'α' for testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$ i.e $H_1 = \theta = \theta_1 \neq \theta_0$ if

$P(x \in W / H_0) = \int_W L_0 \cdot dx = \alpha$

and $P(x \in W / H_1) \geq P(x \in W_1 / H_1)$ $\forall$ $\theta \neq \theta_0$

whatever the region W_1 may be.

To check

$(1+\theta)^n \; \prod x_i$: 1 . n .

Neymann - Pearson Lemma :-

Let $k>0$, be a constant and w be a critical region of size α such that

$$w = \left\{ x \in S : \frac{f(x,\theta_1)}{f(x,\theta_0)} > k \right\}$$

$$w = \left\{ x \in S : \frac{L_1}{L_0} > k \right\}$$

and $\bar{w} = \left\{ x \in S \; \frac{L_1}{L_0} \leq k \right\}$

* where L_0 and L_1 are the likelyhood function of the sample observations $x \in (x_1, x_2, \ldots x_n)$ under H_0 and H_1 respectively. Then w is most powerful critical region of the test hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$

* Proof :- Given $P(x \in w / H_0) = \int_w L_0 \cdot dx = \alpha$

The power of region $1-\beta$ is

$$P(x \in w / H_1) = \int_w L_1 \cdot dx = 1-\beta$$

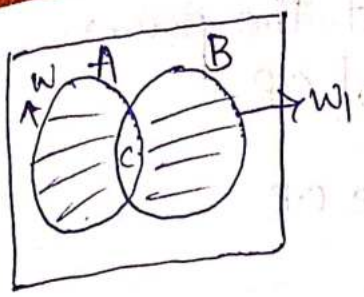
In order to prove the Lemma, we have to prove that $\exists$ no other critical region, of size $\leq \alpha$ which is more powerful than w.

Let w_1 be another C.R of size $\alpha_1 \leq \alpha$ and power $1-\beta_1$, $\ni$

$$P(x \in w_1 / H_0) = \int_{w_1} L_0 \cdot dx = \alpha_1$$

$$P(x \in w_1 / H_1) = \int_{w_1} L_1 \cdot dx = 1-\beta_1$$

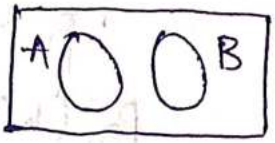
we need to prt $1-\beta \geq 1-\beta_1$



Let $W = A \cup C$

$W_1 = B \cup C$

C may be empty in 'W' and W_1 are disjoint.



If $\alpha_1 \leq \alpha$;

$$\int_{W_1} L_0 \cdot dx \leq \int_{W} L_0 \cdot dx$$

$$\int_{B \cup C} L_0 \cdot dx \leq \int_{A \cup C} L_0 \cdot dx$$

$$\int_{B} L_0 \cdot dx \leq \int_{A} L_0 \cdot dx$$

$$\int_{A} L_0 \cdot dx \geq \int_{B} L_0 \cdot dx$$

Since $A \subset W$, $\int_A L_1 \cdot dx > K \int_A L_0 \cdot dx \geq K \int_B L_0 \cdot dx$ → (+)

Also, w.k.t

$\bar{W}$ $\quad \frac{L_1}{L_0} \leq K \quad \forall \; x \in \bar{W}$

$$\int_{\bar{W}} L_1 \cdot dx \leq K \cdot \int_{\bar{W}} L_0 \cdot dx$$

This holds good for any subset of $\bar{W}$.

$\therefore$ for $\bar{W} \cap W_1 = B$

$\therefore \int_B L_1 \cdot dx \leq K \int_B L_0 \cdot dx \leq \int_A L_1 \cdot dx$ (By (*))

Adding $\int_C L_1 \cdot dx$ on both sides, we get

$$\int_{B \cup C} L_1 \cdot dx \leq \int_{A \cup C} L_1 \cdot dx$$

$$\int_{W_1} L_1 \cdot dx \leq \int_{W} L_1 \cdot dx$$

$$1 - \beta_1 \leq 1 - \beta$$

$1 - \beta \geq 1 - \beta_1$ → Hence proved.

"Unbaised test and unbiased critical region"

The C.R 'W' and consequently the test based

on it is said to be unbiased if the power of the test exceeds the size of C.R.

power of the test $\geq$ size of the C.R

$1-\beta \geq \alpha$

$P_{\theta_1}(w) \geq P_{\theta_0}(w)$

$P(x: x \in w/H_1) \geq P(x: x \in w/H_0)$

$P_\theta(w) \geq P_{\theta_0}(w) \quad \forall\ \theta (\neq \theta_0) \in \Theta$

Theorem: Every MP (or) UMP CR is necessarily unbiased.

Optimum regions & sufficient statistics:-

T in S.S if $L(x,\theta) = \prod_{i=1}^{n} f(x_i,\theta) = g_\theta(t(x)).h(x) \longrightarrow (1)$

where $g_\theta(t(x))$ is the marginal distribution of the statistics $T=t(x)$

By N.P Lemma,

the most powerful C.R is,

$w\{x: L(x,\theta_1) \geq k L(x,\theta_0)\} \forall k>0 \longrightarrow (2)$

From (1) & (2)

$w = \{x: g_{\theta_1}(t(x)).h(x) \geq k.g_{\theta_0}(t(x)).h(x)\} \forall_{k>0}$

$= \{x: g_{\theta_1}(t(x)) \geq k g_{\theta_0}(t(x))\} \forall k>0$

$\therefore$ If $T=t(x)$ is sufficient statistics for θ then MPCR may be defined in-term of marginal distribution of $T=t(x)$.

(Rather than the joint dist. of $X_1, X_2 \cdots X_n$)

N.P Lemma application

Examples:-

Given the freq. fn:

$$f(x,\theta) = \frac{1}{\theta}, \quad 0 \le x \le \theta$$
$$= 0 \quad \text{elsewhere.}$$

1. Testing, $H_0: \theta = 1$ against $H_1: \theta = 2$ by means of a single observed value of 'x', what would be the sizes of type I and type II errors if be the sizes of type I and type II errors if (i) $0.5 \le x$ (ii) $1 \le x \le 1.5$ as the critical regions? Also obtain the power function of test.

