UNIT-4

BIVARIATE DISTRIBUTION

Curve Fitting :-

It is a method of finding equation of a curve that approximates a given set of data

Best fitting by the method of least squares:

Suppose $y=f(x)$ is eqn of aprox. curve for given data. The method of least squares means the fitting line for which the sum of the squares of the verticle distances of the point (x_i, y_i) from the line is minimum.

Linear curve fitting :

Suppose the line equation is $y=a+bx$.

Normal equations are $\Sigma y_i = na + b\Sigma x_i \longrightarrow (1)$

$$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 \longrightarrow (2)$$

Note:-

Suppose the linear curve is of the form $y=ax+b$

Now eqns become $\Sigma y_i = nb + a\Sigma x_i$

$$\Sigma x_i y_i = b\Sigma x_i + a\Sigma x_i^2$$

Problems:

1. By method of least squares, fit a straight line to the following data

x :	0	1	2	3	4
y :	1	1.8	3.3	4.5	6.3

Sol:) let the st.line is of the form $y = a+bx \rightarrow (1)$

then normal eqns are : $\Sigma y_i = na + b\Sigma x_i \longrightarrow (2)$

$$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 \longrightarrow (3)$$

0	1
1	1.8
2	3.3
3	4.5
4	6.3
$\Sigma x_i = 10$	16.9

from (2)

from (3)

solvin

2. Fit a

let th

then

x_i	y_i	$x_i y_i$	x_i^2
0	1	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16
$\Sigma x_i = 10$	16.9	47.1	30

from (2), $16.9 = (5)a + b(10)$

$\Rightarrow 5a + 10b = 16.9 \longrightarrow (4)$

from (3), $47.1 = a(10) + b(30)$

$\Rightarrow 10a + 30b = 47.1 \longrightarrow (5)$

solving (4) & (5)

$$\begin{array}{r} 10a + 20b = 33.8 \\ 10a + 30b = 47.1 \\ \hline -10b = -13.3 \end{array}$$

$\Rightarrow \boxed{b = 1.33}$

from (4), $5a + 10(1.33) = 16.9$

$\Rightarrow 5a = 3.6$

$\Rightarrow \boxed{a = 0.72}$

from (1), $\boxed{y = 0.72 + 1.33x}$

2) Fit a st. line for the following data

x	1	2	3	4	5
y	19	27	40	55	68

let the st. line is of the form $y = a + bx \longrightarrow (1)$

then normal eqns are : $\Sigma y_i = na + b\Sigma x_i \longrightarrow (2)$

$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 \longrightarrow (3)$

Here $n = 5$

x_i	y_i	$x_i y_i$	x_i^2
1	19	19	1
2	27	54	4
3	40	120	9
4	55	220	16
5	68	340	25
15	209	753	55

from (2), $209 = 5(a) + b(15)$

$\Rightarrow 5a + 15b = 209 \longrightarrow (4)$

from (3), $753 = a(15) + b(65)$

$\Rightarrow 15a + 55b = 753 \longrightarrow (5)$

Solving (4) & (5)

$15a + 55b = 753$

$15a + 45b = 627$

$10b = 126$

$\boxed{b = 12.6}$

from (4), $5a + 15(12.6) = 209$

$\Rightarrow \boxed{a = 4}$

from (1), $\boxed{y = 4 + 12.6x}$

Non-linear Curve Fitting

By least squares method, fitting of a quadratic curve are parabola which is of the form

$$y = a + bx + cx^2$$

Now normal eqns are $\Sigma y_i = na + b\Sigma x_i + c\Sigma x_i^2$

$$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 + c\Sigma x_i^3$$

$$\Sigma x_i^2 y_i = a\Sigma x_i^2 + b\Sigma x_i^3 + c\Sigma x_i^4$$

Ex: Fit a polynomial of 2nd degree for the following data:

x	0	1.0	2.0
y	1.0	6.0	17.0

let the polynomial eqn be $y = ax + bx + cx^2 \rightarrow (1)$

Normal eqns are :- $\Sigma y_i = na + b\Sigma x_i + c\Sigma x_i^2 \rightarrow (2)$

$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 + c\Sigma x_i^3 \rightarrow (3)$

$\Sigma x_i^2 y_i = a\Sigma x_i^2 + b\Sigma x_i^3 + c\Sigma x_i^4 \rightarrow (4)$

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
0	1	0	0	0	0	0
1	6	1	1	1	6	6
2	17	4	8	16	34	68
3	24	5	9	17	40	74

from (2), $24 = 3(a) + b(3) + c(5)$

$\Rightarrow 3a + 3b + 5c = 24$

from (3), $40 = a(3) + b(5) + c(9)$

$\Rightarrow 3a + 5b + 9c = 40$

from (4), $74 = a(5) + b(9) + c(17)$

$\Rightarrow 5a + 9b + 17c = 74$

Solving above eqns, we get

$a = 1, b = 2, c = 3$

from (1), $y = x + 2x + 3x^2$

$$\boxed{y = 1 + 2x + 3x^2}$$

2. Fit the parabola for the following data

x	2	4	6	8	10
y	3.07	12.85	31.47	57.38	91.29

Sol) let the parabola eqn be $y = a + bx + cx^2 \longrightarrow (1)$

Normal eqns :- $\Sigma y_i = na + b\Sigma x_i + c\Sigma x_i^2 \longrightarrow (2)$

$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 + c\Sigma x_i^3 \longrightarrow (3)$

$\Sigma x_i^2 y_i = a\Sigma x_i^2 + b\Sigma x_i^3 + c\Sigma x_i^4 \longrightarrow (4)$

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
2	3.07	4	8	16	6.14	12.28
4	12.85	16	64	256	51.4	205.6
6	31.47	36	216	1296	188.82	1132.92
8	57.38	64	512	4096	459.04	3672.32
10	91.29	100	1000	10000	912.9	9129
30	196.06	220	1800	15664	1618.33	14152.12

from (2), $196.06 = 5a + b(30) + c(220)$

$1618.33 = a(30) + b(220) + c(1800)$

$14152.12 = a(220) + b(1800) + c(15664)$

$a = 0.6466$, $b = -0.8350$, $c = 0.9903$

from (1), $y = 0.696 + (-0.85507)x + 0.99196x^2$

Fitting of an exponential Curve

These are 2 types.

1) $y = ab^x$

2) $y = ae^{bx}$

Normal eqns are $\Sigma Y_i = nA + B\Sigma x_i \longrightarrow (1)$

$\Sigma x_i Y_i = A\Sigma x_i + B\Sigma x_i^2 \longrightarrow (2)$

Where $A = \log_e a$, $B = \log_e b$, $Y = \log_e y$

$\Rightarrow a = e^A$, $b = e^B$, $y = e^Y$

Fit a curve is of the form $y = ab^x$ for the following data

x	2	3	4	5	6
y	8.3	15.4	33.1	65.2	127.4

let the curve eqn : $y = ab^x \rightarrow (1)$

Normal eqns are $\Sigma Y_i = nA + B\Sigma x_i \rightarrow (2)$

$\Sigma x_i Y_i = A\Sigma x_i + B\Sigma x_i^2 \rightarrow (3)$

x_i	y_i	$Y_i = \log_e y_i$	$x_i Y_i$	x_i^2
2	8.3	2.116	4.232	4
3	15.4	2.734	8.202	9
4	33.1	3.499	13.996	16
5	65.2	4.177	20.885	25
6	127.4	4.847	29.082	36
20	249.4	17.3733	76.397	90

from (2), $17.3733 = 5A + B(20)$

from (3), $76.397 = A(20) + B(90)$

$A = 0.71314$, $B = 0.69038$

$\Rightarrow a = e^A$, $b = e^B$

$\Rightarrow \boxed{a = 2.0403}$ $\boxed{b = 1.99447}$

from (1), $y = (2.0403)(1.99447)^x$

Fitting :

curve : $y = ae^{bx}$

Normal eqns are $\Sigma Y_i = nA + b\Sigma x_i \rightarrow (1)$

$\Sigma x_i Y_i = A\Sigma x_i + b\Sigma x_i^2 \rightarrow (2)$

1) Fit a curve is of the form $y = ae^{bx}$ for following data:

x	1	5	7	9	12
y	10	15	12	15	21

eqn of curve: $y = ae^{bx}$

Normal eqns are: $\Sigma Y_i = nA + b\Sigma x_i \longrightarrow (1)$

$\Sigma x_i Y_i = A\Sigma x_i + b\Sigma x_i^2 \longrightarrow (2)$

x_i	y_i	$Y_i = \log_e y_i$	$x_i Y_i$	x_i^2
1	10	2.3025	2.3025	1
5	15	2.70805	13.54025	25
7	12	2.4849	17.3943	49
9	15	2.70805	24.3724	81
12	21	3.04452	36.5342	144
34		13.24802	94.1437	300

