19 December 2023 08:03 PM

Correlation and Regression

(X, Y)

Formulation : $\boxed{r_{X,Y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}}$

where $Cov(X,Y) = \frac{1}{n} \Sigma (x-\bar{x})(y-\bar{y})$

or $\left(= \frac{1}{n} \Sigma xy - \bar{x}\bar{y} \right)$

$$\begin{cases} \sigma_x = \sqrt{V(x)} = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2} \\ \sigma_y = \sqrt{V(y)} = \sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2} \end{cases}$$

(1) $-1 \le r \le 1$

(2) $r_{xy} = r_{u,v}$ ~~that~~ i's independent of any change in x and y.

(3) Regression lines → (a) Y on X

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

where b_{yx} = Regression coefficient

$$= \frac{cov(x,y)}{\sigma_x^2}$$

$$= \frac{r\,\cancel{\sigma_x}\,\sigma_y}{\sigma_x^2}$$

(b) Regression Equation of X on Y :

(a) Regression equation of X on Y :

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

where $\boxed{b_{xy} = r\frac{\sigma_x}{\sigma_y}}$

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

(5) Relation between r, b_{yx}, b_{xy}

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

or $$r = \sqrt{b_{yx} \times b_{xy}}$$

Numericals :

Q Height (U) and weight (V) of 5 persons are given

U :	1	4	3	5	5
V :	5	2	1	2	6

Determine the correlation coefficient b/w x and y.

$$r_{U,V} = \frac{Cov(U,V)}{\sqrt{\ldots}\sqrt{\ldots}} = \frac{\frac{1}{n}\sum UV - \bar{U}\bar{V}}{\ldots}$$

$$r_{u,v} = \frac{cov(u,v)}{\sqrt{V(u)}\sqrt{V(v)}} = \frac{\ldots}{\sigma_u \ \sigma_v}$$

U	V	U^2	V^2	UV
1	5	1	25	5
4	2	16	4	32
3	1	9	1	9
5	2	25	4	50
5	6	25	36	150
$\Sigma U = 18$	$\Sigma V = 16$	$\Sigma U^2 = 76$	$\Sigma V^2 = 70$	$\Sigma UV = 246$

$\bar{U} = \frac{1}{n} \Sigma u_i$

$= \frac{1}{5} \times 18 = 3.6$

$\bar{V} = \frac{1}{5} \times 16 = 3.2$

$$\sigma_v = \sqrt{\frac{1}{5} \times 70 - (3.2)^2}$$

$$= \sqrt{3.76}$$

$$= 1.939$$

$$\sigma_u = \sqrt{\frac{1}{n} \Sigma u^2 - \bar{u}^2}$$

$$= \sqrt{\frac{76}{5} - (3.6)^2}$$

$$= \sqrt{15.2 - 12.96}$$

$$= \sqrt{2.24} = 1.496$$

$$cov = \frac{1}{5} \times \overset{49.2}{\cancel{246}} - (3.6)(3.2)$$

$$= 49.2 - 11.52$$

$$= 37.68$$

$$\therefore r = \frac{cov}{\sigma \ \sigma} = \frac{37.68}{\ldots}$$

$$r = \frac{cov}{\sigma_x \sigma_y} = \frac{37.68}{1.939 \times 1.496}$$

$$= ans.$$

Q2 Using the following result on record of age (x) and blood pressure (y) of group 10 women

we have

	x	y
mean	53	142
variance	130	165

$$\Sigma(x-\bar{x})(y-\bar{y}) = 1220.$$

Find the regression equation of y on x and use it to estimate the blood pressure of a woman of age 45.

Regression equation of y on x

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

Given $\bar{x} = 53$ $\quad \sigma_x^2 = 130$

$\bar{y} = 142$ $\quad \sigma_x = \sqrt{130} = 11.401$

$\sigma_y^2 = 165$

$\sigma_y = \sqrt{165} = 12.845$

$$b_{yx} = \frac{cov(xy)}{} = \frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})$$

$\sigma_y = \sqrt{165} = 12.845$

$$b_{yx} = \frac{cov(x,y)}{\sigma_x^2} = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sigma_x^2}$$

$$= \frac{\frac{1220}{10}}{130}$$

$$= \frac{\cancel{1220}}{\cancel{10} \times 130} = 0.94$$

$\therefore$ Regression Equ of y on x is

$$y = 142 + 0.94(x - 53)$$

$$y = 142 + 0.94x - 49.82$$

$$\boxed{y = \underbrace{0.9}_{b_{yx}}4x + 92.18}$$

$x = 45$ & $y = ?$ $\quad y = (0.94 \times 45) + 92.18$

$= \quad$ ans.