Sol: $H_0: \theta = 1$ against $H_1: \theta = 2$

(i) $W = \{x : 0.5 \le x\} = \{x : x \ge 0.5\}$

$\overline{W} = \{x : x \le 0.5\}$

$\alpha = P\{x \in W / H_0\} = P\{x \ge 0.5 / \theta = 1\}$

$P = \{0.5 \le x \le \theta = 1/\theta\}$

$= P\{0.5 \le x \le 1/\theta = 1\}$

$$= \int_{0.5}^{1} [f(x,\theta)]_{\theta=1} dx = \int_{0.5}^{1} 1 \cdot dx = 0.5$$

$$\boxed{\alpha = 0.5}$$

Wkt $\beta = P\{x \in \overline{W} / H_1\}$

$= P\{x \le 0.5 / \theta \in 2\}$

$$= \int_{0}^{0.5} [f(x,\theta)]_{\theta=2} \cdot dx$$

$$= \int_{0}^{0.5} \frac{1}{2} \cdot dx = 0.25 \Rightarrow \boxed{\beta = 0.25}$$

ii) $W = \{x : 1 \le x \le 1.5\}$

$\alpha = P\{x \in W / \theta = 1\}$

$\alpha = \int_1^{1.5} [f(x,\theta)]_{\theta=1} \, dx = 0$

$\beta = P\{x \in \bar{W} / \theta = 2\} = 1 - P\{x \in W / \theta = 2\}$

$= 1 - \int_1^{1.5} [f(x,\theta)]_{\theta=2} \, dx$

$= 1 - \left(\frac{x}{2}\right)_1^{1.5}$

$= 0.75$

Power fn $(1-\beta) = 1 - 0.75 = 0.25$

Example 2:- If $x \ge 1$ is the C.R for testing $H_0 : \theta = 2$ against $H_1 : \theta = 1$ on the basis of the single observation from the population,

$f(x,\theta) = \theta e^{-\theta x}, \quad 0 \le x < \infty$

Find the values of α and β. (Type I, II errors)

Sol:- $W = \{x : x \ge 1\}$; $\bar{W} = \{x : x < 1\}$

$H_0 : \theta = 2$, $H_1 : \theta = 1$

$\alpha = P\{x \in W / H_0\} = P\{x \ge 1 / \theta = 2\}$

$= \int_1^{\infty} [f(x,\theta)]_{\theta=2} \, dx$

$= 2 \int_1^{\infty} e^{-2x} \, dx = 1/e^2$

$\beta = P\{x \in \bar{W} / H_1\}$

$= P\{x < 1 / \theta = 1\}$

$= \int_0^1 e^{-x} \, dx = \frac{e-1}{e}$.

Example 3:- Let p be the probability that a coin will fall head in a single toss in order to test $H_0: p = 1/2$ against $H_1: p = 3/4$. The coin is tossed 5 times and H_0 is rejected if more than 3 heads are obtained. Find α and power of test. [P of type I error] [1-β].

Sol:- Here $H_0: p = 1/2$
$H_1: p = 3/4$

Let the R.V. X denote the no. of heads in 'n' tossess of coin then $X \sim B(n,p) \Rightarrow P(X=x)$

$= {}^nC_r p^r . q^{n-r}$

$= {}^5C_r p^r (1-p)^{5-r}$

Critical reg.

$W = \{x : x \geq 4\} \Rightarrow \bar{W}\{x : x \leq 3\}$

α = P. eq type I error

$= P\{x \geq 4 / H_0\}$

$= P\{x=4/p=1/2\} + P\{x=5/p=1/2\}$

$= {}^5C_4 (1/2)^4 (1/2)^1 + {}^5C_5 (1/2)^5$

$= 3/16$.

β = prob. of type II error

$P\{x \in \bar{W} / H_1\} = 1 - P\{x \in W / H_1\}$

$= 1 - [P\{x=4/p \in 3/4\} + P\{x=5/p=3/4\}]$

$= 1 - [{}^5C_4 (3/4)^4 (1/4)^1 + {}^5C_5 (3/4)^5] = 1 - \frac{81}{128} = \frac{47}{128}$

power of test $1-\beta = 8/125$

Example 4:- let $X \sim N(\mu, 4)$, μ unknown. To test $H_0: \mu = -1$ against $H_1: \mu = 1$ based on a sample of size 10 from this population, we use the C.R $x_1 + 2x_2 + \cdots + 10x_{10} \geq 0$ ... What is its size? What is the power of test?

Sol:- C.R $W = \{x : x_1 + 2x_2 + \cdots + 10x_{10}\} \geq 0\}$

let $U = x_1 + 2x_2 + \cdots + 10x_{10}$.

$U \sim N\left((1+2+\cdots+10)\mu_i, (1^2+2^2+\cdots+10^2)\sigma^2\right]$

$= N(55\mu, 385\sigma^2)$

$U \sim N(55\mu, 385 \times 4) = N(55\mu, 1540)$

$\alpha = P\{x \in W / H_0\} = P(U \geq 0 / H_0)$

$H_0: \mu = -1$, $U \sim N(-55, 1540)$

$\neq N(1,0)$ $Z = \dfrac{U - E(U)}{\sigma_U} = \dfrac{U+55}{\sqrt{1540}}$

$N(1,0)$ – Standard Normal population.

Under H_0, when $U = 0$, $Z = 1.4015$

$$Z = \frac{55}{\sqrt{1540}} = \frac{55}{39.2420} = 1.4015$$

$\alpha = P(Z \geq 1.4015)$

$= 0.5 - P(0 \leq Z \leq 1.4015)$ [Use normal dist. table]

$= 0.5 - 0.4192$

$= 0.0808$

$1-\beta = P(x \in W/H_1) = P(U \geq 0/H_1)$

$H_1 = H=1,\ U \sim N(55; 1540)$

$$2 = \frac{U-[z(0)]}{\sigma(4)} = \frac{-55}{\sqrt{1540}} = -1.4015$$

$1-\beta = P(z \geq -1.40)$

$= P(-1.40 \leq z \leq 0) + 0.5$

$= P(0 \leq z \leq 1.40) + 0.5$

$= 0.4192 + 0.5$

$= 0.9192$

Analysis of Variance (ANOVA)

* ANOVA is a technique that enable to test for the "significance of the difference among more than 2 sample means". This technique was introduced by R.A. Fisher. this technique was originally used in agriculture experiments in which different types of fertilizers were applied to plots of land, different types of feeding methods of animals e.t.c...
* This technique is widely used in different fields, for example to study pattern of average sales by using different sales technique; types of drug manufacturers by different companies to cure a disease e.t.c.
* ANOVA is used to test the significance difference between more than 2 sample mean as well as more than sample variance.

Assumptions of ANOVA:-

* The samples are independently drawn from the population. [Independence of obs]
* The population is normally distributed.
* The variances of all populations are equal. [Homogenity of Variance]

Classifications of ANOVA:-

ANOVA can be divided in to 2 types:

i) One-way ANOVA

ii) Two-way ANOVA

One-way ANOVA:-

In one-way classification, the observations are classified as one factor. this is done as column wise.

Two-way ANOVA:

In two-way classification, the observations are classified as 2 factors., where one is exhibited column and other row-wise

One-way ANOVA:-

Algorithm:-

1) Fix Null (H_0) and Alternate (H_1) hypothesis.

2) let the given samples be $x_1, x_2, x_3 \ldots$

3) Calculate $x_1^2, x_2^2, x_3^2 \ldots$

4) Calculate $T = \Sigma x_1 + \Sigma x_2 + \Sigma x_3 \ldots$

5) Calculate total correlation factor (CF):

$$CF = \frac{T^2}{N}, \quad N = \Sigma n_i.$$

6) Calculate SST [Total sum of squares]

$$SST = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 - \frac{T^2}{N}$$

$$= \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 - CF$$

7) Calculate SSC [Sum of square between samples]

$$SSC = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \cdots - \frac{T.^2}{N}.$$

8) Calculate SSE (Sum of squares for error)

$$SSE = SST - SSC.$$

9) Prepare 1-way ANOVA table with MSC → Sample mean square.
MSE → Error mean Square.
let total no. of columns be = K.