from (1), $13.24802 = 5A + b(34)$

$94.1437 = A(34) + b(300)$

$\boxed{A = 2.2486}$, $\boxed{b = 0.05897}$

$\Rightarrow a = e^A$

$\Rightarrow \boxed{a = 9.475}$

eqn :- $y = (9.475)e^{(0.05897)x}$

2) Fit a curve $y = ae^{bx}$ for following

x	0	1	2	3	4	5	6	7	8
y	20	30	52	77	135	211	326	5+50	1052

Sol) eqn of curve: $y = ae^{bx}$

Normal eqns are: $\Sigma Y_i = nA + b\Sigma x_i$

$\Sigma x_i Y_i = A\Sigma x_i + b\Sigma x_i^2$

x_i	y_i	$Y_i = \log_e y_i$	$x_i Y_i$	x_i^2
0	20	2.9957	0	0
1	30	3.40119	3.40119	1
2	52	3.9512	7.9024	4
3	77	4.3438	13.0314	9
4	135	4.9052	19.6208	16
5	211	5.3518	26.759	25
6	326	5.7868	34.7208	36
7	550	6.3099	44.1693	49
8	1052	6.9584	55.6672	64
36		44.0040	205.27721	204

from (1), $44.0040 = 9A + b(36)$

from (2), $205.27721 = A(36) + b(204)$

$\Rightarrow A = 2.9389 \quad , \quad b = 0.4876$

$\Rightarrow \boxed{a = 18.895}$

eqn is $y = (18.895)e^{(0.4876)x}$

Correlation :

It is a statistical analysis which measures & analysis the degree or extend to which two variables fluctuate with reference to eachother

⇒ The correlation expresses the relationship or interdependence of 2 sets of variables.

One variable may be called subject (independent) & another is relative (dependent)

Types of corelation :-

Correlation is classified into many types

1. Positive & negative.
2. Simple & multiple
3. partial & total

Ex:-

* Heights of a father & son.
* Wage & price index

Coefficient of correlation:-

The degree to which 2 variables are inter-related is measured by a coefficient.

The coefficient of correlation b/w 2 variables X, Y is denoted with $\rho(x,y)$.

Karl Pearson's correlation coefficient:-

Karl Pearson, a british biometrician & statician suggested a mathematical ~~for~~ method for measuring the magnitude of linear relationship b/w 2 variables.

⇒ This is known as pearson's coefficient of correlation. It is denoted with 'r'.

⇒ This method is most widely used.

⇒ It is also called product - movement correlation.

The correlation coefficient is given by the formula:

1) $$r = \frac{[N\Sigma XY] - [\Sigma X \Sigma Y]}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

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

where $x = x - \bar{x}$ & $y = y - \bar{y}$

1) Calculate for

x	
y	

Sol)

X	Y
12	14
9	8
8	6
10	9
11	11
13	12
7	13
70	63

r

Proper

* Corre

* If
* If
* If

* It

1. Calculate coefficient of correlation from the following data

x	12	9	8	10	11	13	7
y	14	8	6	9	11	12	3

Sol) Here $N = 7$

x	y	x^2	y^2	xy
12	14	144	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	13	49	169	91
70	63	728	651	676

$$r = \frac{[N\Sigma XY] - [\Sigma X \Sigma Y]}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

$$= \frac{7(676) - (70)63)}{\sqrt{[7(728) - (70)^2][7(651) - (63)^2]}}$$

$$= \frac{322}{339.48}$$

$$= 0.9485$$

Properties :-

* Correlation coefficient is lies blw -1 & 1
  i.e., $-1 \leq r \leq 1$
* If $r = 1$, the correlation is perfectly positive
* If $r = -1$, then correlation is perfectly negative
* If $r = 0$, then there is no correlation blw 2 Variables & they are said to be independent.
* It is a relative measure of association blw

2, Find if there is any significant correlation blw the heights & Weights given below

Height :	57	59	62	63	64	65	55	58	57
Weight :	113	117	126	126	130	129	111	116	112

Here $N = 9$

Mean of x, $\bar{x} = 60$

Mean of Y, $\bar{y} = 120$

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	x^2	Y^2	XY
57	113	-3	-7	9	49	21
59	117	-1	-3	1	9	3
62	126	2	6	4	36	12
63	126	3	6	9	36	18
64	130	4	10	16	100	40
65	129	5	9	25	81	45
55	111	-5	-9	25	81	45
58	116	-2	-4	4	16	8
57	112	-3	-8	9	64	24
				102	472	216

$$r = \frac{\Sigma XY}{\sqrt{(\Sigma X^2)(\Sigma Y^2)}} = \frac{216}{\sqrt{(102)(472)}}$$

$$= \frac{216}{219.417}$$

$$= 0.98442$$

3, Find Karl Pearson's coefficient of correlation from the following data

Wages :	100	101	102	102	100	99	97	98	96	95
cost of living :	98	99	99	97	95	92	95	96	90	91

Here $N = 10$

Mean of x, $\bar{x} = 99$

mean of y, $\bar{y} = 95$

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
100	98	1	3	1	9	3
101	99	2	4	4	16	8
101	99	3	4	9	16	12
102	97	3	2	9	4	6
100	95	1	0	1	0	0
99	91	0	-3	0	9	0
97	95	-2	0	4	0	0
98	94	-1	-1	1	1	1
96	90	-3	-5	9	25	15
95	91	-4	-4	16	16	16
				54	96	61

$$r = \frac{\Sigma XY}{\sqrt{(\Sigma X^2)(\Sigma Y^2)}} = \frac{61}{\sqrt{(54)(96)}}$$

$$= \frac{61}{72}$$

$$= 0.8472$$

4, Calculate KP coefficient b/w age of carst & anual anual maintaintance cost.

Age (years):	2	4	6	7	8	10	12
Anual maintainance:	1600	1500	1800	1900	1700	2100	2000

$N = 7$, $\bar{x} = 6$, $\bar{y} = 1800$

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
2	1600	-4	-200	16	40000	800
4	1500	-2	-300	4	90000	600
6	1800	0	0	0	0	0
7	1900	1	100	1	10000	100
8	1700	2	-100	4	10000	-200
10	2100	4	300	16	90000	1200
12	2000	6				

$$r = \frac{3700}{\sqrt{(77)(280000)}}$$

$$= 0.7968 \qquad (\text{Ans: } 0.8)$$

Rank-corelation (or) Spearman's rank correlation coefficient:-

This method is based on rank & is useful in dealing with qualitative characteristics such as intelligence, honesty, beauty, character etc.

⇒ It can't be measured quatitatively as in the case pearson's coefficient of correlation.

⇒ It is totally depending on ranks to the given observation.

⇒ This method is applicable only individual observations.

⇒ The formula for spearsman's coefficient is

$$\boxed{\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)}}$$

where N = no.of paired observations.

D = diff of 2 ranks

ρ = rank coefficient of correlation.

Properties:

* The value of ρ always lies b/w -1 & 1
* If $\rho = 1$ there is complete agreement in the order of ranks & the direction of rank is same
* If $\rho = -1$ then there is complete disagreement in the order of ranks & they are in opposite directions.

1) Following are rank obtained by 10 students

Statistics:	1	2	3	4	5	6	7	8	9	10
Maths:	2	4	1	5	3	9	7	10	6	8

To what extent, the knowledge of the students in 2 subjects

Here $N = 10$

Rank correlation coefficient,

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

Rank in stat (x)	Rank in math (y)	$D = x - y$	D^2
1	2	-1	1
2	4	-2	4
3	1	2	4
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	3	9
10	8	2	4
			40

$$\rho = 1 - \frac{6(40)}{10(99)} = 1 - 0.2424 = 0.7576$$

2) The ranks of 16 students in maths & satat are as follows.

(1,1) (2,10) (3,3) (4,4) (5,5) (6,7) (7,2) (8,6), (9,8) (10,11), (11,15), (12,9), (13,14), (14,12), (15,16), (16,13)

x	y	d = x-y	d^2
1	1	0	0
2	10	-8	64
3	3	0	0
4	4	0	0
5	5	0	0
6	7	-1	1
7	2	5	25
8	6	2	4
9	8	1	1
10	11	-1	1
11	15	-4	16
12	9	3	9
13	14	-1	1
14	12	2	4
15	16	-1	1
16	13	3	9
			136

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)} = 1 - \frac{6(136)}{16(256-1)}$$

$$= 0.8$$

3. A random sample of 5 students colleges is are selected & their grades in Maths & statistics are given below:-

	1	2	3	4	5
Maths :	85	60	73	40	90
stat :	93	75	65	50	80

Calculate pearson's rank correlation coefficient

Maths	Rank (x)	stat (Y)	Rank (y)	D = x-y	D^2
85	2	93	1	1	1
60	4	75	3	1	1
73	3	65	4	-1	1
40	5	50	5	0	0
90	1	80	2	-1	1
					4

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)} = 1 - \frac{6(4)}{5(25-1)}$$

$$= 1 - \frac{24}{5(24)} = 1 - 0.2 = 0.8$$

4) 10 competitors in a musical test were ranked by the 3 judges A, B & C in the following order.

