25/08/23

(UNIT-2) **Correlation and Regression**

Topics

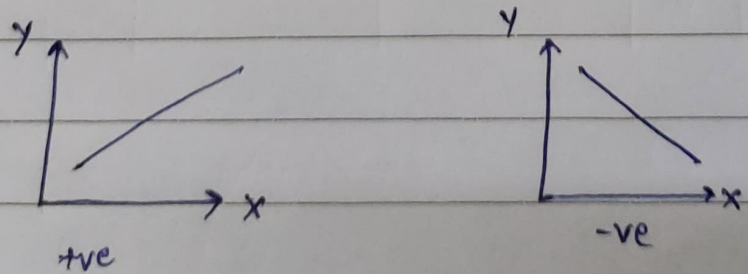
- If X and Y are two RV and when they are inter-related
- Scatter Plots.
- Correlation coefficient and properties.
- Karl pearson's.
- Spearman's Rank Correlation.
- Regression.
- Linear Regression and its Properties.

- X → Demand ↑ ↓
  Y → Supply ↑ ↓
  ↓
  +ve correlation

- X → Placement ↓
  Y → Admission ↑
  ↓
  -ve correlation

- X → attendance
  Y → CGPA
  ↓
  No correlation

* If the change in one variable effect the changes of another variable, then the variables are correlated.
* If the two variables deviate in the same direction, then the correlation is said to be direct or (+ve).
* If two RV defviate in the opposite direction then the correlation is said to be diverse or negative (-ve).



__/__/____

★ Scatter diagram or Plots.

↳ For the bivariate distribution (x_i, y_i) $(i=1, 2, \ldots n)$, if the values of x_i and y_i are plotted along x axis and y-axis in the xy plane, the diagram of dots so obtained is known as scatter diagram/plots.

a) ⇒ If the points are very (dense) i.e, very close to each other, then x and y are correlated. (fairly good data).

b) ⇒ If the points are widely scattered. either a poor correlation or no correlation is expected.

⇒ dense scatter
↓
(+ve) correlation

widely scatter
↙ ↘
(-ve) correlation No correlation.

★ Karl Pearson's Coefficient of Correlation.

Correlation coefficients b/w two RV x and y are denoted by $\gamma(x,y)$ or γ_{xy} which is defined as:-

★ $$\gamma(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$ ——— eq ①.

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

means; no of elements

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

$$\sigma_y^2 = E\{y - E(y)\}^2 = \frac{1}{n} \sum (y_i - \bar{y})^2$$

from eq ①

$$r(x,y) = r_{xy} = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2} \cdot \sqrt{\frac{1}{n}\sum y^2 - \bar{y}^2}}$$

$$= \frac{\sum xy - \bar{x}\bar{y}}{\sqrt{\sum x^2 - \bar{x}^2} \cdot \sqrt{\sum y^2 - \bar{y}^2}}$$

Q Calculate the correlation of height of fathers and their sons in inches.

Father's → X = 65, 66, 67, 67, 68, 69, 70, 72

Son's → Y = 67, 68, 65, 68, 72, 72, 69, 71

X	Y	$U = X - \bar{X}$ (67 → assumed mean) $= X - 68$	$V = Y - \bar{Y}$ (72 → assumed mean) $= Y - 69$	U^2	V^2	UV
65	67	= -3	= -2	9	4	6
66	68	= -2	= -1	4	1	2
67	65	= -1	= -4	1	16	4
67	68	= -1	= -1	1	1	1
68	72	0	= 3	0	9	0
69	72	1	= 3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
				36	44	24

we will assume any random value as a mean instead we will use assumed mean

we will not use it

$$\bar{x} = \frac{\sum x}{n} = \frac{544}{8} = 68.$$

$$\bar{y} = \frac{\sum y}{n} = \frac{552}{8} = 69.$$

$$Cov = \frac{24}{8} = 3$$

$$\sigma_x^2 = \frac{36}{8} = 4.5$$

$$\sigma_y^2 = \frac{44}{8} = 5.5$$

$$r_{xy} = \frac{3}{\frac{6}{\sqrt{8}} \times \sqrt{\frac{44}{8}}}$$

$r_{xy} = \frac{36/8}{6 \times 8 / 8_2 \times \sqrt{44}} \quad \frac{3^1}{\frac{6^2}{\sqrt{8}} \times \frac{\sqrt{44}}{\sqrt{8}}} = \frac{1}{2\sqrt{44}} = \frac{4}{\sqrt{44}}$

$r_{xy} = \frac{1}{2\sqrt{44}}$

$= \frac{4^2}{2\sqrt{11}} = 0.60$ Ans.