Sources of Variations	Sum of Squares	Degrees of freedom	Mean Square	F
Between Samples	SSC	K-1	MSC = SSC/K-1.	
Error	SSE	N-K	MSE = SSE/N-K	MSC/MSE
Total	SST	N-1		

10) Identify the table value of F with degree of freedom.

$$F_\alpha = F(1-\alpha, K-1, N-K)$$

11) If the calculated value is < table value.
Accept null hypothesis.
Else Reject Null hypothesis.

Question①: A consumer Agency wanted to find out if the mean weight loss for each of 3 types of drugs for loosing weight is the same. The following table records weight loss per kg for 15 people after taking drugs for the three months. Using 10% level of significance will you conclude that mean weight loss is same for each of these 3 drugs.

Let:	Drug 1	Drug 2	Drug 3
	7.5	9.5	8.5
	10	9	10
	8	7.5	6.5
	6	10	11
Sol:	6.5	6	8

let, means be denoted as: M_1, M_2, M_3.

H_0: There is no significance difference between the means.

H_0: $M_1 = M_2 = M_3$.

H_1: Atleast one of above is not equal.

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2
7.5	9.5	8.5	56.25	90.25	72.25
10	9	10	100	81	100
8	7.5	6.5	64	56.25	42.25
6	10	11	36	100	121
6.5	6	8	42.25	36	64
38	42	44	298.5	363.75	399.5

$L = (1+\theta)^n \prod_{i=1}^{n} x_i$

$(\Sigma x_1)^2 = (38)^2 = 1444$

$(\Sigma x_2)^2 = (42)^2 = 1764$

$(\Sigma x_3)^2 = (44)^2 = 1936$

$T = \Sigma x_1 + \Sigma x_2 + \Sigma x_3 = 38 + 42 + 44 = 124$

Correction factor $= \frac{T^2}{N}$

$n_1 = 5 : n_2 = 5, n_3 = 5$

$C.F = \frac{(124)^2}{5+5+5} = 1025.066$

$SST = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 - \frac{T^2}{N}$

$= 298.5 + 363.75 + 399.5 - 1025.066$

$= 36.434$

$SSC = \frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n_2} + \frac{(\Sigma x_3)^2}{n_3} - \frac{T^2}{N}$

$= \frac{(1444) + 1764 + 1936}{5} - 1025.066$

$= 1028.8 - 1025.066$

$= 3.734$

Ex:

$SSE = SST - SSC$

$= 36.434 - 3.734$

$= 32.7$

K = No. of columns

K = 3

$MSC = \frac{SSC}{K-1}$

$= \frac{3.734}{3-1}$

$= \frac{3.734}{2}$

$= 1.867$

$MSE = \frac{SSE}{N-K}$

$= \frac{32.7}{15-3}$

$= \frac{32.7}{12}$

$= 2.725$

Source of variations	Sum of squares	Degree of freedom	Mean square	F
Between samples	SSC = 3.734	2	MSC = 1.864	MSC/MSE = 0.68 51
Error	SSE = 32.7	12	MSE = 2.725	
Total	SST = 36.434	14		

Calculated value of F = 0.6851

Table value = 0.281

$0.6851 < 0.281$

Accept H_0. Reject H_1.

∴ We conclude that mean weight loss is same for each of the 3 drugs.

② Research is concerned about the IQ level of student of department in his school. The data of IQ-test result in terms of percentage is recorded. Is there any significant difference between different department of study, Science. Arts, Commerce.

Science	Arts	Commerce.
44	58	65
56	25	45
78	65	58
85	73	38

Sol:

H_0: There is no significant diff blw means.

$\Rightarrow \mu_1 = \mu_2 = \mu_3$

H_1: Atleast one of the above is not equal.

To check if

$$L = (1+\theta)^n \prod_{i=1}^{n} x_i^{\theta}$$

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2
44	58	65	1936	3364	4225
56	25	45	3136	625	2025
78	65	58	6084	4225	3364
85	73	38	7225	5329	1444
263	221	206	18381	13543	11058

$(\Sigma X_1)^2 = 69169$

$(\Sigma X_2)^2 = 48841$

$(\Sigma X_3)^2 = 42436$

$N = 4+4+4 = 12$

$K = 3$

$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3$

$= 263 + 221 + 206 = 690$

Correction factor $= \frac{T^2}{N}$

$= \frac{476100}{12}$

$= 39,675$

$SST = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 - \frac{T^2}{N}$

Ex:

$= 42982 - 39,675$

$= 3307$

$SSC = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} - \frac{T^2}{N}$

$= \frac{69169}{4} + \frac{48841}{4} + \frac{42436}{4} - 39675$

$= 40111.5 - 39675$

$= 436.5$

$SSE = SST - SSC$

$= 3307 - 436.5$

$= 2870.5$

$$MSC = \frac{SSC}{K-1} \quad ; \quad MSE = \frac{SSE}{N-K}$$

$$= \frac{436.5}{2} \qquad = \frac{2870.5}{12-3}$$

$$= 218.25 \qquad = 318.95.$$

$$\frac{MSC}{MSE} = 0.6842$$

Source of variations	Sum of squares	Degree of freedom	Mean square	F
Between samples	436.5	2	218.25	0.6843
Error	2870.5	9	318.95	
Total	3307	11		

Calculated value : $F = \frac{MSC}{MSE} = 0.6842$

Table value = 4.26

$\therefore$ Calculated < Table value

$0.6843 < 4.26$ 5% significance = 0.05.

Accept H_0. Reject H_1.

We conclude, no difference between the different department of study.

2-Way Anova :

When 2 independent factors affect the variable of interest, 2 way Anova is used for testing affect of two factors simultaneously.

Algorithm two-way Anova :-

1) Set up H_0 and H_1.

2) Let row samples be $x_1, x_2, x_3 \ldots$

Column samples be $y_1, y_2, y_3 \ldots$

To check if θ is

$$L = (1+\theta)^n \prod_{i=1}^{n} x_i^{\theta}$$

$$L = (1+\theta)^n \prod_{i=1}^{n} x_i^{\theta-1} \prod_{i=1}^{n} x_i$$

3) Calculate $x_1^2, x_2^2, x_3^2 \ldots$ and $y_1^2, y_2^2, y_3^2 \ldots$

4) Calculate $T_\cdot = \Sigma x_1 + \Sigma x_2 + \Sigma x_3 + \cdots$ (or)

$T_\cdot = \Sigma y_1 + \Sigma y_2 + \Sigma y_3 + \cdots$

5) Find correction factor $C.F = \frac{T_\cdot^2}{N}$; $N = \Sigma n_i$

6) Compute $SST = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \cdots - C.F$ (or)

$SST = \Sigma y_1^2 + \Sigma y_2^2 + \Sigma y_3^2 + \cdots - C.F$

7) $SSR = \frac{(\Sigma x_1)^2}{n_1} + \frac{(\Sigma x_2)^2}{n_2} + \frac{(\Sigma x_3)^2}{n_3} + \cdots - C.F$

$SSC = \frac{(\Sigma y_1)^2}{n_1} + \frac{(\Sigma y_2)^2}{n_2} + \frac{(\Sigma y_3)^2}{n_3} + \cdots - C.F$

8) $SSE = SST - [SSR + SSC]$

9) Prepare a two-way Anova table

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Between row sample	$r-1$	SSR	$MSR = \frac{SSR}{r-1}$	$F_R = \frac{MSR}{MSE}$
Between column sample	$c-1$	SSC	$MSC = \frac{SSC}{c-1}$	$F_C = \frac{MSC}{MSE}$
Error	$(r-1)(c-1)$	SSE	$\frac{SSE}{(r-1)(c-1)}$	

10) Identify the table value accordingly take the decision.

