Module - 2

Regression and Correlation

* Introduction to Correlation :-

In a bivariate distribution we may be interested to find out if there is any correlation or covariation between the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated.

① If the two variables deviate in the same direction i.e. if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be direct or positive.

② But if they constantly deviate in the opposite directions i.e. if increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be diverse or negative.

For e.g. ① the correlation between height and weights of a group of persons is positive correlation.

② The correlation between price and demand of commodity is negative correlation.

* Karl pearson's coefficient of correlation:-

As a measure of intensity or degree of linear relationships between two variables, Karl pearson, a British Biometrician, developed a formula called correlation coefficient.

Correlation coefficient between two random variables x and y, usually denoted by $r(x,y)$ or simply r_{xy} is a numerical measure of linear relationship between them and is defined as-

$$r(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\Sigma x^2 - \bar{x}^2\right)\left(\frac{1}{n}\Sigma y^2 - \bar{y}^2\right)}}$$

Where,

1. n denotes total number of observations
2. $\bar{x}$ is the mean of variable x given by $\bar{x} = \frac{1}{n}\Sigma x$
3. $\bar{y}$ is the mean of variable y given by $\bar{y} = \frac{1}{n}\Sigma y$

Que] Calculate the Karl Pearson's coefficient of correlation for the following heights (in inches) of Fathers (X) and their sons (Y).

Father (X)	65	66	67	67	68	69	70	72
Son (Y)	67	68	65	68	72	72	69	71

Answer:-

We know that,

$$r(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\Sigma x^2 - \bar{x}^2\right)\left(\frac{1}{n}\Sigma y^2 - \bar{y}^2\right)}} \rightarrow (1)$$

Here n = Number of observations = 8

The calculation of Karl Pearson's coefficient of skewness is given as-

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
$\Sigma X = 544$	$\Sigma Y = 552$	$\Sigma X^2 = 37028$	$\Sigma Y^2 = 38132$	$\Sigma XY = 37560$

$\therefore$ Here, $n = 8$

$\Sigma x = 544$, $\Sigma y = 552$, $\Sigma x^2 = 37028$, $\Sigma y^2 = 38132$,

$\Sigma xy = 37560$.

$$\therefore \bar{x} = \frac{1}{n}\Sigma x = \frac{1}{8} \times 544 = 68$$

$$\bar{y} = \frac{1}{n}\Sigma y = \frac{1}{8} \times 552 = 69$$

$\therefore$ ① $\Rightarrow$

$$r(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\Sigma x^2 - \bar{x}^2\right)\left(\frac{1}{n}\Sigma y^2 - \bar{y}^2\right)}}$$

$$= \frac{\frac{1}{8} \times 37560 - (68 \times 69)}{\sqrt{\left[\frac{1}{8} \times 37028 - (68)^2\right]\left[\frac{1}{8} \times 38132 - (69)^2\right]}}$$

$$= \frac{4695 - 4692}{\sqrt{(4628.5 - 4624)(4766.5 - 4761)}}$$

$$= \frac{3}{\sqrt{4.5 \times 5.5}}$$

$$= \frac{3}{\sqrt{24.75}}$$

$$= \frac{3}{4.97}$$

$$= 0.6036$$

$$\therefore \boxed{r(x,y) = 0.6036}$$

Que] calculate karl pearson's coefficient of correlation from following data.

X	1	3	4	5	7	8	10
Y	2	6	8	10	14	16	20

Answer:-

We know that,

$$r(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\Sigma x^2 - \bar{x}^2\right)\left(\frac{1}{n}\Sigma y^2 - \bar{y}^2\right)}} \rightarrow (1)$$

Here,

n = No. of observations = 7

The calculation of karl pearson's coefficient of correlation is given in following table.

X	Y	X^2	Y^2	XY
1	2	1	4	2
3	6	9	36	18
4	8	16	64	32
5	10	25	100	50
7	14	49	196	98
8	16	64	256	128
10	20	100	400	200
$\Sigma X = 38$	$\Sigma Y = 76$	$\Sigma X^2 = 264$	$\Sigma Y^2 = 1056$	$\Sigma XY = 528$

