Correlation

Lesson 1: Karl Pearson's Correlation Coefficient

In a bivariate distribution, we are interested to find out if there is a degree of association between the variables under study. If the change in one variable affects a change in the other variable, they are said to be correlated. If the increase (or decrease) in one variable results in a corresponding increase (or decrease) in the other, correlation is said to be direct or positive. But if the increase (or decrease) in one variable results in a decrease (or increase) in the other, then correlation is said to be diverse or negative.

Eg. The correlation between height and weight of people is positive; the correlation between price and demand of an item is negative

Correlation is said to be perfect if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

Correlation can be linear and non-linear.

Karl Pearson's Correlation Coefficient:

This method is known as the covariance method. Let us consider that a bivariate distribution $x_i|y_i$ $(i = 1, 2, \ldots, n)$ is given. Then the correlation coefficient between the two variables x and y is defined as

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

where, $\text{cov}\,(\text{x}, \text{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}$ is the covariance between the two variables. Here $\bar{x}$ and $\bar{y}$ are the arithmetic means of the two distributions, σ_x and σ_y are the standard deviations of the two distributions.

Note:

(i) This is a numerical measure of linear relationship between the two variables x and y

(ii) $-1 \leq r \leq 1$

(iii) If $r = 1$, then the correlation is positive perfect and if $r = -1$, then the correlation is negative perfect

(iv) Correlation coefficient is independent of change of origin and scale

(v) Two independent variables are uncorrelated [*but the converse is not true]

(vi) For large values of the observations, we may substitute

$$u = \frac{x - a}{b}, v = \frac{y - c}{d}$$

Here a, b, c, d are constants. Then the correlation coefficient becomes

$$r = \frac{\frac{\sum_{i=1}^{n} u_i v_i}{n} - \bar{u}\bar{v}}{\sigma_u \sigma_v}$$

(vii) If $\bar{x}$ and $\bar{y}$ are whole numbers, then we can substitute

$$u = x - \bar{x} \text{ and } v = y - \bar{y}$$

In that case, the correlation coefficient becomes

$$r = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} {u_i}^2} \times \sqrt{\sum_{i=1}^{n} {v_i}^2}}$$

Problems:

Ex.1. Find correlation coefficient:

x	1	2	3	4	5
y	3	2	5	4	6

Solution: Let us make the table for calculation of correlation coefficient:

						Total
x	1	2	3	4	5	15
y	3	2	5	4	6	20
x^2	1	4	9	16	25	55
y^2	9	4	25	16	36	90
xy	3	4	15	16	30	68

The means of the two distributions are given by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{15}{5} = 3 \text{ and } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{20}{5} = 4$$

Then

$$\text{cov (x, y)} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} = \frac{68}{5} - 12 = 1.6$$

Also

$$\sigma_x{}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i{}^2 - \bar{x}^2 = \frac{55}{5} - 9 = 2 \rightarrow \sigma_x = \sqrt{2}$$

Similarly,

$$\sigma_y{}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i{}^2 - \bar{y}^2 = \frac{90}{5} - 16 = 2 \rightarrow \sigma_y = \sqrt{2}$$

Then the correlation coefficient is given by

$$r = \frac{\text{cov (x, y)}}{\sigma_x \sigma_y} = \frac{1.6}{2} = 0.8$$

Ans.

Ex.2. Find the correlation coefficient of the following data:

x	3	5	7	8	9	15	16
y	15	18	22	24	19	25	31

Solution: Let us make the table for calculation of correlation coefficient:

								Total
x	3	5	7	8	9	15	16	63
y	15	18	22	24	19	25	31	154
$u = x - 9$	-6	-4	-2	-1	0	6	7	0
$v = y - 22$	-7	-4	0	2	-3	3	9	0
u^2	36	16	4	1	0	36	49	142
v^2	49	16	0	4	9	9	81	168
uv	42	16	0	-2	0	18	63	137

The means of the two distributions are given by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{63}{7} = 9 \text{ and } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{154}{7} = 22$$

Since the two A.M's are whole number, let us substitute

$$u = x - \bar{x} \text{ and } v = y - \bar{y}$$

Then the correlation coefficient becomes

$$r = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} {u_i}^2} \times \sqrt{\sum_{i=1}^{n} {v_i}^2}} = \frac{137}{\sqrt{142} \times \sqrt{168}} = 0.887$$

Ans.

Ex.3. In a 12 pairs of observations, $\sum x = 30$, $\sum y = 5$, $\sum x^2 = 670$, $\sum y^2 = 285$ and $\sum xy = 334$. Later, it was found that a pair (10, 14) had been wrongly taken as (11, 4). Find the correct correlation coefficient.

Solution: Since the pair (10, 14) was wrongly taken instead of the correct pair (11, 4), we need to rectify all the summations. They are given by

$$\sum x_{|\text{corrected}} = 30 - 11 + 10 = 29$$

$$\sum y_{|\text{corrected}} = 5 - 4 + 14 = 15$$

$$\sum x^2{}_{|\text{corrected}} = 670 - 11^2 + 10^2 = 649$$

$$\sum y^2{}_{|\text{corrected}} = 285 - 4^2 + 14^2 = 465$$

$$\sum xy_{|\text{corrected}} = 334 - (11 \times 4) + (14 \times 10) = 430$$

Hence the corrected means are $\bar{x} = \frac{29}{12} = 2.416$ and $\bar{y} = \frac{15}{12} = 1.25$

Corrected covariance is $\text{cov}(x, y) = \frac{430}{12} - (2.416 \times 1.25) = 35.83 - 3.02 = 32.81$. Corrected standard deviations will be given by

$${\sigma_x}^2 = \frac{649}{12} - 2.416^2 = 54.08 - 5.84 = 48.24 \rightarrow \sigma_x = 6.95$$

and

$$\sigma_y{}^2 = \frac{465}{12} - 1.25^2 = 38.75 - 1.5625 = 37.19 \rightarrow \sigma_y = 6.098$$

Then the correlation coefficient will be given by

$$r = \frac{\text{cov (x, y)}}{\sigma_x \sigma_y} = \frac{32.81}{6.95 \times 6.098} = 0.774$$

Ans.

Ex.4. Comment on the relationship between x and y by calculating the correlation coefficient:

x	-4	-3	-2	-1	0	1	2	3	4
y	16	9	4	1	0	1	4	9	16

Solution: Let us make the table for calculation of correlation coefficient:

										Total
x	-4	-3	-2	-1	0	1	2	3	4	0
y	16	9	4	1	0	1	4	9	16	60
xy	-64	-27	-8	-1	0	1	8	27	64	0

Here we can see that $\text{cov (x, y)} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} = 0$, which means that Karl Pearson's correlation coefficient $r = 0$. This statement implies that there exists no linear relationship between the two variables x and y.

However, the two variables x and y are related through a non-linear (quadratic) relation given by $y = x^2$.

Ans.