

# CORRELATION

## (CONCEPT)

### METHODS OF STUDYING CORRELATION

There are two methods which visualize the relationship between the two variables i.e.

- Scatter Diagram Method and Graphic Method :These are based on graphs and diagrams
- Karl Pearson's Coefficient of Correlation, Rank Correlation Method, Concurrent Deviation Method, Method of Least Squares:These are based on mathematical methods.

### SCATTER DIAGRAM:

This is a simple diagrammatic method to establish correlation between a pair of variables. Scatter diagram can be applied for any type of correlation – linear as well as non-linear i.e. curvilinear.

Scatter diagram can distinguish between different types of correlation although it fails to measure the extent of relationship between the variables.

Each data point which in this case a pair of values  $(x_i, y_i)$  is represented by a point in the rectangular axes of coordinates. The totality of all plotted points forms the scatter diagram.

The platform of the plotted points reveals the nature of correlation.

In case of a positive correlation, the plotted points lie from lower left corner to upper right corner.

In case of a negative correlation the plotted points concentrate from upper left to lower right corner

In case of zero correlation, the plotted points would be equally distributed without depicting any particular pattern.

### KARL PEARSON'S PRODUCT MOMENT CORRELATION COEFFICIENT

This is by far the best method for finding correlation between two variables provided the relationship between two variables is linear.

Pearson's correlation coefficient may be defined as the ratio of covariance between the two variables to the product of the standard deviations of the two variables.

If the two variables are denoted by  $x$  and  $y$  and if the corresponding bivariate data are  $(x_i, y_i)$  for  $i = 1, 2, 3, \dots, n$ , then the coefficient of correlation between  $x$  and  $y$ , due to Karl Pearson, is given by

$$r = r_{xy} = \frac{\text{cov}(x, y)}{S_x \times S_y}$$

where

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

$$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2}$$

$$S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}} = \sqrt{\frac{\sum y_i^2}{n} - \bar{y}^2}$$

A single formula for computing correlation coefficient is given by

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

In case of bivariate frequency distribution, we have

$$\text{cov}(x, y) = \frac{\sum_{i,j} x_i y_j f_{ij}}{N} - \bar{x} \bar{y}$$

$$S_x = \sqrt{\frac{\sum_i f_{i0} x_i^2}{N} - \bar{x}^2}$$

$$S_y = \sqrt{\frac{\sum_j f_{0j} y_j^2}{N} - \bar{y}^2}$$

Where

$x_i$  = Mid value of the  $i$ th class interval of  $x$ .

$y_j$  = Mid value of the  $j$ th class interval of  $y$ .

$f_{i0}$  = Marginal frequency of  $x$

$f_{0j}$  = Marginal frequency of  $y$

$f_{ij}$  = frequency of the  $(i, j)$ th cell

$$N = \sum_{ij} f_{ij} = \sum_i f_{i0} = \sum_j f_{0j} = \text{Total Frequency}$$

### Properties of Correlation Coefficient:

- The coefficient of correlation is a unit –free measure
- The coefficient of correlation remains invariant under a change of origin and/or scale of the variables under consideration depending on the sign of scale factors.

This property states that if the original pair of variables  $x$  and  $y$  is changed to a new pair of variables  $u$  and  $v$  by effecting a change of origin and scale for both  $x$  and  $y$  i.e.

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

where  $a$  and  $c$  are the origins of  $x$  and  $y$  and  $b$  and  $d$  are the respective scales and then we have

$$r_{xy} = \frac{bd}{|b||d|} r_{uv}$$

- The coefficient of correlation always lies between -1 and 1, including both the limiting values.

### A) SPEARMAN'S RANK CORRELATION COEFFICIENT:

- When we need finding correlation between two qualitative characteristics, we take recourse to using rank correlation coefficient.
- Spearman's rank correlation coefficient is given by

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

Where  $r_R$  = Rank correlation coefficient and it lies between -1 and 1 inclusive of these two values.

$d_i = x_i - y_i$  represents the difference in ranks for  $i$ th individual and  $n$  denotes the number of individuals.

AUTHOR: SUPRITHA MARLA