Ranks by A:	1	6	5	10	3	2	4	9	7	8
Ranks by B:	3	5	8	4	7	10	2	1	6	9
Ranks by C:	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pair of judges has the nearest approach to common linkings in music.

Rank by A (x)	Rank by B (y)	Rank by C (z)	$D_1^2=(x-y)^2$	$D_2^2=(y-z)^2$	$D_3^2=(z-x)^2$
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
			200	214	60

$$\rho_1 = 1 - \frac{6\Sigma D_1^2}{N(N^2-1)} = 1 - \frac{6(200)}{10(100-1)} = -0.2121$$

$$\rho_2(y,z) = 1 - \frac{6\Sigma D_2^2}{N(N^2-1)} = 1 - \frac{6(214)}{10(99)} = -0.2969$$

$$\rho_3(z,x) = 1 - \frac{6\Sigma D_3^2}{N(N^2-1)} = 1 - \frac{6(60)}{10(99)} = \frac{33}{99} = \frac{11}{33}$$

$$= 0.6364$$

Since $\rho_3(z,x)$ is maximum.

Hence the pair of judges A & C has nearest approach

Rank correlation coefficient for equal or repeated ranks :-

In this case, common ranks are given to repeated items.

The commonrank is the average of the rank which these items would have assume if they were different from each other & the next item will get the rank.

⇒ In this case, the rank correlation coefficient is given by the formula :

$$\rho = 1 - 6\left\{\frac{\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \dots}{N^3 - N}\right\}$$

Where m = no. of items whose ranks are common.

1. From the following data, calculate the rank correlation coef after making adjustment for tide ranks.

x :	48	33	40	9	16	16	65	24	16	57
y :	13	13	24	6	15	4	20	9	6	19

X	Rank (x)	y	Rank (y)	D = x - y	D^2
48	3	13	5.5	-2.5	6.25
33	5	13	5.5	-0.5	0.25
40	4	24	1	3	9
9	10	6	8.5	1.5	2.25
16	8	15	4	4	16
16	8	4	10	-2	4
65	1	20	2	-1	1
24	6	9	7	-1	1
16	8	6	8.5	-0.5	0.25
57	2	19	3	-1	1
					41

Here 16 repeated 3 times in x- items i.e., $m_1 = 3$

$$\rho = 1 - 6\left\{\frac{\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots}{N^3 - N}\right\}$$

$$\rho = 1 - 6\left\{\frac{41 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)}{10^3 - 10}\right\}$$

$$\rho = 1 - 6\left[\frac{43.4}{990}\right]$$

$$\boxed{\rho = 0.7334}$$

2. Obtain the rank correlation coef for following

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

X	Rank (x)	Y	rank (y)	D = x - y	D^2
68	4	62	5	-1	1
64	6	58	7	-1	1
75	2.5	68	3.5	-1	1
50	9	45	10	-1	1
64	6	81	1	5	25
80	1	60	6	-5	25
75	2.5	68	3.5	-1	1
40	10	48	9	1	1
55	8	50	8	0	0
64	6	70	2	4	16
					72

5+6+7 = 18/3

Here $m_1 = 3$, $m_2 = 2$, $m_3 = 2$

$$\rho = 1 - 6\left[\frac{72 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(8 - 2) + \frac{1}{12}(8 - 2)}{10^3 - 10}\right]$$

$$= 1 - 6\left(\frac{72 + 3}{990}\right)$$

$$= 1 - \frac{6(75)}{990} = 1 - 0.4545 = 0.5455$$

Regression:

The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called regression.

⇒ The line describes in the avg relationship blw 2 variables is known as line of regression or estimating line.

Uses:

* It is used to estimate the relation blw 2 economic variables like income & expenditure.
* It is highly valuable tool in economics & buisiness.
* we can calculate coef of correlation & coef of determination with the help of the regression coefficient.

Procedure to calculate regression line

⇒ It is a st. line fitted to the data by the method of least squares.

1, Regression eqn of Y on x : $(y = a+bx)$

$$\Sigma y = Na + b\Sigma x$$

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

2, Regression eqn of x on Y : $(x = a+by)$

$$\Sigma x = Na + b\Sigma y$$

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

1) Determine the eqn of the st.line which best fits the data.

X :	10	12	13	16	17	20	25
Y :	10	22	24	27	29	33	37

X	Y	X^2	XY^2
10	10	100	100
12	22	144	264
13	24	169	312
16	27	256	432
17	29	289	493
20	33	400	660
25	37	625	925
113	182	1983	3186

Regression eqn of Y on x :

$$Y = a + bx$$

Normal eqns :- $\Sigma Y = Na + b\Sigma x$

$$182 = 7(a) + b(113) \longrightarrow (1)$$

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

$$3186 = a(113) + b(1983) \longrightarrow (2)$$

Solving (1) & (2), $a = 0.7985$, $b = 1.561$

$$\therefore \quad Y = 0.7985 + b(1.561)$$

NOTE:-

In The line $Y = a + bx$, b is called regression coefficient & also in the line $x = a + bY$, b is called regression coefficient.

1) Heights of fathers & sons are given in inches

Ht of fathers: 65 66 67 67 68 69 71 73
Ht of sons : 67 68 64 68 72 70 69 70

From the 2 lines of regression & calculate the expected avg height of the son when the ht of the father is 67.5 inches.

Sol) Regression eqn.

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	64	4489	4096	4288
67	68	4489	4624	4556
68	72	4624	5184	4896
69	70	4761	4900	4830
71	69	5041	4761	4899
73	70	5329	4900	5110
546	548	37314	37578	37422

i) Regression eqn of Y on X : $(Y = a + bX)$

$$\Sigma Y = Na + b\Sigma X$$

$$548 = 8(a) + b(546) \longrightarrow (1)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

$$37422 = a(546) + b(37314) \longrightarrow (2)$$

from (1) & (2), $a = 39.545$, $b = 0.4242$

$$\therefore Y = 39.545 + (0.4242)X \longrightarrow (5)$$

ii) Regression eqn of X on Y : $(X = a + bY)$

$$\Sigma X = Na + b\Sigma Y$$

$$546 = 8(a) + b(548) \longrightarrow (3)$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

$$37422 = a(548) + b(37578) \longrightarrow (4)$$

from (3) & (4), $a = 32.287$, $b = 0.525$

$$x = 32.287 + 0.525\,y \longrightarrow (4)$$

when $x = 67.5$ inches, the height of son

from eqn (5),

$$Y = 39.545 + (0.4242)\,67.5$$

$$\boxed{Y = 68.1785}$$

from eqn (6),

$$67.5 = 32.287 + (0.525)\,Y$$

$$\boxed{Y = 67.07}$$

Regression Coefficient:

The regression coefficient of Y on x is

$$\gamma\left(\frac{\sigma y}{\sigma x}\right)$$

and regression coefficient of x on Y is $\gamma\left(\frac{\sigma x}{\sigma y}\right)$

Where γ = correlation coefficient

$$\gamma = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}}$$

$$\sigma x = \sqrt{\frac{\Sigma x^2}{N} - (\bar{x})^2}$$

$$\sigma y = \sqrt{\frac{\Sigma y^2}{N} - (\bar{y})^2}$$

Note:-

Regression equation Y on x $\rightarrow$ $y - \bar{y} = \gamma \frac{\sigma y}{\sigma x}(x - \bar{x})$

x on Y $\rightarrow$ $x - \bar{x} = \gamma \frac{\sigma x}{\sigma y}(y - \bar{y})$

UNIT-5

SAMPLING DISTRIBUTIONS

Population:

The term population or universe is an aggregate or totality of statistical data forming a subject of investigation.

Examples:-

National banks in India, heights of Indians, lengths of fish in a particular lake etc.

⇒ A no.of observations in the population is defined to be size of the population & is denoted with 'N'

Example:-

1. In a school, 600 students are classified according to blood group then we say that population size is 600 & is finite.
2. The observations obtained by measuring the atmospheric pressure everyday from the past into the future, which is infinite.

Sampling:

A portion of a population which is examined with a view to determine the population characteristics is called a sampling.

⇒ i.e., a sample is a subset of population

⇒ The process of selecting a sample is called sampling.

⇒ The no.of objects in the sample is the size of sample.
(denoted by n)

cars produced in India is the population & nanocars is the sample.

Note:-

If $(n < 30)$ then it is called small sample.

If $(n \geq 30)$ then it is called large sample.