1. For arranging christmas party, Raj, Ram & Ravi were collecting funds from Monday, Tues, Wednes as 1000's of money as tabulated below.
"Is there any significant difference blw their collection and on different days at 5% level".

	Raj	Ram	Ravi
Monday	10	14	8
Tuesday	12	9	15
Wednesday	12	12	16

Sol: ~~Row wise~~: "Column Wise":
H_0: The funds of Raj, Ram, Ravi are same
H_1: The funds of Raj, Ram, Ravi are different.

"Row wise":
H_0: funds collected on 3 days are same.
H_1: funds collected on 3 days are different.

	X_1	X_2	X_3	
Y_1	10	14	8	32
Y_2	12	9	15	36
Y_3	11	12	16	39
	33	35	39	107

$X_1^2 = 100$; $X_2^2 = 144$; $X_3^2 = 121$

Y_1^2 -

	X_1^2	X_2^2	X_3^2	
Y_1^2	100	196	64	360
Y_2^2	144	81	225	450
Y_3^2	121	144	256	521
	365	421	545	1331

$T. = \Sigma X_1 + \Sigma X_2 + \Sigma X_3$

$= 33 + 35 + 39 = 107$

Correction factor $= \frac{T.^2}{N} = \frac{107^2}{9} = 1272.111$

$SST = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 - C.F$

$= 1331 - 1272.111$

$= 58.889$

$SSR = \frac{(\Sigma X_1)^2 + (\Sigma X_2)^2 + (\Sigma X_3)^2}{3} - 1272.111$

$= \frac{[33]^2 + (35)^2 + (39)^2}{3} - 1272.111$

$= 6.22$

$SSE = SST - [SSR + SSC]$

$SSR = \frac{(32)^2 + (36)^2 + (39)^2}{3} - 1272.111$

$= 1280.33 - 1272.111$

$= 8.223$

$SSE = 58.889 - [6.22 + 8.22]$

$SSE = 58.889 - 14.44 = 44.45$

Source of variations	Degree of freedom	Sum of Squares	Mean Square	F.
Between row sample	2	8.22	$\frac{8.22}{2} = 4.11$	.369
Between column sample	2	6.22	$\frac{6.22}{2} = 3.11$	0.279
Error.	(2)(2) = 4	44.45	$\frac{44.45}{4} = 11.112$	

$F_r = 0.369$; $F_c = 0.279$

Table value = 6.94

Table value = 6.94

H_0 is accepted in both the cases.

11/5

Non-parametric Test:-

→ Most of (non-parametric) statistical tests require an assumption that the population of data from which a sample (or) samples are drawn is normally distributed.

→ Most experimental situations yield data which can be tested in usual way, if data comes from a distribution which is bounded on one end. where there is a good chance the distribution will not be normal.

Advantages of Non-parametric test:-

i) Non-parametric tests are distribution free. They donot require any assumptions to be made about popolation following normal (or) any other distribution.

ii) They are simple to understand, easy to apply when sample sizes are small.

iii) Most non-parametric tests donot require lengthy, laborious computations and hence less-tym consuming.

iv) Non-param. tests are applicable to all types of data.

v) Many non-parametric methods make it possible to work with very small samples. which particularly help researches collecting pilot study data (or) to medical researchers working with rear disease.

vi) Non-param. methods make fewer assumptions

* Some non-parametric tests are :-

i) Sign test
ii) Wilcoxom Signed Rank test [Paired T-test]
iii) Man-Whitney test. [Rank-Sum test]
iv) Run test (Wilcoxm Rank Sum test)
v) Kolmogorov-Smirnov test.
vi) Spearsman's test
vii) Kendall-test.

Differences between parametric & Non-parametric:

	Parametric Tests.		Non-Parametric Tests.
1)	It makes assumption above the parameters of pop. distri. from which the data are draw	1)	It makes no such assumptions.
2)	The information about the population is completely known by means of its parameters.	2)	There is no information about the population. But still it is required to test hypothesis of population.
3)	The data should be normally distributed.	3)	Data doesn't follow any specific distribution {Distribution free test}
4)	H_0 is made on parameters of population distribution.	4)	H_0 is free from parameters.
5)	Applicable only for variables	5)	Applicable for both variables & athibutes.
6)	It is the most powerful.	6)	It is not powerful as parametric test.

Dis-advantages of non-parametric-test:

i) Not so efficient.
ii) May waste information.
iii) Difficult to compute manually.
iv) Requires a larger sample size.
v) Statistical-tables are not ready available.

1. Sign Test:

→ The sign-test is particularly useful in situations in which quantitative measurement is impossible (or) inconvinient, but on basis of superior (or) inferior performance it is possible to rank w.r.t each other the two numbers of each pair. The assumption underline this-test is the variable under investigation has continous distribution.

* The Sign-test can be of two types:
i) One-sample sign-test.
ii) The paired sample sign test.

* In one-sample sign test, we test null-hypothesis $H = H_0$ against and appropriate alternative on basis of random sample of size 'n', we replace each sample value $> H_0$ with a '+' sign and less than H_0 is '-' sign and discard the sample value exactly $= H_0$

* In paired sample sign test, each pair of

sample values can be replaced with '+' sign. if first value is greater than a second and '-' sign if the first value smaller than second. (or) we discarded the two value are equal.

*. The sign test is simplest non-parametric test based on the direction of the pair of obs. not on their numerical magnitude

Steps to solve the problems:-

1) In sign test count the no. of '+' signs, '-' sings, $\{H_0\} \in p = 0.05$

2) Let 'S', the no. of times the less frequent sign occurs. Then S has binomial distribution with $p = 1/2$

3) Critical value (Test statistic), for a two-sided alternative is given by: [z]

$$K = \frac{(n-1)}{2} - 0.98\sqrt{n}$$

4) H_0 is rejected if $S \leq K$
H_0 is accepted if $S > K$

5) For, large samples $[n \geq 25]$, the test statistic

$$Z = \frac{(X - np_0)}{\sqrt{n \cdot p_0 (1-p_0)}}$$

X = No. of '+' signs.

Example: One-Sample sign Test

The following are marks of 15 students in a subject use sign test to check if the median mark is 60.

Marks	Signs
66	+
58	−
70	+
60	0
56	−
55	−
81	+
76	+
53	−
71	+
66	+
59	−
88	+
73	+
80	+

$H_0 : M_d = 60$

$H_1 : M_d > 60$

$N = 14$

No. of + signs $= 9$

No. of − signs $= 5$

No. of 0's $= 1$

$S = 5$ [∵ less frequent one]

$$K = \frac{n-1}{2} - 0.98\sqrt{n}$$

$$K = \frac{13}{2} - 0.98\sqrt{14}$$

$$K = 6.5 - 0.98(3.74)$$

$$K = 2.83$$

∴ $S > K$

$5 > 2.83$

∴ H_0 is accepted.