* also when elements are factor of 5, 10, 15, 25

$U = \frac{x-a}{h}, \quad V = \frac{y-b}{k}$

a, b → mean origin

h, k → scale

31/08/23

Q A computer while calculating correlation coefficient b/w two variables x and y from 25 pairs of observation, obtained the following results.

$n = 25, \quad \Sigma X = 125, \Sigma X^2 = 650, \Sigma Y = 100, \Sigma Y^2 = 460, \Sigma XY = 508,$

it was how, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values are :-

X	Y
8	12
6	8

obtain correct correlation coefficient.

→ $r_{(x,y)} = \frac{Cov(x,y)}{\sigma_x, \sigma_y} = \frac{\Sigma xy - \bar{x}\bar{y}}{\sqrt{\Sigma x^2 - \bar{x}^2}\sqrt{\Sigma y^2 - \bar{y}^2}}$ — ①.

$\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

$\Sigma Y = 100 + 12 + 8 - 14 - 6 = 100$

$\Sigma X^2 = 650 - 36 - 64 + 64 + 36 = 650$

$\Sigma Y^2 = 460 - 196 - 36 + 144 + 64 = 486$ (196 − 36 = 160; 144 + 64 = 208; 460 − 160 + 208 = 508 − 182 = 326)

$\Sigma xy = 508 - 96 - 48 + 84 + 48$

$= 496$

$= 508 - 84 + 48 + 96$

$= 520$

$\bar{x} = \frac{\Sigma X}{n} = \frac{125}{25} = 5 \quad \bar{y} = \frac{\Sigma Y}{n} = \frac{100}{25} = 4.$

$= \frac{1}{25}(520) - 25$

$\sqrt{\frac{1}{25}(650) - 25}\ \sqrt{\frac{1}{25}(436) - 20}$

$= \frac{520 - 25}{\sqrt{25}\sqrt{416}}$

$= \frac{495}{\sqrt{25}\sqrt{416}}$

$= 2$

$Cov(X,Y) = \frac{1}{25}(\Sigma XY) - \bar{X}\bar{Y}$

$= \frac{1}{25}(520) - 20$

$= 20.8 - 20$

$= 0.8$

$\sigma_x = \sqrt{\frac{1}{n}(\Sigma x^2) - \bar{X}^2}$

$= \sqrt{\frac{1}{25}(650) - 25}$

$= \sqrt{\frac{25}{25}} = 1$

$r_{(x,y)} = \frac{0.8 \times 5}{6} = \frac{4}{6} = \frac{2}{3} = 0.67$

$\sigma_y = \sqrt{\frac{1}{n}(436) - 16}$

$= \sqrt{\frac{436 - 400}{25}} = \sqrt{\frac{36}{25}} = \frac{6}{5}$

★ **Theorem:-**

The Variable X and Y are connected by the equation $aX + bY + c = 0$. Show that the correlation coeff b/w them is -1, if the signs of 'a' & 'b' are alike (same) and $+1$, if the sign of 'a' and 'b' are different.

⇒ Let $ax + by + c = 0$ — ①

Taking Expection on both side.

$aE(X) + bE(Y) + c = 0$ — ②.

__/__/__

① - ②

$Cov(X,Y) = ?$

$\sigma_X^2 = ?$

$$a(X - E(X)) + b(Y - E(Y)) + c - c = 0$$

__/__/____

★ Properties of Correlation :-

⇒ Correlation coefficient always lies b/w "-1 to +1;"
i.e $\boxed{-1 \le r \le 1}$

a) If $r = +1$, ⇀ then co-relation is perfect and positive (+ve)

b) If $r = -1$, ⇀ then co-relation is perfect and negative (-ve)

★★ ⇒ Co-relation coefficient is independent of change of origin and change of scale.

i.e $U = \frac{X-a}{h}$, $V = \frac{Y-b}{k}$, $a, b \rightarrow$ origin (mean assumed)

$k, h \rightarrow$ scale (factor)

⇒ Two independent variables are uncorrelated.