Types of Sampling:

1. purposive
2. Random
3. Simple
4. Stratified

Notations in Population

⇒ Mean of population (μ)

⇒ Variance (σ^2)

⇒ standard deviation (σ)

⇒ Population Proportions (P)

Notations in sample

⇒ Sample mean $(\bar{x})$

⇒ Sample Variance (s^2)

⇒ standard deviation (s)

⇒ Sample proportion (p)

Notations in Sampling distribution

⇒ Mean of sampling distribution $(\mu_{\bar{x}})$

⇒ Variance $(\sigma_{\bar{x}}^2)$

⇒ standard deviation $\frac{\sigma}{\sqrt{n}}$

⇒

* If each element of a population is selected more than once then it is called sampling with replacement.

and if the element can't be selected then it is called sampling without replacement.

Note:

⇒ If N is the size of a population & n is the sample size then the no.of samples with replacement is given by N^n (infinite)

⇒ The no.of samples without replacement is N_{C_n} (finite)

⇒

Sample mean:-

let $x_1, x_2, \ldots, x_n$ represents random samples of size (n) then mean of sample is given by the formula :-

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ (or) } \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Sample Variance :-

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \text{ (or) } \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n}$$

Sample standard deviation:

$$S.D = (s) = \sqrt{\text{variance}}$$

* Mean of sampling distribution

Case (i) :

Infinite population / with replacement.

⇒ The mean of sampling distribution of means is

$$\mu_{\bar{x}} = \mu$$

⇒ The variance of sampling distribution of means is

$$\boxed{\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}}$$

⇒ The standard deviation of sampling distribution of means is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

case-(ii)

Finite population / without replacement

⇒ The mean of sampling distribution of means is

$$\mu_{\bar{x}} = \mu$$

⇒ The variance is

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1}\right)$$

⇒ The standard deviation of sampling distribution of means is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}$$

* Here, SD = standard error

* The factor $\frac{N-n}{N-1}$ is called finite population correction factor.

1) Find the value of population correction factor when $n = 10$ & $N = 1000$

We know that,

population correction factor $= \frac{N-n}{N-1}$

$= \frac{1000-10}{1000-1}$

$= \frac{9990}{999}$

$= 0.999$

2) A sample is collected from the items produced by factory. The sample size is 81, SD of the popalation is 3. Find the standard error of the mean of sampling distribution.

given, $n = 81$

$\sigma = 3$

∴ Standard error of sampling distribution

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{3}{\sqrt{81}}$$

$$= 0.333$$

3) How many different samples of size 2 can be chosen from a finite population of size 25? How can we select

Sol) Given that $N = 25$ & $n = 2$

The no.of samples we can take from the finite population is ${}^{N}C_{n} = {}^{25}C_{2} = 300$.

4) A population consists of 5 members 2, 3, 6, 8, 11. Consider all possible samples of size 2 which can be drawn with replacement from this population. Then find

i) mean of the population

ii) Variance of the population

iii) Standard deviation of the population

iv) standard deviation of sampling distribution of means

v) mean of the sampling distribution of means.

vi)

Sol) Given that,

Sample size $(n) = 2$

and population size $(N) = 5$

i) Mean of population, $\mu = \frac{2+3+6+8+11}{5} = 6$

ii) Variance of population,

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

$$\sigma^2 = \frac{(2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2}{5}$$

$$= \frac{16 + 9 + 0 + 4 + 25}{5}$$

$$= \frac{54}{5}$$

$$= 10.8$$

iii) SD, $\sigma = \sqrt{10.8} = 3.286$

iv) The total no. of samples from infinite (with replacement) is given by $N^n = 5^2 = 25$

The sample set is

$$\begin{Bmatrix} (2,2) & (2,3) & (2,6) & (2,8) & (2,11) \\ (3,2) & (3,3) & (3,6) & (3,8) & (3,11) \\ (6,2) & (6,3) & (6,6) & (6,8) & (6,11) \\ (8,2) & (8,3) & (8,6) & (8,8) & (8,11) \\ (11,2) & (11,3) & (11,6) & (11,8) & (11,11) \end{Bmatrix}$$

The means of samples are

$$\begin{Bmatrix} 2 & 2.5 & 4 & 5 & 6.5 \\ 2.5 & 3 & 4.5 & 5.5 & 7 \\ 4 & 4.5 & 6 & 7 & 8.5 \\ 5 & 5.5 & 7 & 8 & 9.5 \\ 6.5 & 7 & 8.5 & 9.5 & 11 \end{Bmatrix}$$

The mean of sampling distribution of means is

$$\mu_{\bar{x}} = \mu = \frac{\sum_{i=1}^{25} \bar{x}_i}{n}$$

$$= \frac{2 + 2.5 + 4 + 5 + 6.5 + \cdots + 11}{25}$$

$$= \frac{150}{25} = 6$$

iii) Variance, $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{\sum_{i=1}^{n} (\bar{x}_i - \mu_{\bar{x}})^2}{n}$

$$= \frac{(2-6)^2 + (2.5-6)^2 + \cdots + (11-6)^2}{25}$$

$$= \frac{135}{25}$$

5) Same problem but withoutreplacement.

Total no.of samples $N_{C_n} = {}^5C_2 = 10$

6) A population consists of 5, 10, 14, 18, 13, 24. Consider all possible samples of size 2. which can be drawn without replacement from the population. Then find

i. mean of popopulation ii. SD of population

iii. mean of sampling distribution of means

iv. SD of " " " "

Sol) Here $n = 2$ & $N = 6$

i. mean of population $(\bar{x})$ $\mu = \frac{5+10+14+18+13+24}{6}$

$= 14$

ii. Variance $(\sigma^2) = \frac{(5-14)^2 + (10-14)^2 + \cdots + (24-14)^2}{6}$

$= 35.667$

$\Rightarrow \sigma = \sqrt{35.667} = 5.972$

iii. Total no.of samples $= {}^N C_n = {}^6C_2 = 15$

Sample set is

$$\left\{ \begin{matrix} (5,10) & (5,14) & (5,18) & (5,13) & (5,24) \\ (10,14) & (10,18) & (10,13) & (10,24) & \\ (14,18) & (14,13) & (14,24) & & \\ (18,13) & (18,24) & (13,24) & & \end{matrix} \right\}$$

The mean samples are

$$\left\{ \begin{matrix} 7.5, & 9.5, & 11.5, & 9, & 14.5 \\ 12, & 14, & 11.5, & 17, & 16 \\ 13.5, & 19, & 15.5, & 21, & 18.5 \end{matrix} \right\}$$

The mean of sampling distribution of means is

$$\mu_{\bar{x}} = \frac{7.5 + 9.5 + 11.5 + \cdots + 18.5}{15}$$

$= 14$

iv) Variance $\sigma_{\bar{x}}^2 = \frac{\Sigma (x_i - \mu_{\bar{x}})^2}{n}$

$= \frac{(7.5-14)^2 + \cdots + (18.5-14)^2}{15}$

$= 14.2667$

$\Rightarrow SD = \sqrt{14.2667}$

$= 3.7712$

(or) $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{(5.972)^2}{2}\left(\frac{6-2}{6-1}\right)$

$= 14.2668$

7) samples of size 2 are taken from the population 1, 2, 3, 4, 5, 6

i) with replacement ii) without replacement.

then find

(a) mean of population (b) S.D of population

(c) mean of sampling distribution of means

(d) SD " "

Verify that means of sampling dist = mean of populat & SD of means of sampling dist are not equal SD of population.

Sol) Here, $n = 2$, $N = 6$

i) mean of population $(\mu) = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$

ii) Variance $(\sigma^2) = \frac{(1-3.5)^2 + \cdots + (6-3.5)^2}{6}$

$= 2.9166$

$SD = \sqrt{2.9166} = 1.7078$

iii) with replacement :-

The means of samples are

$$\left\{\begin{array}{cccccc} 1, & 1.5, & 2, & 2.5, & 3, & 3.5 \\ 1.5, & 2, & 2.5, & 3, & 3.5, & 4 \\ 2, & 2.5, & 3, & 3.5, & 4, & 4.5 \\ 2.5, & 3, & 3.5, & 4, & 4.5, & 5 \\ 3, & 3.5, & 4, & 4.5, & 5, & 5.5 \\ 3.5, & 4, & 4.5, & 5, & 6, & 5.5 \end{array}\right\}$$

The mean of sampling distribution $= \frac{1+1.5+2+\cdots+5.5}{36}$

$= 3.5$

Variance $\sigma_{\bar{x}}^2 = \frac{(1-3.5)^2+(1.5-3.5)^2+\cdots+(5.5-3.5)^2}{36}$

$= 1.4583$

$\sigma_{\bar{x}} = 1.2076$

Without replacement:

Total no. of samples $= {}^{N}C_{n} = {}^{6}C_{2} = 15$

The sample set is

$$\left\{\begin{array}{ccccc} (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,3) & (2,4) & (2,5) & (2,6) & (3,4) \\ (3,5) & (3,6) & (4,5) & (4,6) & (5,6) \end{array}\right\}$$

The means of samples are :-

$$\left\{\begin{array}{ccccc} 1.5, & 2, & 2.5, & 3, & 3.5 \\ 2.5, & 3, & 3.5, & 4, & 3.5 \\ 4, & 4.5, & 4.5, & 5, & 5.5 \end{array}\right\}$$

The means of sampling distribution $= \frac{1.5+2+2.5+\cdots+5.5}{15}$

$= 3.5$

Variance $\sigma_{\bar{x}}^2 = \frac{(1.5-3.5)^2+\cdots}{15}$

8) If the population is 3, 6, 9, 15, 27.

i) list of all possible samples of size 3 that can be taken without replacement from the finite population.

ii) Calculate the mean of each of the sampling distribution of means

iii) Find SD of sampling distribution of means.