also find x on y

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

$$x = 53 + b_{xy}(y - 142)$$

$$b_{xy} = cov(x,y)/\sigma_y^2 = \frac{1220/10}{}$$

$$b_{xy} = \mathrm{cov}(x,y)/\sigma_y^2 = \frac{1220/10}{165}$$

$$= \frac{122}{165}$$

$$= 0.739.$$

$$x = 53 + 0.739(y - 142)$$

$$x = 53 + 0.739y - 104.938$$

$$\boxed{x = 0.739y - 51.938}$$

Q if $b_{yx} = -3/2$, $b_{xy} = -1/5$

find $v(y)$ and $v(x)$

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

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \text{ans}$$

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

$$r = \sqrt{\frac{-3}{2} \times \left(-\frac{1}{5}\right)} = \sqrt{3/10} = 0.547.$$

Binomial Distribution

Note: r.v x — discrete

continuous

Note: r.v (x) — discrete / continuous.

discrete r.v x → probab distribn → probab mass funct (pmf)

Cont r.v x → " " → probab density fn (pdf).

① If x is a discrete r.v with n no. of trial and p is the probab of success, then the pmf of x is

$$f(x) = {}^{n}C_{x}\, p^{x} q^{n-x}$$

when $x = 0, 1, \ldots, n$

$$f(x) = 0 \quad \text{otherwise.}$$

This is Binomial Distribution.

① mean of Binomial Distribution

$$E(x) = \Sigma x p = n.p$$

② variance of B.D $V(x) = npq$
$= np(1-p)$

and S.D $= \sqrt{V(x)} = \sqrt{npq}$

Q: If a random variable X follows Binomial Distribution with

mean $= 2$

$E(x^2) = 28/5$

find $n = ?$

and ~~probabilit~~ $P(x=0)$

Given: mean $= E(x) = 2$

$E(x^2) = 28/5$

$Var(x) = E(x^2) - E(x)^2$

$Var(x) + E(x)^2 = E(x^2)$

$npq + (np)^2 = \frac{28}{5}$

$npq + 4 = 28/5$

$npq = \frac{28}{5} - 4$

$2q = \frac{28-20}{5}$

$\boxed{q = \frac{8}{5 \times 2} = 4/5}$

$\therefore \boxed{p = 1 - q = 1 - \frac{4}{5} = \frac{1}{5}}$

$E(x) = 2$

$np = 2$

$n \times \frac{1}{5} = 2$

$\boxed{n = 10}$

$\therefore P(x=0) = {}^nC_x \, p^x q^{n-x}$ ... $\sqrt{npq}$ ...

$$\therefore P(x=0) = {}^nC_x\, p^x\, q^{n-x}$$
$$= {}^{10}C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{10-0}$$
$$= 1 \cdot 1 \cdot \left(4/5\right)^{10}$$
$$= (0.8)^{10} \quad \text{(ans)}$$

$${}^nC_x = \frac{n!}{x!(n-x)!}$$

$${}^5C_3 = \frac{5!}{3!\,2!} = \frac{5\times4\times3\times2\times1}{3\times2\times1\times2\times1}$$

Q2 The probability of a patient recovering from a certain disease is 0.75. x
What is the probability of ~~getting~~ 3 $=x$ patients getting recovered out of 4. n
Also find the mean and variance.

$p = 0.75 \qquad n = 4$

$q = 0.25$

$$P(x=3) = {}^nC_x\, p^x\, q^{n-x}$$
$$= {}^4C_3\,(0.75)^3\,(0.25)^1$$
$$= \frac{4!}{\ldots}\,(0.75)^3\,(0.25)^1$$

$$= \frac{4!}{3!\,1!}(0.75)^{-}(0.25)$$

$$= \frac{\cancel{4}\times\cancel{3!}}{\cancel{3!}}\left(\frac{3}{4}\right)^3\cancel{\left(\frac{1}{4}\right)^1}$$

$$= \frac{27}{64} = 0.421 \text{ (ans)}.$$

Mean, $E(x) = np$
$= \cancel{4}\times\frac{3}{\cancel{4}} = 3$ (ans)

variance $= npq$
$= \cancel{4}\times\frac{3}{\cancel{4}}\times\frac{1}{4} = \frac{3}{4}$ (ans)

$s.d = \sqrt{v} = \frac{\sqrt{3}}{2}$ (ans)

— * —

(1) central Tendency
(1) mean (with & without frequency)
(2) median & mode (with frequency)
(3) Relation mean, median mode. (empirical)
(4) missing frequency

(5) Combined mean

$$\bar{X} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

(6) Quartiles

(Q_1, Q_2, Q_3 and Q_4)

(Numericals)

Dispersion: First 2 classes in Stat

follow p.d.f.

Time Series: Follow p.d.f

4:30 pm - 5:30 pm