Proof :- If two rv's are independent, then

$$Cov(X,Y) = 0$$

$$\therefore r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{0}{\sigma_X \sigma_Y}$$

$$\boxed{r_{(X,Y)} = 0}$$

★ Rank Co-relation :-

→ Spearman's rank correlation Coefficient :-

⇒ The rank corr. is denoted by 'ρ'.

$$\boxed{\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}}$$

where $d_i = x_i - y_i$

{when rank is not repeated.}

1/09/23

Q The rank of same say, 16 student in maths and Physics are as:-
Two numbers with bracket, denote the rank of the student in math and physics.

{(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).}

calculate the rank corr coff for coffficency of this group for Maths. and Physics.

X	Y	$d_i = x - y$	d_i^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

$$\Rightarrow \rho = 1 - \left(\frac{6\sum d_i^2}{n(n^2-1)}\right)$$

$$= 1 - \left(\frac{6 \times 136}{16(255)}\right) \quad \frac{816}{64} \quad \frac{17}{4}$$

$$= 1 - \frac{1}{24.285} \qquad 36.42 \qquad 6.07$$

$$= 1 - 0.2$$

$$= 0.8$$

NOTE

→ Rank can be (+ve) also.

__/__/____

Q. 10 competitor in a musical test were ranked by three judges (A, B, C) in following order.

Rank by A : 1 6 5 10 3 2 4 9 7 8

Rank by B : 3 5 8 4 7 10 2 1 6 9

Rank by C : 6 4 9 8 1 2 3 10 5 7

using rank correlation matter, discuss which pair of judges has the nearest approach to common liking in music.

⇒

X	Y	Z	$d_1 = X - Y$	$d_2 = X - Z$	$d_3 = Y - Z$	d_1^2	d_2^2	d_3^2
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	+6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	-1	2	1	1	4
						200	60	214

$$f(X,Y) = 1 - \left(\frac{6\sum d_1^2}{10(100-1)}\right) = 1 - \frac{6 \times 200}{10 \times 99} = 1 - \frac{40}{33} = \frac{-7}{33}$$

$$f(X,Z) = 1 - \frac{6 \times 60}{10 \times 99} = 1 - \frac{36}{33} = \frac{7}{11}$$

$$f(Y,Z) = 1 - \frac{6 \times 214}{10 \times 99} = 1 - \frac{214}{165} = \frac{-49}{165}$$

$f(x, z)$ having the max. value, so

⇒ the judges A and C have common liking in music.

★ Repeated Rank :-

$$\rho = 1 - \frac{6\left[\sum d^2 + \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} + \dots\right]}{n(n^2-1)}$$

∴ where $m_1, m_2 \dots$ no. of times rank is repeated.

Q. Obtain the rank co-relation for the following data :-

where data are :-

⇒ X : 68, 64, 75, 50, 64, 80, 75, 40, 55, 64

Y : 62, 58, 68, 45, 81, 60, 68, 48, 50, 70

75 → 2, 3 → write these in series

75 ⇒ $\frac{2+3}{2} = 2.5$

64 ⇒ $\frac{5+6+7}{3} = \frac{18}{3} = 6$

68 ⇒ $\frac{3+4}{2} = 3.5$

→

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

→ 72

__/__/____

$$1 - \frac{6\left[\Sigma d^2 + \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} + \cdots\right]}{n(n^2-1)}$$

In X series

↳ 75 is repeated 2 times

↳ 64 is repeated 3 times.

$$\Rightarrow 1 - 6\left[72 + \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12}\right]$$

$$\frac{m_1(m_1^2-1)}{12} = \frac{2(4-1)}{12} = 5/12$$

$$\frac{m_2(m_2^2-1)}{12} = \frac{3(9-1)}{12} = \frac{24}{12}$$

$$= 5/12 + 24/12 = \frac{29}{12} = 2\,5/12$$

$$= 5/2$$

In Y-Series

↳ 68 is repeated 2 times.