Here $N = 5$

i) The total no. of samples of size 3 taken from finite population is $NC_n = {}^5C_3 = 10$:

The samples are :-

$\{(3,6,9), (3,6,15), (3,6,27), (3,9,15), (3,9,27)$
$(3,15,27), (6,9,15), (6,9,27), (6,15,27), (9,15,27)\}$

The means of samples are

$\{6, 8, 12, 13, 9, 15, 10, 14, 16, 17\}$

The means of sampling distribution

$$\mu_{\bar{x}} = \frac{6+8+12+13+9+15+\cdots+17}{10}$$

$$\mu_{\bar{x}} = 12$$

$$\text{Variance, } \sigma_{\bar{x}}^2 = \frac{(6-12)^2 + (8-12)^2 + \cdots + (17-12)^2}{10}$$

$$= 12$$

$$\boxed{\sigma_{\bar{x}} = 3.464}$$

9) when a sample is taken from an infinite population. what happened to the standard error of the mean when the sample size is decreased from 800 to 200?

we know that,

standard error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

given that $n_1 = 800$ & $n_2 = 200$.

Now, $SE_1 = \frac{\sigma}{\sqrt{n_1}} = \frac{\sigma}{\sqrt{800}} = \frac{\sigma}{20\sqrt{2}}$

$SE_2 = \frac{\sigma}{\sqrt{n_2}} = \frac{\sigma}{\sqrt{200}} = \frac{\sigma}{10\sqrt{2}}$

$= 2\left(\frac{\sigma}{20\sqrt{2}}\right)$

$= 2SE_1$

$\Rightarrow \boxed{SE_2 = 2SE_1}$

Hence if sample size is reduced from 800 to 200 then standard error of mean will be multiplied by 2.

Central Limit theorem

let $\bar{x}$ be the mean of random sample taken from a population having the mean μ & SD (σ) then

$$\boxed{z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}$$

is a random variable whose distribution function approaches to the standard normal distribution (N) as $n \to \infty$

1) The mean height of students in a college is 155cm and standard deviation is 15. Now what is the probability that the mean height of 36 students is less than 157 cm.

Sol) given that,

mean of population $\mu = 155$cm

Standard deviation $\sigma = 15$

Here sample size $(n) = 36$

We know that, $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

when $\bar{x} = 157$, $z = \frac{157 - 155}{\frac{15}{\sqrt{36}}}$

$$z = \frac{2}{15/6}$$

$$z = \frac{2 \times 2}{5}$$

$$z = 4/5$$

$$z = 0.8$$

Now $P(\bar{x} < 157) = P(z < 0.8)$

$= A(0.8)$

$= 0.7881$

2) A random sample of size 100 taken from an infinite population having the mean 76 & variance is $\sigma^2 = 256$. What is the probability that $\bar{x}$ will be b/w 78, 75 & 78

Sol) given,

mean of population $\mu = 76$

standard deviation, $\sigma = \sqrt{256} = 16$

Here $n = 100$

We know, $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

Now at $\bar{x}_1 = 75$, $z_1 = \frac{75 - 76}{\left(\frac{16}{\sqrt{100}}\right)} = -1 \times \frac{10}{16} = -0.625$

at $\bar{x}_2 = 78$, $z_2 = \frac{78 - 76}{\left(\frac{16}{\sqrt{100}}\right)} = 2 \times \frac{10}{16_8} = 1.25$

Now $P(75 \leq \bar{x} \leq 78) = P(-0.625 < z < 1.25)$

$= A(1.25) - A(-0.625)$

$= 0.8944 - 0.2676$

$= 0.6268.$

3) A random sample of size 64 is taken from a normal population with $\mu = 51.4$ & $\sigma = 6.8$ what is the probability that the mean of the sample will

i, exceed 52.9 ii, b/w 50.5 & 52.3

iii, less than 50.6

Sol,)

given, $\mu = 51.4$; $\sigma = 6.8$, $n = 64$

We know, $z = \frac{\bar{x} - \mu}{(\frac{\sigma}{\sqrt{n}})}$

Now at $\bar{x}_1 = 52.9$, $z_1 = \frac{52.9 - 51.4}{(\frac{6.8}{\sqrt{64}})} = 1.76$

at $\bar{x}_2 = 50.5$, $z_2 = \frac{50.5 - 51.4}{(\frac{6.8}{\sqrt{64}})} = -1.0588$

at $\bar{x}_3 = 52.3$, $z_3 = \frac{52.3 - 51.4}{(\frac{6.8}{\sqrt{64}})} = 1.0588$

at $\bar{x}_4 = 50.6$, $z_4 = \frac{50.6 - 51.4}{(\frac{6.8}{\sqrt{64}})} = -0.94117$

i, $P(\bar{x} > 52.9) = P(z > 1.76)$

$= 1 - P(z < 1.76)$

$= 1 - 0.9608$

$= 0.0392$

ii, $P(50.5 \leq \bar{x} \leq 52.3) = P(-1.0588 \leq z < 1.0588)$

$= A(1.0588) - A(-1.0588)$

$= 0.71$

$P(\bar{x} < 50.6) = P(z < -0.94117)$

$= A(-0.94117)$

$= 0.1736$

4) A normal population has a mean of 0.1 & S.D is 2.1 find the probability that mean of a sample size 900 will be negative?

Sol) given, $\mu = 0.1$, SD $(\sigma) = 2.1$, $n = 900$

We know, $z = \frac{\bar{x} - \mu}{(\frac{\sigma}{\sqrt{n}})}$

$$z = \frac{\bar{x} - 0.1}{(\frac{2.1}{\sqrt{900}})}$$

$$z = \frac{\bar{x} - 0.1}{0.07}$$

$$\Rightarrow \bar{x} = 0.07z + 0.1$$

Now $P(\bar{x} < 0) = P(0.07z + 0.1 < 0)$

$= P(0.07z < -0.1)$

$= P(z < \frac{-0.1}{0.07})$

$= P(z < -1.4285)$

$= A(-1.4285)$

$= 0.0778$

5) Mean voltage of a battery is 15 & SD is 0.2 Find the probability that such 4 batteries connected in series will have combined voltage of 60.8 (or) more volts

$\mu = 15$, $\sigma = 0.2$, $n = 4$

$$z = \frac{\bar{x} - \mu}{(\frac{\sigma}{\sqrt{n}})} = \frac{\bar{x} - 15}{(\frac{0.2}{\sqrt{4}})} = \frac{\bar{x} - 15}{0.1}$$

$$\Rightarrow \boxed{\bar{x} = 0.1z + 15}$$

Now $P(\bar{x} \geq 60.8) = P(0.1z + 15 \geq 60.8)$

$= P(0.1z > 60.8 - 15)$

$= P(0.1z > 45.8)$

$= P(z > 458)$

Given, $\mu = 15$, $\sigma = 0.2$, $n = 4$

let us take four batteries are x_1, x_2, x_3, x_4

$P(x_1 + x_2 + x_3 + x_4 \geq 60.8) = ?$

Now $P(x_1 + x_2 + x_3 + x_4 \geq 60.8) = P\left(\frac{x_1 + x_2 + x_3 + x_4}{4} \geq \frac{60.8}{4}\right)$

$= P(\bar{x} \geq 15.2)$

at $\bar{x} = 15.2$, $z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{15.2 - 15}{\left(\frac{0.2}{\sqrt{4}}\right)} = \frac{0.2}{0.1} = 2$

$\therefore P(x_1 + x_2 + x_3 + x_4 \geq 60.8) = P(\bar{x} \geq 15.2)$

$= P(z \geq 2)$

$= 1 - P(z < 2)$

$= 1 - 0.9772$

$= 0.0228$

6) The mean lifetime of light bulbs produced by a company is 1500 hours and SD is 150 hours Find the probability that lighting will take place for

i. Atleast 5000 hours

ii. Atmost 4200 hours

if 3 bulbs are connected such that when one bulb burns out, another bulb will glow on. Assume that lifetimes are normally distributed.