Here, $n = 7$

$\Sigma x = 38$, $\Sigma y = 76$, $\Sigma x^2 = 264$, $\Sigma y^2 = 1056$,

$\Sigma xy = 528$

$\therefore \bar{x} = \frac{1}{n}\Sigma x = \frac{1}{7} \times 38 = 5.43$

$\bar{y} = \frac{1}{n}\Sigma y = \frac{1}{7} \times 76 = 10.86$

① $\Rightarrow$

$$r(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\sqrt{\left[\frac{1}{n}\Sigma x^2 - \bar{x}^2\right]\left[\frac{1}{n}\Sigma y^2 - \bar{y}^2\right]}}$$

$$\therefore r(x,y) = \frac{\frac{1}{7} \times 528 - (5.43 \times 10.86)}{\sqrt{\left[\frac{1}{7} \times 264 - (5.43)^2\right]\left[\frac{1}{7} \times 1056 - (10.86)^2\right]}}$$

$$\therefore r(x,y) = \frac{75.43 - 58.97}{\sqrt{[37.71 - 29.48][150.85 - 117.94]}}$$

$$\therefore r(x,y) = \frac{16.49}{8.23 \times 32.91} = \frac{16.49}{270.85}$$

$$\therefore \boxed{r(x,y) = 0.0608}$$

Que] Find the Karl pearson's coefficient of correlation from following data.

Ans:

X	10	15	12	17	13	16	24	14	22	20
Y	30	42	45	46	33	34	40	35	39	38

Q.11e] A computer while calculating correlation coefficient between two variables x and y from 25 pairs of observations obtained the following results.

$n = 25$, $\Sigma x = 125$, $\Sigma x^2 = 650$, $\Sigma y = 100$, $\Sigma y^2 = 460$, $\Sigma xy = 508$.

It was, however, 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 where

x	y
8	12
6	8

obtain the correct value of correlation coefficient.

Answer:-

~~corrected Σx~~

It is given that,

$n = 25$, $\Sigma x = 125$, $\Sigma x^2 = 650$, $\Sigma y = 100$, $\Sigma y^2 = 460$, $\Sigma xy = 508$

corrected $\Sigma x = 125 - 6 - 8 + 8 + 6 = 125 - 14 + 14 = 125$

corrected $\Sigma y = 100 - 14 - 6 + 12 + 8 = 100 - 20 + 20 = 100$

corrected $\Sigma x^2 = 650 - 6^2 + 8^2 + 6^2 + 8^2 = 650$

corrected $\Sigma y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2$

$= 460 - 196 - 36 + 144 + 64$

$= 436$.

corrected $\Sigma xy = 508 - (6 \times 14) - (8 \times 6) + (8 \times 12) + (6 \times 8)$

$= 508 - 84 - 48 + 96 + 48$

$= 520$

$\therefore \quad \bar{x} = \frac{1}{n}\Sigma x = \frac{1}{25} \times 125 = 5$

$\bar{y} = \frac{1}{n}\Sigma y = \frac{1}{25} \times 100 = 4$

$\therefore \quad cov(x,y) = \frac{1}{n}\Sigma xy - \bar{x}\bar{y}$

$= \frac{1}{25} \times 520 - (5 \times 4)$

$= 20.8 - 20$

$\therefore \quad \boxed{cov(x,y) = 0.8}$

$\sigma_x = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2} = \sqrt{\frac{1}{25} \times 650 - (5)^2}$

$= \sqrt{26 - 25} = \sqrt{1}$

$\therefore \quad \boxed{\sigma_x = 1}$

$\sigma_y = \sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2} = \sqrt{\frac{1}{25} \times 436 - (4)^2}$

$= \sqrt{17.44 - 16} = \sqrt{1.44}$

$\therefore \quad \boxed{\sigma_y = 1.2}$

$\therefore \quad \text{corrected } r(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$

$= \frac{0.8}{1 \times 1.2} = \frac{0.8}{1.2} = 0.67$

$\therefore \quad \boxed{\text{corrected } r(x,y) = 0.67}$

* Rank correlation:-

Let us suppose that a group of "n" individuals is arranged in order of merit or proficiency in possession of two characteristics A and B. These rank in the two characteristics will, in general, be different.

Let (x_i, y_i) for $i = 1, 2, 3, \ldots n$ be the ranks of the ith individual in two characteristics A and B respectively. These correlation between the ranks x_i and y_i is called the rank correlation coefficient between A and B for that group of individuals.

* Spearman's Rank Correlation Coefficient:-

Assuming that no two individuals are bracketed equal in either classification, each of the variables X and Y takes the values 1, 2, 3, ---- n.