$$= \frac{m_3(m_3^2-1)}{12}$$

$$= \frac{2(4-1)}{12} = 1\,5/12$$

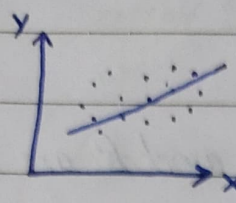
$$\Rightarrow \rho = 1 - \frac{6\left[72 + 5/2 + 1/2\right]}{10(99)} = 1 - \frac{6 \times 75}{10 \times 99} = 1 - \frac{15}{33}$$

$$= 18/33$$

$$= 0.545$$ Ans

4/09/23 __/__/____

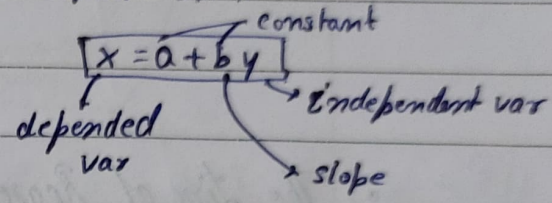
* Regression



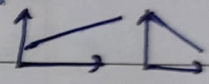
* -> Best fit data
* -> Give future value and estimated.

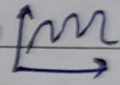
eshun → x depends on y

-> Regression analysis is a mathematical measure of the average relationship b/w two or more variables in terms of the original units of data.



* Types of Regression :-

1) Linear Regression 

2) Curvi - Linear Regression 

↳ any line except straight line.

* Linear Regression.

-> If the variables of bi-variate distribution are related, then the points in the scattered diagram will cluster round some curve called the curve's of Regression.

-> If the curve is a straight line it is called Line of regression and said to be Linear regression b/w the variables, otherwise it will be curvi-linear regression.

-> The line of regression is the line which gives the best estimate to the value of one variable for any specific value of other variable. Thus the line of regression is the line of "best fit."

__/__/____

- Let us suppose that in bi-variate distribution, such that Y is ~~infit~~ dependent variable and X is independent variable.

(x_i, y_i) ; $i = 1, 2, 3, \ldots, n$

Y on X is $\Rightarrow \boxed{Y = a + bx}$, where a and b are arbitrary const;

b → is the scope of line of Regression

- The line of Regression passes through the points $(\bar{x}, \bar{y})$ is

Y on X

$$\boxed{Y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})}$$

means; coefficient of co-relation; Slope

X on Y

$$\boxed{X - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})}$$

- Remark

★ ↳ In a particular case of perfect co-relation (r = +1 or -1) that is +ve or -ve i.e, $r = \pm 1$, then the line of regression for Y on X is :-

Y on X

$$\boxed{Y - \bar{y} = \pm\frac{\sigma_y}{\sigma_x}(x - \bar{x})}$$

★ { coincide in case of perfect correlations r = ±1 or -1

X on Y

$$\boxed{X - \bar{x} = \pm\frac{\sigma_x}{\sigma_y}(y - \bar{y})}$$

both Line coincide.

means=[const]

★ Since both lines of regression pass through the point $(\bar{x}, \bar{y})$ (mean), so they can't be parallel, Hence in this case of perfect correlation (+ve or -ve) the two Lines will coincide.

__/__/____

* Regression coeff.

⇒ "b" → it is the slope of the line of regression of Y on X is also called the coefficient of regression of Y on X.

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \text{Reg coff of } Y \text{ on } X$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \text{Reg coff of } X \text{ on } Y.$$

Q In a partially destroyed lab, record of an analysis of co-relation data. The following data are ~~for~~ only legible:-

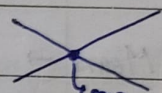
Var of $x = 9 \rightsquigarrow \sigma_x^2 = 9, \sigma_x = 3$.

Regression eq^n are $8x - 10y + 66 = 0$.
$40x - 18y = 214$

Find i) what are the mean values of X and Y.

ii) Correlation coff b/w x and y.

iii) σ_y.



since the line of Reg passes through a point $(\bar{x}, \bar{y})$ so, means are :-

$8x - 10y = -66 \times 5$
$40x - 18y = 214$
$\overline{\quad\quad\quad\quad}$
$32y = 544$
$y = 544/32$ (+36, 17, 8)