i) given that,

$\mu = 1500$, $\sigma = 150$, $n = 3$

let us take 3 bulbs as x_1, x_2, x_3

i. Now $P(x_1 + x_2 + x_3 \geq 5000) = P\left(\frac{x_1 + x_2 + x_3}{3} \geq 5000\right)$

$= P(\bar{x} \geq 1666.66)$

at $\bar{x} = 1666.66$, $z = \frac{1666.67 - 1500}{\left(\frac{150}{\sqrt{3}}\right)} = 1.91911$

$\therefore P(x_1 + x_2 + x_3 \geq 5000) = P(\bar{x} \geq 1666.66)$

$= P(z \geq 1.9191)$

$= 1 - P(z < 1.9191)$

$= 1 - 0.9719$

$= 0.0281$

ii) $P(x_1 + x_2 + x_3 \leq 4200) = P(\bar{x} \leq 1400)$

at $\bar{x} = 1400$, $z = \frac{1400 - 1500}{\left(\frac{150}{\sqrt{3}}\right)} = -1.1547$

$\therefore P(\bar{x} \leq 1400) = P(\bar{z} < -1.1547)$

$= 0.1251$

Estimation:

⇒ Quantity is appearing in distribution such as p in the binomial distribution & μ, σ are in the normal distribution are called parameters.

Estimate:

An estimate is a statement made to find an unknown population parameter.

Estimater:

The procedure or rule to determine an unknown population parameter is called an estimater.

Types of estimation:

There are 2 kinds of estimates to determine

Point Estimation:

If an estimate of population is given by a single value then it is called point estimation of the parameter of population.

Interval Estimation:

If an estimate of population is given by 2 values blw which the parameter may consider to lie is called Interval estim

Example:

If the ht of student is measured as 162 cm then the measurement gives point estimation but ht is lies blw (163 ± 3.5) cm i.e., 166.5 cm & 159.5 cm. then the measarement gives an interval estimation.

Note:

The sample mean $\bar{x}$ is point estimate to the population mean (μ) & sampling Variance (s^2) is a point estimate of the population Variance (σ^2)

Def:

⇒ A point estimator is a statistic for estimating the population parameter (θ) & is denoted by ($\hat{\theta}$)

Properties:

* A good estimator is one, which is a closer Value to the true value of the parameter as possible.

* The imp properties of a good estimator are

→ consistency:

A estimator ($\hat{\theta}$) of a parameter θ

is consistent

→ Unbiasedness:

A statis

if es expecta

For example:

⇒ Sampling

i.e., E(

⇒ sample v

of σ2

→ Efficiency

A

efficient

if

a) both

b) and

→ sufficien

An e

for

inform

the

→ Con

Confi

In an

if

on

popu

incl

spe

int

is consistent if it converges to θ as $n \to \infty$

→ Unbiasedness:

A statistic $(\hat{\theta})$ is said to be unbaised estimate of θ if expectation $E(\hat{\theta}) = \theta$

For example:

⇒ sampling mean $(\bar{x})$ is unbaised estimator of μ

i.e., $E(\bar{x}) = \mu$

⇒ sample variance (s^2) is unbaised estimator of σ^2

i.e., $E(s^2) = \sigma^2$

→ Efficiency:

A statistic $(\hat{\theta}_1)$ is said to be more efficient unbaised estimator than statistic $(\hat{\theta}_2)$ if

a) both $\hat{\theta}_1$ & $\hat{\theta}_2$ are unbaised estimators of θ.

b) and variance, $V(\hat{\theta}_1) < V(\hat{\theta}_2)$

→ Sufficiency

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

→ Con

Confidence Interval:

In an interval of the population paramter (θ), if we can find 2 quantities t_1 & t_2 based on sample observation drawn from the population such that the parameter (θ) is included in the interval $[t_1, t_2]$ in a specified percentage of cases then this interval is called a confidence interval.

Confidence Interval :	98 %	95 %	90 %
LOS (α) :	1 %	5 %	10 %
$z_{\alpha/2}$:	2.58	1.96	1.64

NOTE:

⇒ The maximum error of estimate E with $1-\alpha$ probability for large samples is given by

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

⇒ The maximum error estimate E with $1-\alpha$ probability for small samples is given by

$$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Where n = sample size

s = SD of sample

⇒ If $\bar{x}$ is the mean of random sample of size n from the population with known variance σ^2 and $(1-\alpha)100\%$ confidence interval μ is given by (for large samples)

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

where $z_{\alpha/2}$ is the z value leaving an area of $\alpha/2$ to the right

⇒ If $\bar{x}$ & s are the mean & SD of random sample from a normal population with unknown variance (σ^2), $(1-\alpha)100\%$ confidence interval μ is given by (for small samples).

$$\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

Where $t_{\alpha/2}$ is the t value with $\nu = n-1$ degrees of freedom leaving an area $\alpha/2$ to the right

⇒ Confidence interval for single proportion is

$$\left(p - z_{\alpha/2}\sqrt{\frac{pq}{n}}, p + z_{\alpha/2}\sqrt{\frac{pq}{n}}\right)$$

1) A random sample of size 100 has a SD of 5 What can you say about the maximum error with 95% confidence ?

given, sample size (n) = 100 (large sample)

SD (σ) = 5

confidence interval percentage = 95%

⇒ $\alpha = 5\%$

⇒ $\alpha = 0.05$

Now $z_{\alpha/2} = 1.96$

Now maximum error (E) = $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$= (1.96)\left(\frac{5}{\sqrt{100}}\right)$

$= (1.96)\frac{5}{10}$

$= 0.98$

2) what is the maximum error one can expect to make with probability 0.90 when using the mean of a sample size 64 to estimate the mean of population with $\sigma^2 = 2.56$

Sol:) given, $n = 64$, $\sigma^2 = 2.56$

$\Rightarrow \sigma = 1.6$

confidence interval, $= 0.90$

$\Rightarrow \alpha = 10\%$

$\Rightarrow \alpha = 0.01$

$\Rightarrow z_{\alpha/2} = 1.64$

Maximum error $(E) = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$= 1.64 \frac{1.6}{\sqrt{64}}$

$= (1.64) \frac{1.6}{8}$ 0.2

$= 0.328$

3) Assuming that $\sigma = 20.0$. How large a sample be taken to assert with probability 0.95 that the sample mean will'nt differ from the true mean by more than 3.0 points.

given, $\sigma = 20$, $\alpha = 5\% = 0.05$

$z_{\alpha/2} = 1.96$

$E = 3.0$

$\Rightarrow E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$\Rightarrow 3 = (1.96) \frac{20}{\sqrt{n}}$

$\Rightarrow \sqrt{n} = 13.06$

$\Rightarrow n^2 = 170.56$

$\Rightarrow n = 171$

4) If we can assert with 95% that the maximum error is 0.05% of P = 0.2. Find the size of the sample.

Sol) confidence interval, = 95%

$\alpha = 5\% = 0.05$

$\Rightarrow z_{\alpha/2} = 1.96$

maximum error, $E = 0.05$

given, $P = 0.2$

$\Rightarrow q = 1 - P = 1 - 0.2 = 0.8$

We know, $\sigma = \sqrt{npq}$

maximum error, $E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$\Rightarrow 0.05 = (1.96) \frac{\sqrt{0.2 \times 0.8}}{\sqrt{n}}$

$\Rightarrow \sqrt{n} = \left(\frac{1.96}{0.05}\right)^2 (0.2 \times 0.8)$

$\Rightarrow n = 246$

5) A sample of size 300 was taken whose variance is 225 & mean 54. construct 95% confidence interval of the mean.

Sol) given, sample size $(n) = 300$

mean $(\mu) = 54$

Variance $(\sigma^2) = 225$

$\sigma = 15$

confidence interval percentage = 95%

$\Rightarrow \alpha = 5\% = 0.05$

$\Rightarrow \alpha/2 = 0.025$

at $\alpha/2$, $z_{\alpha/2} = 1.96$

95% confidence interval $= \left(\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \right)$

$= 54 - (1.96)\left(\frac{15}{\sqrt{300}}\right), 54 + (1.96)\left(\frac{15}{\sqrt{300}}\right)$

$= (52.3025, 55.69)$

6) A random sample of 500 coins on a heated plate resulted in an average temperature of 73.54°F with a SD of 2.79°F. Find a 99% confidence interval for the avg temperature of the plate.