The Spearman's Rank correlation coefficient is given by-

$$\boxed{\rho = 1 - \left[\frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}\right]}$$

where $d_i = X_i - Y_i$

n = Number of observations.

Que] The ranks of same 16 students in mathematics and physics are as follows. Two numbers within brackets denote the ranks of the students in mathematics 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 correlation coefficient for proficiencies of this group in mathematics and physics.

Answer:- The calculation for rank correlation coefficient is given as -

We know that,

$$\rho = 1 - \left[\frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}\right] \rightarrow (1)$$

Rank in maths (X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		
Rank in physics (Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13		
$d = X-Y$	0	-8	0	0	0	-1	5	2	1	-1	-4	3	-1	2	-1	3		$\Sigma d = 0$
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9		$\Sigma d^2 = 136$

Here $n = 16$, $\Sigma d^2 = 136$

$$(1) \Rightarrow \rho = 1 - \left[\frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}\right] = 1 - \left[\frac{6 \times 136}{16(16^2-1)}\right]$$

$$= 1 - \left[\frac{6 \times 136}{16(256-1)}\right] = 1 - \left[\frac{6 \times 136}{16 \times 255}\right] = 1 - \frac{816}{4080}$$

$$= 1 - 0.2 = 0.8$$

$$\therefore \boxed{\rho = 0.8}$$

Que] Ten competitors in a musical test were ranked by three judges A, B and C in the 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 method, discuss which pair of judges has the nearest approach to common linking in music.

Answer:- We know that,

By Spearman's rank correlation coefficient

$$\rho = 1 - \left[\frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)} \right]$$

Here $n = 10$

Rank by A (X)	Rank by B (Y)	Rank by C (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
			$\Sigma d_1 = 0$	$\Sigma d_2 = 0$	$\Sigma d_3 = 0$	$\Sigma d_1^2 = 200$	$\Sigma d_2^2 = 60$	$\Sigma d_3^2 = 214$

$\therefore$ Here, $n=10$, $\Sigma d_1^2 = 200$, $\Sigma d_2^2 = 60$, $\Sigma d_3^2 = 214$

$\therefore$ ① $e(X,Y) = 1 - \left[\frac{6\Sigma d_1^2}{n(n^2-1)}\right] = 1 - \left[\frac{6\times 200}{10(10^2-1)}\right]$

$e(X,Y) = 1 - \left[\frac{6\times 200}{10(100-1)}\right] = 1 - \left[\frac{6\times 200}{10\times 99}\right]$

$e(X,Y) = 1 - \left[\frac{1200}{990}\right] = 1 - 1.21 = -0.21$

② $e(X,Z) = 1 - \left[\frac{6\Sigma d_2^2}{n(n^2-1)}\right] = 1 - \left[\frac{6\times 60}{10(10^2-1)}\right]$

$\therefore e(X,Z) = 1 - \left[\frac{6\times 60}{10(100-1)}\right] = 1 - \left[\frac{6\times 60}{10\times 99}\right]$

$\therefore e(X,Z) = 1 - \left[\frac{360}{990}\right] = 1 - 0.36 = 0.64$

③ $e(Y,Z) = 1 - \left[\frac{6\Sigma d_3^2}{n(n^2-1)}\right] = 1 - \left[\frac{6\times 214}{10(10^2-1)}\right]$

$\therefore e(Y,Z) = 1 - \left[\frac{6\times 214}{10(100-1)}\right] = 1 - \left[\frac{6\times 214}{10\times 99}\right]$

$\therefore e(Y,Z) = 1 - \left[\frac{1284}{990}\right] = 1 - 1.30 = -0.30$

Since $e(X,Z)$ is maximum.
We conclude that the pair of judges A and C has the nearest approach to common linking in music.

Que] Ten competitors in a beauty contest
H.W. are ranked by three judges as follows –

Judges	1	2	3	4	5	6	7	8	9	10
A	6	5	3	10	2	4	9	7	8	1
B	5	8	4	7	10	2	1	6	9	3
C	4	9	8	1	2	3	10	5	7	6

using rank correlation method, discuss which pair of judges has the nearest approach to common tastes of beauty.

Que] A sample of 12 fathers and their eldest
H.W. sons gave the following data about their height in inches.

Father	65	63	67	64	68	62	70	66	68	67	69	71
Son	68	66	68	65	69	66	68	65	71	67	68	70

calculate coefficient of rank correlation.