⇒ $\boxed{M_d = 60}$

The following data gives the marks of 15 students in a particular subject before and after the revision class. Check if there is any signi. diff before and after revision class.

Before	After	Sign
	41	−
38	37	−
34	44	+
45	27	+
28	33	−
27	30	−
25	38	+
41	36	0
36	32	−
30	29	−
28	33	+
34	32	+
35	37	+
40	43	−
42	40	−
33		

No. of '+' signs = 6

No. of '−' signs = 8

No. of '0's = 1

$S = \min(6, 8) = 6 \; ; \; N = 14$

$K = \frac{13}{2} - 0.98\sqrt{14} = 2.83$

H_0: There is no difference in marks

H_1: There is increase / decrease in marks

$6 > 2.83$

$S > K$

H_0 is accepted.

A Typing school claims that in a 6-week course, it can train students to type on the average atleast 60 words permanent. A random

Sample of 16 graduates is given a typing test and medium no. of words per minute typed by each of these students given below. Test the hypothesis that the median typing speed of graduates is atleast 60 words per minute.

Students	Words per Minute	Signs
A	81	+
B	76	+
C	63	+
D	71	+
E	66	+
F	59	−
G	88	+
H	73	+
I	80	+
J	66	+
K	58	−
L	70	+
M	60	0
N	56	−
O	55	−

No. of '+' values = 10
No. of '−' values = 4
No. of 0's = 1 ; N = 14
$S(\min(4,10)) = 4$

$K = \frac{N-1}{2} - 0.98\sqrt{N}$

$= 13/2 - 0.98\sqrt{14} = 2.83$

$4 > 2.83 \Rightarrow S > K.$

Accept H_0.

Use the sign test to see if there is difference blw the no. of days until correction of an account recievable before and after a new collection policy. Use 5% level of significance.

Before	After	Sign
30	32	−
28	29	−
34	33	+
35	32	+
40	37	+
42	43	−
33	40	−
38	41	−
34	37	−
45	44	+
28	27	+
27	33	−
25	30	−
41	38	+
36	36	0

H_0: There is no significance difference

H_1: There is significance differen

No. of +ve signs = 6

−ve signs = 8

0's = 1 $\Rightarrow$ $8 (\min 6, 8) = 6$

$N = 14$

$k = \frac{N-1}{2} - 0.98\sqrt{14}$

$= 13/2 - 0.98(\sqrt{14}) = 2.83$

$6 > 2.83$

H_0 is accepted.

$8 > k$

The following data relate to daily production of cement in metric tonns for large plans of 30 days. Use sign test to test null hypothesis that the plans average daily production is 11.2 metric tonns against the alternative hypothesis that its less than 11.2 m.tonns. Here 5% level of sign.

Sol:

11.5	+
10	−
11.2	0
10	−
12.3	+
11.1	−
10.2	−
9.6	−
8.7	−
9.3	−
9.3	−
10.7	−
11.3	+
10.4	−
~~10.4~~	−
11.4	+
12.3	+
11.4	+
10.2	−
11.6	+
9.5	−
10.8	−
11.9	+
12.4	+
9.6	−
10.5	−
11.6	+
8.3	−
9.3	−
10.4	−
11.5	+

No. of + signs = 11

− signs = 18

0's signs = 1

$P_0 = 1/2$ $N = 29$, $X = 11$

$$Z = \frac{(X - np_0)}{\sqrt{n p_0 (1-p_0)}}$$

$$= \frac{(11 - 29(1/2))}{\sqrt{29(1/2)(1-1/2)}}$$

$$= \frac{-3.5}{\sqrt{1.346}} = \frac{-3.5}{2.69}$$

$$= -1.3011$$

$Z = -1.3011$

Table value $= -1.64$

$\therefore -1.64 < -1.3011$

Hence, we reject H_0 and accept H_1.

H_1: Avg production is less than 11.2 metric tonns.

Wilcoxm Signed-Rank Test:-

* This test is more powerful than sign test because it tests not only directions but also magnitudes of differences within pairs of matched groups. Like the sign test, deals with dependent groups made up of matched pairs of individuals and not applicable to independent groups.
* We test the null hypothesis $H_0: M_1 = M_2$ the null hypothesis H_0 is rejected if the computed value of lowest signed value is less than or equal to appropriate table value.

Note: $n \geq 15$ is considered as large sample.

the test statistic $Z = \frac{W_+ - M}{\sigma}$

W_+ = Sum of +ve ranks

$M = \frac{n(n+1)}{4}$

$\sigma^2 = \frac{n(n+1)(2n+1)}{24}$

1) The Body weight of 10 patients before and after the weight loss treatment given below:

Use Wilcoxm Signed rank test; to check if there is any significant different before and after treatment.

Before	After	Dif	Absdif	Rank
58.5	60	-1.5	1.5	2 (-ve)
60.3	54.9	5.4	5.4	7
61.7	58.1	3.6	3.6	6
69.0	62.1	6.9	6.9	10
64.0	58.5	5.5	5.5	8
62.6	59.9	2.7	2.7	4
56.7	54.4	2.3	2.3	3
63.6	60.2	3.4	3.4	5
68.2	62.3	5.9	5.9	9
59.4	58.7	0.7	0.7	1

Ho: There is no significant difference in weight before and after weight lose treatment

Hi: There is significance difference. [2-tailed]

W_- = Sum of rank from -ve diff. = 2

W_+ = Sum of rank from +ve diff = 53

$$\mu = \frac{n(n+1)}{4} = \frac{110}{4} = 27.5$$

$$\sigma^2 = \frac{n(n+1)(2n+1)}{24} = \frac{10 \times 11 \times 21}{24} = 96.5$$

$$\sigma = \sqrt{96.25} = 9.81$$

~~$Z = W_+ - \mu_0$~~

~~$\neq 53 +$~~

Calculated value = Min (53, 2) = 2

$n = 10$, $\alpha = 0.05$

Table value = 8

Calculated value < Table value

∴ We reject H_0.

There is difference before and after weight lose prgrm.

A survey of American themographis estimated the avg household spending on healthcare. On Average was 1800 dollars. 6 families in 2 locations are matched demographically and data is given below.

Family pair	Location 1 ($)	Location 2 ($)
1	1950	1760
2	1840	1870
3	2015	1810
4	1580	1660
5	1790	1840
6	1925	1765

Use $\alpha = 0.05$ to test whether there is significant difference in annual hosehold healthcare spending blw 2 cities.

Sol: H_0: There is no significant difference.

H_1: There is significant difference.

X	Y	Dif	Abs dif	Rank
1950	1760			
1840	1870			
2015	1810			
1580	1660			

A company implemented a quality control program the president wants to find if worker productivity increased after implemented of quality control prgrm. Use $\alpha = 0.05$.

Ho: There is no significant difference.

H1: There is increase in the productivity.