Sol) given, $n = 500$, $\bar{x} = 73.54°F$, SD $(\sigma) = 2.79$

$\alpha = 1\% = 0.01$

$z_{\alpha/2} = 2.58$

confidence interval $= \left(\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right)$

$= \left(73.54 - 2.58\left(\frac{2.79}{\sqrt{500}}\right), 73.54 + 2.58\left(\frac{2.79}{\sqrt{500}}\right)\right)$

$= (73.218, 73.86)$

7) The mean & SD of a population are 11795 & 14054 respectively. What can one assert with 95% confidence about the maximum error if $\bar{x} = 11,795$ & $n = 50$. And also construct 95% confidence interval for the true mean.

Sol) given, $\bar{x} = 11,795$, $n = 50$, $\sigma = 14054$

confidence % = 95%

$\Rightarrow \alpha = 5\%$

$z_{\alpha/2} = 1.96$

Maximum error $(E) = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

$= 1.96\left(\frac{14054}{\sqrt{50}}\right)$

$= 253463$ 389557.6

confidence interval $= (\bar{x} - E, \bar{x} + E)$

$= (15690.57, 7899.43)$

8) A sample of 400 items is taken from a population whose SD is 10. The mean of sample is 40. Calculate 95% confidence interval for the population.

Sol) given, $\bar{x} = 40$, $\sigma = 10$, $n = 400$

$\alpha = 5\%$

$z_{\alpha/2} = 1.96$

Error $(E) = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = 1.96\left(\frac{10}{\sqrt{400}}\right)$

$= 1.96\left(\frac{10}{20}\right)$

$= 0.98$

$CI = (40.98, 39.02)$

$= (39.02, 40.98)$

9) In a random sample of size 400 industrial accidents, it was found that 231 were due to atleast partially to unsafe working conditions. Construct 99% confidence interval for the corresponding true proportion.

Sol) Given,

$n = 400$, $x = 231$

Sample proportion $(P) = \frac{x}{n} = \frac{231}{400}$

$q = 1 - P = 1 - \frac{231}{400} = \frac{169}{400}$

confidence $= 99\%$

$\Rightarrow \alpha = 1\%$

$z_{\alpha/2} = 2.58$

confidence interval $= \left(P - z_{\alpha/2}\sqrt{\frac{pq}{n}}, P + z_{\alpha/2}\sqrt{\frac{pq}{n}}\right)$

$= (0.5128, 0.6412)$

10) Among 900 people in a state, 90 are found to be chapathi eaters. Construct 99% CI for the true proportions.

Sol)

$n = 900$

$p = \frac{90}{900} = \frac{1}{10}$

$q = 9/10$

$\alpha = 1\%$, $z_{\alpha/2} = 2.58$

$CI = \left(p - z_{\alpha/2}\sqrt{\frac{pq}{n}}, \; p + z_{\alpha/2}\sqrt{\frac{pq}{n}} \right)$

$= \left(\frac{1}{10} - 0.01, \; \frac{1}{10} + 0.01 \right)$

$= (0.11, 0.09)$

11) In a study of an automobile insurence, a random sample of 80 body repair cars had a mean of ₹472.36 Rs. And SD is ₹62.35, if $\bar{x}$ is used as a point estimate to the true avg repair cars with that confidence we can assert

Sol) given, ~~$\bar{x} = 472.36$, $\sigma = 62.35$~~

that the maximum error doesn't exceed ₹10

Sol) given, $\bar{x} = 472.36$, $\sigma = 62.35$, $n = 80$

maximum error, $E = 10$

$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$10 = z_{\alpha/2} \frac{62.35}{\sqrt{80}}$

$10 = z_{\alpha/2}$

$z_{\alpha/2} = \frac{10 \times \sqrt{80}}{62.35} = 1.4338$

at $t_{\alpha/2} = 1.4338$, $\alpha_1 = 0.9236$

$$\alpha/2 = 1 - 0.9236$$
$$= 0.0764$$
$$\Rightarrow \alpha = 0.1528$$

Confidence $= (1-\alpha)100\% = (1-0.1528)100\%$

$= 84.72\%$

UNIT-6

TEST OF HYPOTHESIS

Def:-

In many circumstances, we arrive at decisions about the population on the basis of sample into, we make assumptions about the population parameters Such is called statistical hypothesis which may be true or not.

⇒ The procedure which enables us to decide on the basis of sample results whether a hypothesis is true or not is called as test of hypothesis / test of significance.

Ex:-

* A drug chemist is to decide whether a new drug really effective in curing a disease
* A quality control manager is to determine whether a process is working properly.
* A statistian is to decide whether a given coin is baised.

Working procedure for the test of hypothesis:-

Step-1 (Null hypothesis):

A definite statement about the population parameter

⇒ A Null hypothesis is the hypothesis which assert that there is no significant difference b/w the statistic & population parameter & whatever observed difference is there.

Null hypothesis $H_0 : \mu = \mu_0$

step 2 :- (Alternative hypothesis)

Any hypothesis which contradicts the null hypothesis is called alternative hypothesis.

$H_1 : \mu \neq \mu_0$ (two tailed test)

$H_1 : \mu < \mu_0$ (left tailed) } one tailed

$H_1 : \mu > \mu_0$ (Right tailed) }

Step-3: (Level of significance)

Select the appropriate level of significance α, depending on the reliability of & permissible ~~tist~~ risk.

Step-4:

test statistics :-

We have to ~~right~~ select the right test depending on the nature of information.

⇒ Then we construct the test criterian & select the appropriate probability distribution like z, t, f, χ^2 .... etc.

Step-5:

Making decision / conclusion for the problem

Comparing z & z_α :-

→ If $|z| < z_\alpha$ we accept the NULL hypothesis.

→ if $|z| > z_\alpha$ we reject the NULL hypothesis.
(i.e., alternative hypothesis accepted)

Errors of Sampling

Type-1 Error :-

Reject H_0 when it is true.

If the null hypothesis H_0 is true but it is rejected by the test procedure. then the error made is called type-1 error (or) α-error.

Type-2 Error :-

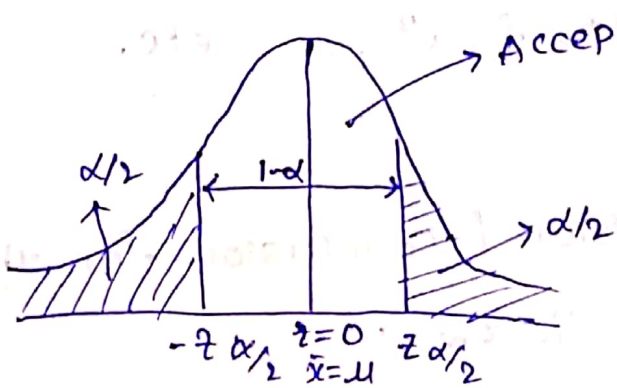
Accept H_0 when it is wrong.

If the Null hypothesis is false but it is accepted by the test then the error made is called type-2 error (or) β-error.

NOTE:

* Type-1 error is producers risk
* Type-2 error is consumers risk

Two tailed test at level of significance (α)

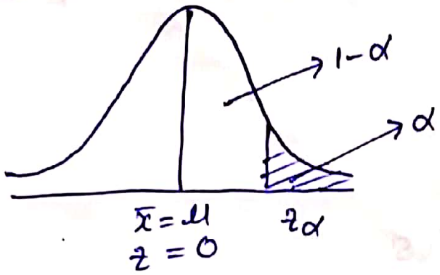


i.e., $P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1$.

In this case we rejected $\alpha/2$ area in right & $\alpha/2$ in left end

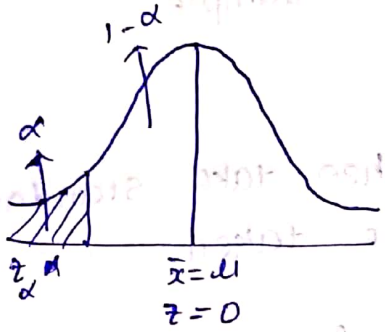
One tailed test:

Right tailed test:



$$P(z < z_\alpha) = 1-\alpha$$

left tailed test:



$$P(z > z_\alpha) = 1-\alpha$$

Critical Values

	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
two tailed test	$\lvert z_\alpha \rvert = 2.58$	$\lvert z_\alpha \rvert = 1.96$	$\lvert z_\alpha \rvert = 1.645$
Right tail-test	$z_\alpha = 2.33$	$z_\alpha = 1.645$	$z_\alpha = 1.28$
left tail test	$z_\alpha = -2.33$	$z_\alpha = -1.645$	$z_\alpha = -1.28$

Formulas for z-test

1. Test of significance for single mean :-

$$Z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

Z = Test statistics.
μ = mean of the population
σ = standard deviation
n = size of sample
$\bar{x}$ = mean of the sample

Note:
If σ is unknown then take standard deviation of sample (s) can be taken.