$\bar{y} = 17$

$8x - 10 \times 17 = -66$
$8x - 170 = -66$
$8x = 104$
$x = 13$

$\bar{x} = 13$

i) $(\bar{x}, \bar{y}) = (13, 17)$

ii) $8x - 10y + 66 = 0$ | $40x - 18y = 214$

Y on X | X on Y

$-10y = -66 - 8x$ | $40x = 214 + 18y$

$y = \frac{66}{10} + \frac{8}{10}x$ | $x = \frac{18}{40}y + \frac{214}{40}$

$y = \frac{8x}{10} + \frac{66}{10}$ | $b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{18}{40}$

$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{8}{10}$

$r^2 = b_{yx} \cdot b_{xy} = \frac{8}{10} \cdot \frac{18}{40}$

$r = \pm 0.6$

iii) $\sigma_y = b_{yx} = \frac{8}{10} = \pm 0.6 \cdot \frac{\sigma_y}{3}$

=

7/09/23

(iii) σ_y $\quad b_{yx} = 8/10$

$b_{xy} = 18/40$ $\quad r = \pm 0.6$

$\sigma_y =$ $\quad b_{yx} = r \frac{\sigma_y}{\sigma_x}$ (0.6, $\sigma_x \to 3$)

$8/10 \quad \frac{8}{10} \times \frac{3}{0.6} \times 0.6$

$\boxed{\sigma_y = 4}$ //

H.W

Q. Find the most likely price in mumbai corresponding to the price of ₹70 at kolkata from following :-

	Kolkata (X) (independent)	Mumbai (Y) (dependent)
Mean → $\bar{x}$ and $\bar{y}$	65	67
standard deviation	$\sigma_x = 2.5$	$\sigma_y = 3.5$

$Y = a + bx$

Co-relation Coeff (r) b/w the price of commodities in the two cities is 0.8.

→ $y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$ (0.8, 3.5 → 7)

$y \to 67$, $\sigma_x \to 2.5$, $\bar{x} \to 65$

$y - \bar{y} = 0.8 \times \frac{7}{5} (x - 65)$ ($x \to 70$)

$y - 67 = 0.8 \frac{7}{5} \times 5$

$y = 5.6 + 67$

$\boxed{y = 72.6}$ //

__/__/____

* Properties of Co-Regression.

1) The co-relation coefficient is the geometric mean b/w the regression co-efficients.

↳ $r = \sqrt{b_{xy} \cdot b_{yx}}$

a & b

gm → $\sqrt{ab}$

geometric gm → $\sqrt{ab}$

arithmetic am → $\frac{a+b}{2}$

$$b_{xy} \cdot b_{yx} = \left(r \frac{\sigma_y}{\sigma_x}\right) \cdot r\left(\frac{\sigma_x}{\sigma_y}\right)$$

$$= r^2$$

$$\sqrt{b_{xy} \cdot b_{yx}} = \sqrt{r^2} = r \quad \text{proved}$$

2) If one of the Regression co-efficient is greater than unity, other must be less than unity. → 1.

↳ $b_{yx} > 1$ → given

$$\Rightarrow \frac{1}{b_{yx}} < 1 \quad \text{—①}$$

r always lies → $-1 \leq r \leq 1 \Rightarrow r^2 \leq 1$

$$\Rightarrow b_{xy} \cdot b_{yx} \leq 1$$

$$b_{xy} \leq \frac{1}{b_{yx}}$$

from eq①

$$b_{xy} \leq \frac{1}{b_{yx}} < 1$$

↳ $b_{xy} \leq 1$

$b_{yx} > 1$ //.

3) The modulus value of the arithmetic mean of regression coefficient is not less than the modulus value of correlation coefficient.

$$\Rightarrow \left| \frac{b_{xy} + b_{yx}}{2} \right| > |r|$$

$$\Rightarrow \left| \frac{1}{2} \left(r \frac{\sigma_x}{\sigma_y} + r \frac{\sigma_y}{\sigma_x} \right) \right| > |r|$$

$$\left| \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \right| > 2$$

$$\left| \frac{\sigma_y^2 + \sigma_x^2}{\sigma_x \sigma_y} \right| > 2 \Rightarrow \sigma_y^2 + \sigma_x^2 > 2\sigma_x\sigma_y$$

$$\Rightarrow \sigma_y^2 + \sigma_x^2 - 2\sigma_x\sigma_y > 0$$

$$\Rightarrow \boxed{(\sigma_x - \sigma_y)^2 > 0}$$

which is true,
real quantity is always ≥ 0.