Worker	Before	After	Diff	Rank
1	5	11	-6	19
2	4	9	-5	17
3	9	9	0	-
4	6	8	-2	9
5	3	5	-2	9
6	8	7	1	3.5
7	7	9	-2	9
8	10	9	1	3.5
9	3	7	-4	14.5
10	7	9	-2	9
11	2	6	-4	14.5
12	5	10	-5	17
13	4	9	-5	17
14	5	7	-2	9
15	8	9	-1	3.5

	7	6	1	3.5
16	9	10	-1	3.5
17	5	8	-3	12.5
18	4	5	-1	3.5
19	3	6	-3	12.5
20				

$W+ = (3.5+3.5+3.5) = 10.5$

$W- = 179.5$

$\mu = \frac{n(n+1)}{4} = \frac{19 \times 20}{4}^{5} = 95$

$\sigma^2 = \frac{n(n+1)(2n+1)}{24} = \sqrt{\frac{19(20)(39)}{24}} = 24.8$

$Z = \frac{10.5 - 95}{24.8}$

$Z = -3.407 = -3.41$

Table value = 2.33

$-3.41 < -2.33 \Rightarrow -2.33 > -3.41$.

We accept H_1.

Reject H_0.

Mann-Whitney U-test: [Rank-Sum test]:

It is a non-parametric test. It is used to compare the difference b/w two independent groups when dependent variable is either discrete (or) continous but not normally distributed.

It can be used for very small samples

Let the sample size n_1 & n_2 respectively. U

$U = \text{Min}(U_1, U_2]$ for $n_1, n_2 > 0$ {large sample}

Where $U_1 = n_1n_2 + \frac{n_1(n_1+1)}{2} - R_1$

$U_2 = n_1n_2 + \frac{n_2(n_2+1)}{2} - R_2$

where R_1 = Sum of ranks of first sample.

R_2 = " " " second "

The test statistic for Mann-Whitney's U-test.

$$Z = \frac{U - \frac{n_1n_2}{2}}{\sqrt{\frac{n_1n_2(n_1+n_2)}{12}}}$$

1) 23 applications for a position of interview by 3 administrators rated on a scale of 5 as to suitability for position. Each one is given a - Use Mann-Whitney U-test to know whether there is difference of two groups A & B, One with educational background less than 2-years and another with atleast 2-years : $\alpha = 0.05$.

Group A	Group B
7	8
17 ($n_1 = 12$)	9
9	13
4	14
8	11
6	10 ($n_2 = 12$)
12	12
11	14
9	13
10	9
11	10
11	8

Sol:

A	B	R_A	R_B
4		1	
6		2	
7		3	
8	8,8	5	5,5
9,9	9,9	8.5, 8.5	8.5, 8.5
10	10,10	12	12,12
11,11,11,11	11	16,16,16,16	16
12	12	19.5	19.5
	13,13	123.5	21.5, 21.5
	14,14		23.5, 23.5
			176.5

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$= (12)(12) + \frac{12(13)}{2} - 123.5$$

$$= 144 + 6(13) - 123.5$$

$$= 98.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$= 12(12) + \frac{(12)(13)}{2} - 176.5$$

$$= 45.5.$$

$$U = \text{Min}\{98.5, 45.5\} = 45.5$$

$$Z = \frac{U - n_1 n_2/2}{\sqrt{\frac{n_1 n_2 (n_1 + n_2)}{12}}}$$

$$= \frac{45.5 - 72}{\sqrt{\frac{12(12)(24)}{12}}} = \frac{45.5 - 72}{\sqrt{12(24)}}$$

$$= -1.56.$$

Table value = 1.96

$-1.56 < 1.96$.

H_0 is accepted.

There is no significant difference of b/w A (less than 2 years of edu) and B (atleast 2 years of education).

eg- Display a String in Rconsole.

```
>"Hey. Welcome to Sastra"
[1] "Hey Welcome to Sastra"
```
(display symbol)

```
>
```

* Variable declaration

(>age<-)

instead we can use "=" also

```
>age<-20
>name<-"Leela"
>age
[1] 20
>name
[1] "Leela"
```

* >ls() (lists all the stored variables that are created during a particular r session.)

```
[1] "age", "name".
>Salary
Error: object "Salary" not found.
```

* >height <-2 → #create 2 new var height & width

```
>width<-4
>height
[1] 2
>width
[1] 4 — #Find area
>area<-height*width
>area
[1] 8
>ls()
```

> "area", "height", "width", "area".

* The Comments in R starts with #

Remove function – rm() :

* > rm(area) – area will be removed.

(continue above program)

* > rm (area)

> ls()

> "height", "width".

Basic data types –

* R fundamental data types are also known as atomic Veetor types. and are used extensively in R programs. R uses a function named class() to determine the type of a Variable.

* The Variables can be logical, numeric, integer, character, Complex, raw, double.

* > class (TRUE) } [1] "logical"
> class (FALSE) } output

> class (4) – [1] "numeric"

> class (4L) – [1] "Integer".

> class ("Welcome to Sastra") – [1] "character"

> z = 10 + 2i

> class (z) – [1] "complex".

* class(T) (or) (F) - indicates logical (True) not character

* class (TTT) - character

* NA - represents missing value

```
>TRUE
[1] TRUE
>class (TRUE)
 [1] "logical"
>class (FALSE)
 [1]" logical"
>T
[1]TRUE
>F
[1]FALSE
>class (NA)
 [1]"logical"
>4
[1]4
>4.5
[1]4.5
>4L
>4
>class (4)
 [1]" numeric"
>class (4L)
 [1] "integer"
>a <= double (3)
>a
 [1] 0 0 0
>is. double (a)
```

```
[1] TRUE
```

Create d.p.v of ↑length 5

```
>b <- double (5)
>b
[1] 0 0 0 0 0
> is. integer (b)
 [1] FALSE
```

it creates a double precision vector of length 3

15. Function – (is dot func)

* It is used to check whether the given input is numeric / Integer / double -- etc.

* > is.numeric (2)
  [1] TRUE
  > is.integer (2)
  [1] FALSE
  > is.integer (2L)
  [1] TRUE
  > is.numeric (2L) (all integer var are numeric but not all numeric var are integers).
  [1] TRUE
  > z <- 10+2i
  > class (z)
  [1] "complex".

Conversion of any given datatype to Raw data type.

* R provides flexibility to programmers to convert any given data type to special data type called raw data type.
* Raw data type is intented to hold any data as a sequence of bytes, where it is possible to extract subsequences of bites and replaced in as elements of vector.
* In R raw vectors are used to store fixed length sequences of bytes.

```
>name <- "Leela"
>class(name)
[1] "character"
>r <- charToRaw(name)        → Converting character to Raw.
>class(r)
[1] "raw"
>convert Raw to char(r)
>name
[1] "Leela"
```

A vector is a sequence of data elements of same data type. There are 6 types of atomic vectors:-

i) logical
ii) integer
iii) Double
iv) complex
v) charaiter
vi) Raw.

Programmers can create character vectors & numeric vectors, logical vectors etc.

Creating and naming vectors:

C() paranthesis is used to create vector in R.

Eg:- >earnings <- c(50,100,30)
>earnings * 3
[1] 150 300 90
>earnings /5

[1] 10 20 6

>earnings + 20

[1] 2500 10000 900

Coercion - [adding two vectors | attaching smthg]

>remain <- c (11,12,11,13)

> remain

[1] 11 12 11 13

> Suits < c- ("spades", "hearts", "diamonds," clubs

> names (remain) c- Suits

> remain

Spades hearts diamonds clubs
11 12 11 13

Attach labels to elements:

>remain<- c ["spades"-11, "hearts-12, "diamonds
-11, "clubs"-13]

>length (remain)

[1] 4

'str' function displays complete structure of R objects with details of all the variables and their data types.