2. Test of significance for 2 means :-
(difference of means)

Test statistics $(Z) = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1^2} + \frac{\sigma_2^2}{n_2^2}}}$

$\bar{x_1}$ = mean of Sample 1
$\bar{x_2}$ = mean of sample 2
n_1, n_2 = Size of sample 1 & sample 2
μ_1 = mean of population 1
μ_2 = mean of population 2
σ_1, σ_2 = SD of population 1 & 2

(or) $Z = \frac{(\bar{x_1} - \bar{x_2}) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1^2} + \frac{\sigma_2^2}{n_2^2}}}$

where $\delta = \mu_1 - \mu_2$

NOTE:

If σ_1 & σ_2 are not known then best statistics becomes $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

3, Test of significance for single proportion :-

Test statistics, $z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}}$

where $Q = 1 - P$

P = population proportion

p = sample proportion

n = size of sample.

4, Test of significance for 2 proportions :-

Test statistics; $z = \dfrac{P_1 - P_2}{\sqrt{PQ\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

where $P = \dfrac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$

P_1, P_2 = sample 1 & 2 proportion

n_1, n_2 = size of sample 1 & 2

P_1, P_2 = population 1 & 2 porportion

1) A sample 400 items is taken from a population whose SD is 10. The mean of the sample is 40. Test whether the sample has come from the population with mean 38 and also calculate 95% confidence interval for the population.

Sol) Given that,

Sample size $(n) = 400$

SD of sample $(\sigma) = 10$

mean of sample $(\bar{x}) = 40$

mean of population $(\mu) = 38$

⇒ Null hypothesis (H_0):

$\mu = 38$

⇒ Alternative hypothesis (H_1): (Two tailed test)

$\mu \neq 38$

⇒ level of significance

$\alpha = 5\%$

$\Rightarrow \alpha = 0.05$

Now $z_{\alpha/2} = 1.96$

⇒ Test statistics:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{40 - 38}{\frac{10}{\sqrt{400}}}$$

$$= 2 \times \frac{20}{10}$$

$$= 4$$

$\Rightarrow |z| = |4| = 4$

⇒ conclusion:

Here $|z| > z_{\alpha/2}$

i.e, reject the null hypothesis.

That means, accepting the alternative hypothesis.

95% confidence interval for the population

$$= \left(\bar{x} - z\alpha/2 \frac{\sigma}{\sqrt{n}} , \bar{x} + z\alpha/2 \frac{\sigma}{\sqrt{n}} \right)$$

$$= \left(40 - 1.96 \frac{10}{\sqrt{400}} , 40 + (1.96) \frac{10}{20} \right)$$

$$= (40 - 0.98, 40 + 0.98$$

$$= (39.02, 40.98)$$

2) An ambulance service claims that it takes an avg less than 10 min to reach its destination in emergency calls. A sample of 36 cars has a mean of 11 min & variance is 16 min. Test the significance at 0.05 level.

Sol) Given that,

Sample size (n) = 36

SD (σ) = $\sqrt{16}$ = 4

mean of sample ($\bar{x}$) = 11

mean of population (μ) = 10

i) Null hypothesis (H_0) : $\mu = 10$

ii) Alternative hypothesis (H_1) : $\mu < 10$

iii) level of significance

$\alpha = 0.05$

$\Rightarrow z\alpha = 1.96 = -1.645 = 1.645$

iv) Test statistics

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{11 - 10}{\frac{4}{\sqrt{36}}}$$

$$= 1 \times \frac{6}{4}3$$

$$= 1.5$$

M Conclusion:

$|z| < z_\alpha$

i.e., Accepting the null hypothesis.

8) In a random sample of 60 workers, the avg tail time taken by them to get to work is 33.8 min with S.D is 6.1 min. Can we reject the null hypothesis $\mu = 32.6$ min in favour of alternative hypothesis $\mu > 32.6$ at $\alpha = 0.025$

Sol) Given,

$n = 60$

$\bar{x} = 33.8$

$\sigma = 6.1$

$\mu = 32.6$

i. null hypothesis (H_0) : $\mu = 32.6$

ii. Alternative hypothesis (H_1) : $\mu > 32.6$ (right tail)

iii. level of significance :

$\alpha = 0.025$

$\Rightarrow$ ~~$z_\alpha =$~~

$1-\alpha = 1-0.025 = 0.975$

$\Rightarrow z_\alpha = 1.96$

iv. Test statistics,

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

$$= \frac{33.8 - 32.6}{\frac{6.1}{\sqrt{60}}}$$

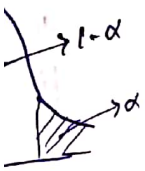
$$= 1.2 \times \frac{7.74}{6.1}$$

$$z = 1.523$$

Conclusion:

$|Z| = 1.523 < Z_\alpha = 1.96$.

i.e., accepting the null hypothesis.



5) A researcher wants to know the intelligence of students in a school. He selects 2 groups of students. In the 1st group, 150 students having mean IQ 75 with SD 15. In the 2nd group there are 250 students having mean IQ 70 with SD of 20.

Sol) Sample 1 size $(n_1) = 150$

$n_2 = 250$

$\bar{x}_1 = 75$

$\bar{x}_2 = 70$

$s_1 = 15$

$s_2 = 20$

Null hypothesis : $\mu_1 = \mu_2$

Alternative hypothesis : $\mu_1 \neq \mu_2$ (2 tail test)

level of significance :

$\alpha = 5\%$ (our assumption)

$\Rightarrow \alpha = 0.05$

$\Rightarrow \alpha/2 = 0.025$

$z_{\alpha/2} = 1.96$

Test statistics :

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{75 - 70}{\sqrt{\frac{15^2}{150} + \frac{(20)^2}{250}}}$$

$$= 2.8399$$

Conclusion:

Here $|Z| > Z_{\alpha/2}$

i.e., rejecting the null hypothesis.

Single Proportion:

1) In a sample of 1000 people in Karnataka, 540 rice eaters & the rest of people are wheat eaters. Can we assume that both rice & wheat are equally popular in the state at a 1% level of significance.

Sol) given that,

Sample size $(n) = 1000$

p = sample proportion $= \frac{540}{1000} = 0.54$

P = population proportion of rice eaters $= \frac{1}{2} = 0.5$

$$Q = 1 - P = 1 - \frac{1}{2} = \frac{1}{2} = 0.5$$

i) Null hypothesis : $P = \frac{1}{2}$

ii) Alternative hypothesis : $P \neq \frac{1}{2}$ 2-tail (2-tailed)

iii) level of significance

$$\alpha = 1\% = 0.01$$

$$\alpha/2 = 0.005$$

$$Z_{\alpha/2} = 2.58$$

iv) Test significance statistic:

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

$$= \frac{0.54 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{1000}}}$$

$$= 2.5316.$$

v) Conclusion:

$$|z| = 2.5316 < z_{\alpha/2} = 2.58$$

i.e., accepting the null hypothesis.

2) In a big city 325 men out of 600 men are found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers.?

Size sample, $n = 600$

$$p = \frac{325}{600} = 0.5$$

p = sample proportion of smokers $= \frac{1}{2} = 0.5$

$\therefore Q = 1 - P = 1 - \frac{1}{2} = 0.5$

i) Null hypothesis: $P = \frac{1}{2}$

ii) Alternative hypothesis: $P > \frac{1}{2}$ (right tailed)

iii) Test level of significance.

$$\alpha = 5\%$$

$$z_\alpha = 1.64$$

iv) Test statistis:

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0.5417 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{600}}}$$

$$= 2.041205$$

Conclusion:

$|z| = 2.04205 > z_\alpha = 1.64$

i.e., rejecting the null hypothesis.

i.e., accepting the alternative hypothesis.

∴ Majority men are smokers.

3) In a city-A, 20% of a random sample of 900 school boys have a certain physical defect. In another city-B, 18.5% of a random sample of 1600 school boys had the same defect. Is the diff b/w the proportion significant at 0.05 level of significance.

Sol) p_1 = sample proportion of city-A = $\frac{20\% \text{ of } 900}{900}$

$= \frac{\frac{20}{100} \times 900}{900} = 0.2$

p_2 = sample proportion of city-B

$= \frac{\frac{18.5}{100} \times 1600}{1600} = 0.185$

$n_1 = 900$, $n_2 = 1600$

$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(900)(0.2) + (1600)(0.185)}{900 + 1600}$

$= 0.1904$

$Q = 1 - P = 0.8096$

i) null hypothesis: (H_0): $p_1 = p_2$

ii) Alternative hypothesis (H_1): $p \neq p_2$

iii) level of significance,

$\alpha = 0.05$

$\alpha/2 = 0.025$

$z_{\alpha/2} = 1.96.$

iv. Test statistics

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.2 - 0.185}{\sqrt{(0.1904)(0.8096)\left(\frac{1}{900} + \frac{1}{1600}\right)}}$$

$$= 0.91743$$

v. Conclusion:

$|z| < z_{\alpha/2}$

i.e., accepting the null hypothesis.