4) Regression coefficient are independent of change of origin but not of scale.

↳ Let $U = \frac{X-a}{h}$ & $V = \frac{Y-b}{k}$

$\Rightarrow X = a + hU$ & $Y = b + kV$

where $a, b, h > 0$ & $k > 0$ are const.

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \cdot \frac{\sigma_y}{\sigma_x}$$

$$= \frac{Cov(X,Y)}{\sigma_X^2}$$

$$= \frac{hk\, Cov(U,V)}{\sigma_U^2}$$

$$\boxed{b_{yx} = \frac{k}{h}\frac{Cov(U,V)}{\sigma_V^2}}$$

Similarly

$$\boxed{b_{xy} = \frac{h}{k}\frac{Cov(U,V)}{\sigma_V^2}}$$

$$\boxed{b_{yx} = \frac{k}{h}\, b_{UV}} \quad \text{or} \quad \boxed{b_{xy} = \frac{h}{k}\, b_{VU}}$$

5) Angle b/w two lines of regression is

$$\theta = \tan^{-1}\left\{\frac{1-r^2}{|r|}\left(\frac{\sigma_x\,\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right\}$$

* Case I → If $r=0$, $\tan\theta = \infty$

$$\Rightarrow \boxed{\theta = \pi/2}$$

Thus if two variables are un-correlated i.e $\boxed{r=0}$, the line of regression perpendicular to each other.

* Case II → $r = \pm 1$, $\tan\theta = 0$ $\Rightarrow \boxed{\theta = 0 \text{ or } \pi}$

In this case, the two lines of regression either coincide or they are parrallel to each other.

Remark → Since both line of regression passes through the point $(\bar{x}, \bar{y})$. So they can't be parallel. Hence, in the case of perfect correlation (+ve or −ve) the two lines of regression coincides.

__/__/____

* Proof of correlation coefficient property ②.

Proof :- Let $U = \frac{X-a}{h}$ & $V = \frac{Y-b}{k}$

$\Rightarrow X = a + Uh$ & $Y = b + kV$ —— ①

Taking Expection :-

↳ $E[X] = a + E(U).h$ & $E[Y] = b + k.E(V)$ —— ②

① - ②

$X - E[X] = h(U - E[U])$ & $Y - E[Y] = k(V - E[V])$

$Cov(X,Y) = (X - \underbrace{E[X]}_{\mu_X})(Y - \underbrace{E[Y]}_{\mu_Y})$

$= h(U - E(U)).k(V - E(V))$

$= hk.\{(U - E(U))(V - E(V))\}$

$$\boxed{Cov(X,Y) = hk.\,Cov(U,V)} \quad —— ①$$

$\sigma_X^2 = E\{[X - E(X)]^2\}$

$= E\{h.(U - E(U))\}^2$

$= h^2.(U - E(U))^2$

$= h^2.\sigma_U^2$

$$\boxed{\sigma_X^2 = h.\sigma_U} \quad — ②$$

$\sigma_Y^2 = E\{[Y - E(Y)]^2\}$

$= E\{k[V - E(V)]^2\}$

$= k^2 E\{[V - E(V)]^2\}$

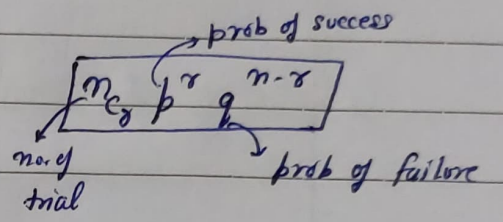
$\sigma_Y^2 = k^2 \sigma_V^2$

$$\boxed{\sigma_Y = k\sigma_V} \quad — ③$$

8/09/23

(UNIT-3) Some Discrete distributions

- The Bernoulli process.
- Binomial distribution and its Mgf [Moment Generating func^n] to find mean and Variance
- Negative Binomial distribution and its Mgf.
- Geometric distribution and its Mgf.
- Poisson distribution & its Mgf.



* Bernoulli process

- An Experiment often consist of repeated trials, each with two possible outcome that may be denoted as success or failure.

every outcome can be seen as success or failure.