> str (remain)

Named num [1:4] 11 12 11 13

attr (*, "Remain") = chr [1:4] "spades" "hea
"diamonds" "clubs"

Coercion of vector element

```
> college <- c (13, 54, "A", "B", 13, 32, "K", "R")
> college
[1] "13" "54" "A" "B" "13" "32" "K" "R"
> class (college)
[1] characters.

> earnings <- c(50, 100, 30)
> expenses <- c (30, 40, 80)
> earnings - expenses
[1] 20 60 -50
> earnings + c (10, 20, 30)
[1] 60 120 50
> earnings * c (1, 2, 3)
[1] 50 200 90
> earnings / c (1, 2, 3)
[1] 50, 50, 10.

> Bank <- earnings - expenses
> bank
[1] 20 60 -50
> Sum (bank)
[1] 30.
```

Vector Subsetting:-

```
> remain <- c (Spades = 11, hearts - 12, diamonds - 11
                clubs - 13)
> remain [1]
```

Spades
11
>remain [3]
diamonds
11
>remain ["hearts"]
hearts
12
>remain_black <- remain [c(1,4)]
Spades clubs
> remain_black <- remain [c(4,1)]
> remain_black
clubs spades
13 11
>remain [-1] (removes 1st one) clubs)
hearts diamonds clubs
12 11 13
* >remain [-(1,2)]
>remain [-"Spades"]
Error
>remain [c(FALSE, TRUE, FALSE, FALSE)]
(hearts) 12
> selection vector ← c(F,T,F,F)
>remain [selection vector]
(hearts)
12

4. Spades hearts diamonds clubs > remain [c(FTFT)]
hearts clubs
12 13

>remain [c(FTF)]

Matrices are objects in which the elements of same atomic type are arranged in 2D rectangular layout.

Basic syntax for creating matrix:

Matrix (data, nrow, ncol, byro, dimension).

Data:- Data is input vector which becomes the data elements of matrix.

nrow:- No. of rows to be created

ncol:- No. of columns to be created.

Byro:- It's a logical clue. If true, then the input vector elements are arranged by row.

Dimension names:- Are the names assigned to the rows and columns.

Note:- The users cannot have logicals and numericals in matrix. It can only be atomic vector type (or) only characters (or) only logicals (or) only numericals.

* If you don't mention byro, by default it will print byro is false.

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix} \quad \begin{pmatrix} 1 & 6 & 11 \\ 2 & 7 & 12 \\ 3 & 8 & 13 \\ 4 & 9 & 14 \\ 5 & 10 & 15 \end{pmatrix}$$

Byro is false.

Matrix (1:15, nrow=5, ncol=3, byrow=T,
dimnames = list (c($R_1 R_2 R_3 R_4 R_5$)
c(C_1, C_2, C_3)))

```
>n<-matrix(1:6, nrow:2)
 [1,] 1 3 5
 [2,] 2 4 6

>n<-matrix (1:6, nrow=2, byrow=T)
> n[1][,2][,3]

 [1,] 1 2 3
 [2,] 4 5 6

>matrix (1:6, ncol=2)
      [,1] [,2]

 [1,] 1 4
 [2,] 2 5
 [3,] 3 6

>matrix (1:5, nrow=2, ncol=3)
      [1,] [,2] [,3]

 [1,] 1 3 1
 [2,] 2 4 2
```

Columnbind (cbind), Row bind (Rbind):-

```
> cbind (1:3, 4:6) 7:9)
   1 4 7
   2 5 8
   3 6 9
```

```
> rbind (1:3, 4:6)-7:9)
     1 2 3
     4 5 6
```

```
>m<- matrix(1:6, nrow=2, bryrow=T)
>m    [,1] [,2] [,3]
  [1,]  1    2    3
  [2,]  4    5    6
```

```
rbind (m; 7:9)
     [,1] [,2] [,3]
[1,]  1    2    3
[2,]  4    5    6
[3,]  7    8    9.
```

```
>rownames (m)<- c("Row1", "Row2")
>colnames (m)<- c("col1", "col2", "col3")
>m  col1 col2 col3
Row1  1    2    3
Row2  4    5    6.
```

Matrix subsetting:-

Selecting elements randomly in to a matrix:

```
>m<- matrix (sample (1:15, 12) , nrow=3)
>m
     [1] [2] [3] [4]
[1,]  2   12  10  11
[2,]  6   7   3   1
[3,]  9   8   13  5
```

```
>m<-[,3] 10 3 13
>m<-[2,] 6 7 3 1
>m[c(1,3),c(1,4)]
     2 11
     9 5
>m[c(2,3) c(2,3,4)]
     9 3 1
     8 13 5
```

```
>m[5]
[1] 7
> m[10]
[1] 11
```

* r_1 r_2 r_3

a b c d

```
>remain(m)<-c("r1","(r2)","r3")
>olumn(m)<-c("a","b","c","d")
```

```
> m
     a  b  c  d
r1   2  12 14 11
r2   6  7  3  1
r3   9  8  13 5
```

```
>m[c("r1".("r3")),c("a","b")]
     2 12
     9 8
```

```
>m[c(T,F,T),c(T,T,F,F)]
>T(m
     2  6  9
     12 7  8
     14 3  13
     11 1  5
```

Array:- It is a collection of similar data of same type.

Array function () is used to create array by passing vectors as input and by using values of dimension parameters.

Basic syntax for creating an array ():

array (data, dim, dimname)

where data = input values of vector.

dim - used to create dimension.

dimname = used to assign a name to the dimension.

```
1> Vec1 <- c(5,9,3)
> vec2 <- c(10,11,12,13,14,15)
> result <- array (c(vec1,vec2), dim(3,3,2))
> result
  5 10 13
  9 11 14
  3 12 15

  5 10 13
  9 11 14
  3 12 15
```

dim(3,3,2): 3 → no. of rows, 3 → no. of cols, 2 → no. of times repeat.

Introduction to data frame:-

R is a statistical programming where large data sets are used. Data sets comprise observations (or) instances that have some variable associated with them. For example:

name, age, no. of children, salary, saving....
A matrix cannot be used for such information as name would be charaiter and age numeric.
Dataframe is fundamental data structure to store datasets. It is similar to a spreadsheet that has rosen columns where each column can be a different vector.

Read
write
work &
Manipulate.

Creating a dataframe:

Create a dataframe with Id number 1 to 15. name of person and his/her age.
We have 3 independent equal length vectors, 2 numerical and one charaiter.

Syntax to create data frame():

data.frame()
data.frame (Id, name, age)

```
>Id<-c(1:15)
>name<-c("A","B","C","D","E","F","G","H",
"I","J","K","L","M","N","O")

>age<-c(15,14,21,27,10,12,25,22,19,18,
19,20,21,22,23).
```

```
>data.frame (Id, Name, age)
>details <- data.frame (Id, name, age)
>details
  Id  name  age
1      AA    15
2      BB    14
3      CC    21
4      DD
5      EE
>nrow (details)
[1] 15
>ncol (details)
[1] 3
>head (details) (first 6)

>tail (details) (~~first~~ last 6)
>details [14, 2]
   "NN"
>details[10, "Name"]
>details [14, ]
>class (details)
  "data.frame"
>class (details [, "age"])
   "Numeric"
>details [c(1:12), c("Id", "Name")]
```

Extending data frames:

* To add a new column, a new variable, new row (or) new observation to existing data: use $ (or) [[ ]]. cbind function, rbind function can also be used.

A vector named height is created and is added using $ (or) [[ ]] as follows:

```
>height <- c (.....)
>details $ height <- height
        (or)
>details [[height]] <-height
> details
    Id Name age height
     :
     :
> cbind (details, height)
> karan < data frame (16, "Karan", 20,168)
> rbind (details, karan)
```

Sorting in data frames:-

To sort (), sort function is used -> Sort (details $age)

↓ ascending.

```
>sort (details $ height).
>details [order (details $age, decreasing TRUE)
                    ↳ descending
```

Conditionals & Control flow:

1. Equality check.
2. Comparing strings, numbers and logical values.
3. Relational operators & vectors.
4. logical operators.

1. Equality check:

```
> TRUE == T
[1] TRUE
> T == FALSE
[1] FALSE
```

2.

```
> "HELLO" == "GOODBYE"
[1] FALSE
> 3 == 2
[1] FALSE
> 3 == 2+1
[1] TRUE
> TRUE != FALSE
[1] TRUE
> "HELLO" != "GOODBYE"
[1] TRUE
> "Hello" > "GOODBYE"
[1] TRUE
```

3. Relational operators & vectors:

```
> linkedin <- c(16,9,13,5,2,17,14)
> linkedin
[1] 16  9 13  5  2 17 14
```

> facebook <- c(14, 7, 15, 16, 8, 13, 14)

> linkedin > 10

[1] T F T F F T T

> facebook < > 10

[1] T F T T F T T

> facebook <= linkedin

[1] F T F F F T T

If the elements are added in the one vector then it compares with the repeatition of elements in other vector and gives result with a warning message.

4. Logical operators:-

AND	OR	NOT
&	|	!
T & T = T	T | T = T	T ! = F
T & F = F	T | F = T	F ! = T
F & T = F	F | T = T	
F & F = F	F | F = F	

> c(T, T, F) & c(T, F, F)

[1] T F F

> c(T, T, F) | c(T, F, F)

[1] T T F

> ! c(T T F)

[1] F F T

> c(T, T, F) && c(T, F, F)

[1] T

> c(T, T, F) || c(T, F, F)

[1] T

Conditional Statements:-

1. if else, if

Syntax: (IF)

```
If (condition) {
        expression
    }
```

If condition evaluates to True, code in curly brackets is executed.

(If else)

```
If(condition)
    {
      expression 1
    }
else (condition)
    {
      expression 2
    }
```

Nested if: Syntax:

```
if (condition) {
      expression1
    }
else if (condition 2) {
        expression 2
      }
  else {
      expression 3
    }
```

Eg:- >x<-5
>if (x<0){
+print ("x is negetive as"):
+} else
+print ("x is either +ve (or) 0")
+}
[1] "x is eithere +ve (or) 0.

2. >x<-5
> if (x>0){
+print("x is negetive")
+} else if (x==0)
+print ("x is 0")
+} else {
+print ("x is true")
[1] "x is tve".

Iterative programming in R:-

1. while loop
2. for loop.

→ while loop Syntax:-

```
while (test-expression) {
        statement
    }
```

for loop:-

```
for (variable in start : end) {
```

statement

```
}
> v<-c("Best wishes", "Take care")
> cnt <- 2
> while (cnt<7){
+ print(v)
+ cnt = cnt+1
+ }
```

[1] "Best wishes" "Take care"
[2] "Best wishes" "Take care"
[3] "Best wishes" "Take care"
[4] "Best wishes" "Take care"
[5] "Best wishes" "Take care"

```
> v<-c("Hello friends", "keep smiling")
> for (i in 2:7){print(v)}
```

6 times printing.

looping over list:-

1. Loops for vectors
2. loops for matrices
3. loops for lists
4. loops for data frames.

```
> v<-c(1,2,"one","True")
> for (i in v){print(i)}
```

for loop.

[1] 1
[2] 2
[3] "one"
[4] "True"

```
> v<-c(1, 2, "one", "True")
> cnt <-1
while (cnt < length (v))
+{
+ print (v)
+ cnt = cnt+1
+}
[1] 1 2 "one" "True"
[2] 1 2 "one" "True"
[3] 1 2 "one" "True".
```

The difference blw for and while loops is that the control variable in while loop needs to be initialised and controlled.

1) Explain the use of length function(), sum function(), l.s function, class function, str functi.

* Give examples to delase variables in R.
* Give examples for performing arithmetic operations in R.
* Give exampler to use l.s function?
* How do you convert any given data type to raw-data type. give examples
* Write syntax to create vector, name, coefficient and subsetting of vectors.
* Write syntax to create a matrix,
  a) Create a matrix using numbers 51-75

with 5 rows? Byrow = True, Give dimension names?

b) Create another matrix with alphabets A-Z with 2 rows.

c) Use Rcode to find transpose of first matrix

d) Use Rcode to pickup vowels from second matrix

e) Use cbind and rbind to add columns and rows to second matrix.

Answers

1. length function - determine the length of the function.

ls() - lists all the stored variables that are created during particular session.

class function - To determine type of variable.

str() - displays the complete structure of R object with details of all.

sum() - Used to combine 2 vectors.

Examples:

1) >Name <- leela
>age <- 18
>Name
[1] leela

2) >length <- 4
>Breadth <- 5

Loops of matrices:-

Two loops are needed, one for iterating over the rows, and another iterating over the columns.

```
i)>y<-matrix (1:20, nrow=5, ncol=4)
  > for (i in 1:nrow(y))
+}
+ >for (j in 1:ncol(y)
+}
+ print{y (i,j)}
 +}
 +}
```

Ex:-

1	6	11	16
2	7	12	17
3	8	13	18
4	9	14	19
5	10	15	20

```
[1] 1
[1] 6
[1] 11
    16
    2
    7
    :
    20
```

Loops for dataframes :-

Write an r programme to create 4 vectors namely patient Id, age, type of diabetes, status of patient. put these 4 vectors in to a dataframe named patient data and print the values using for loop.

Sol:- dataframe()

data.frame (Id, age, typeof diabetes, status of patient)

```
> Id <- c(1:5)
> age <- c(15,22,37,42,56)
> Types of diabetes <- c(type 1, type 2, type1,2,1)
> Status of patient <- c("Good improvement",
  "Steady", "Poor", "Improved", "Bad")
> dataframe (Id, age, types of diabetes,
                        Status of patient):
>patient data > - data.frame ( " " ):
>patient data for (i in names (patient data)){
+print (patient data [i]
+}
```

patient Id Type of diabetes

age Status of patient

Recursion:- function calls itself.

Ex:-

fibonacci series.

> area <- length * Breadth
> area
[1] 20

3) > is.Numeric(2L)
[1] True
> is.integer(3L)
[1] True
> is.Integer(5)
[1] FALSE

2) Replaced in as elements of vectors:
> name <- "leela"
> class(name)
[1] "character"
> r <- charToRaw(name)
> class(r)
[1] "raw"
> Raw to chan(r).
> name
[1] "leela"