- Remark -

1) The Experiment consist of repeated trial. → Ex:- for tossing 2 coin.
2) The probability of success is denoted by → "p", which remains constant.

Heads can be
$x = \{0, 1, 2\}$
↓
No. of successes.

3) The repeated trials are independent of each other.

4) The number of trials → "n" is always finite → for bernoulli
infinite → for poisson.

→ can be applied when 'n' is small.

* Binomial Distribution

A bernoulli trial can result in success with probability "p" and a failure with probability "q", i.e, $\boxed{q = 1-p}$, then the $\boxed{p+q=1}$

probability distribution of the binomial RV 'x', the no. of success in 'n' independent trial is—

$$\boxed{b(x; n, p) = {}^{n}C_{x}\, p^{x} \cdot q^{n-x}}$$

Q. Three items are selected at random from a manufacturing process, inspected and classified as defective or non-defective. Find the probability distribution for the number of defective items, ~~the~~ assuming that 25% of items are defective.

⇒ Let X :- defective. $n = 3$

~~25% of 3~~ $\frac{25}{100} \times 3 = \frac{75}{100}$

$X = \{0, 1, 2, 3\}$

$p = \frac{25}{100} = \frac{1}{4}$ probability of defective.

$b(x; n, p) = {}^{3}C_{x} \left(\frac{1}{4}\right)^{x} \cdot \left(\frac{3}{4}\right)^{3-x}$

$x = \{0, 1, 2, 3\}$

$x = 0 \quad = {}^{3}C_{0} \cdot \left(\frac{1}{4}\right)^{0} \cdot \left(\frac{3}{4}\right)^{3}$

$= 1 \cdot 1 \cdot \frac{27}{256}$

$x = 1 = {}^{3}C_{1} \cdot \left(\frac{1}{4}\right)^{1} \cdot \left(\frac{3}{4}\right)^{2}$

$x = 2 = {}^{3}C_{2} \left(\frac{1}{4}\right)^{2} \cdot \left(\frac{3}{4}\right)^{1}$

$x = 3 \quad = {}^{3}C_{3} \left(\frac{1}{4}\right)^{3} \cdot \left(\frac{3}{4}\right)^{0}$

31/11.21

$b = 1 \cdot \left(\frac{1}{4}\right)^{0} \cdot \frac{27}{256} + 3 \cdot \frac{1}{4} \cdot \left(\frac{3}{4}\right)^{2} + 3 \cdot \frac{1}{4}^{2} \cdot \frac{3}{4} + 1 \cdot \left(\frac{1}{4}\right)^{3}$

$= \frac{27}{256} + \frac{27}{64} + \frac{9}{64} + \frac{1}{64} = \frac{27 + 108 + 36 + 4}{256} = \frac{175}{256}$ ≠ 1

$\frac{27+27+9+1}{64}$

11/09/23 __/__/____

Q The prob that a patient recovers from a rare blood disease is 0.4. If 15 people are known to have contracted this disease. What is the prob of:-

(i) atleast 10 survive.

(ii) Exactly 5 survive.

(iii) from 3 to 8 patient are survive.

$n = 15$

Let X = Survive. $x = \{0, \ldots, 15\}$

prob = (0.4) = p of recovery.

$q = 0.6$

$$b(x; n, p) = {}^{n}C_x \cdot p^x \cdot q^{n-x} = {}^{15}C_x \cdot p^x \cdot q^{15-x}$$

(i) $P(X \geq 10) \Rightarrow x = 10, 11, 12, 13, 14, 15.$

$$b(x; 15, 0.4) = {}^{15}C_x \cdot p^x \cdot q^{n(15)-x}$$

$b_{10} \Rightarrow x = 10 = {}^{15}C_{10} \cdot p^{10} \cdot q^5$

$= \frac{15 \times 14 \times 13 \times 12 \times 11}{5 \times 4 \times 3 \times 2 \times 1}$

$= 3003 \times (4/10)^{10} \cdot (6/10)^5$

$x = 11 = {}^{15}C_{11} \cdot p^{11} \cdot q^4$

$= \frac{15 \times 14 \times 13 \times 12}{4 \times 3 \times 2 \times 1}$

$= 1365$

$x = 12 = {}^{15}C_{12} \cdot p^{12} \cdot q^3$

$= \frac{15 \times 14 \times 13}{3 \times 2 \times 1}$

$= 455$

$x = 13 \rightarrow {}^{15}C_{13} \cdot p^{13} \cdot q^2$

$= \frac{15 \times 14}{2 \times 1}$

$= 105$

$x = 14 \rightarrow {}^{15}C_{14} \cdot p^{14} \cdot q^1$

$= \frac{15}{1}$

$= 15$

$x = 15 \rightarrow {}^{15}C_{15} \cdot p^{15} \cdot q^0 \rightarrow 1$

__/__/____

★ Mean of binomial
↳ $\mu = np$ ★

★ Variance of binomial
↳ $Var = npq$ ★

★ (Moment Generating function) = MGF. → used to find mean and variance of any distribution

↳ The MGF of a RV "X" about origin having the probability function $f(x)$ is given by

$$M_X(t) = E[e^{tx}] = \begin{cases} \int e^{tx} f(x)\, dx, & \text{continuous} \\ \sum e^{tx} f(x), & \text{discrete.} \end{cases}$$

$E[x] = \sum x f(x)$

$$M_X(t) = E[e^{tx}] = E\left[1 + \frac{e^{tx}}{1!} + \frac{t^2 e^{tx}}{2!} + \dots + \frac{t^r e^{tx}}{r!} \dots\right]$$

$$= E\left[1 + \frac{tx}{1!} + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + \dots + \frac{t^r x^r}{r!} \dots\right] \quad e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots$$

$$= 1 + t\,E[X] + \frac{t^2 E[X^2]}{2!} + \dots + \frac{t^r}{r!} E[X^r] + \dots$$

$E[X^r]$ = moment of r^{th} term about origin

$$= 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \frac{t^3}{3!}\mu_3' + \dots \frac{t^r}{r!}\mu_r' + \dots$$

$$= \sum_{r=0}^{\infty} \frac{t^r}{r!} \cdot \mu_r'$$

$$\Rightarrow \boxed{M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu_r'}$$

or

$$\boxed{M_X^{\mu_r}(t) = \left| \frac{d^r}{dt^r} M_X(t) \right|_{t=0}}$$

UNIT 2- UNIT 3 → Bernoulli
Binomial
MGF mean and Variance.

__/__/____

Find mean and Variance of a binomial distribution by using MGF :-

Let 'X' be a binomial R.V. such that;

$$b(x; n, p) = {}^{n}C_x \, p^x \cdot q^{n-x} \quad ; \quad x = 0, 1, 2, \ldots n$$

$$M_x(t) = E[e^{tx}] = \sum e^{tx} \cdot {}^{n}C_x \cdot p^x \cdot q^{n-x}$$

$$= \sum {}^{n}C_x \cdot (pe^t)^x \cdot q^{n-x}$$

$$(1+x)^n = 1 + x\frac{n}{1!} + n\frac{(n-1)}{2!}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \ldots$$

$$(x+a)^n = {}^{n}C_0 a^n x^0 + {}^{n}C_1 a^{n-1} x^1 + \ldots$$

$$= \sum_{x=0}^{n} (pe^t + q)^n$$

Mean → $\mu = E[x^1] = \mu_1'$ $\qquad p+q=1$

$$\mu_r' = \left| \frac{d^r}{dt^r} M_x(t) \right|_{t=0} \Rightarrow = \left| \frac{d}{dt} (q+pe^t)^n \right|_{t=0}$$

$$\mu_1' = \left| \frac{d}{dt} (q+pe^t)^n \right|_{t=0} = \left| n(q+p)^{n-1} \cdot p \right|$$

$$= \left| n(q+pe^t)^{n-1} \cdot pe^t \right|_{t=0} = \left| n(1)^{n-1} \cdot p \right|$$

$$\boxed{E[x] = np}$$

$\mu_r' = E[X^r]$

$\downarrow$

$\frac{d^r}{dt^r} M_X(t)$

$$\text{Var} = E[X^2] - \underbrace{\left(E(X)\right)^2}_{\text{mean}}$$

$$E[X^2] = \mu_2'$$

$$= \left| \frac{d^2}{dt^2} M_X(t) \right|$$

$$= \left| \frac{d^2}{dt^2} (q + pe^t)^n \right|_{t=0}$$