Name: Dr.P K Sivakumaran,
Assistant Professor

Subject Code : 18BST46S

Subject Name :
Psychological Statistics

## 12·1. CORRELATION RATIO

As discussed earlier, when variables are linearly related, we have regression lines of one variable on another variable and correlation coefficient can be computed to tell us about the extent of association between them. However, if the variables are not linearly related but some sort of curvilinear relationship exists between them, the use of $r$ which is a measure of the degree to which the relation approaches straight line "law" will be misleading. We might come across bivariate distributions where $r$ may be very low or even zero but the regression may be strong, or even perfect. *Correlation ratio* '$\eta$' is the appropriate measure of curvilinear relationship between the two variables. Just as $r$ measures the concentration of points about the straight line of best fit, $\eta$ measures the concentration of points about the curve of best fit. If regression is linear $\eta = r$, otherwise $|\eta| > |r|$ (*c.f.* Remark 2, § 12·1).

**12·1·1. Measure of Correlation Ratio.** In the previous articles we have assumed that there is a single observed value $Y$ corresponding to the given value $x_i$ of $X$ but sometimes there are more than one such value of $Y$.

Suppose corresponding to the value $x_i$ $(i = 1, 2, ...,m)$ of the variable $X$, the variable $Y$ takes the values $y_{ij}$ with respective frequencies $f_{ij}$; $j = 1, 2, ..., n$.

Though all the $x$'s in the $i$th vertical array have the same value, the $y$'s are different. A typical pair of values in the $i$th array is $(x_i, y_{ij})$, with frequency $f_{ij}$. Thus the first suffix $i$ indicates the vertical array while the second suffix $j$ indicates the positions of $y$ in that array. Let

$$\sum_{j=1}^{n} f_{ij} = n_i \quad \text{and} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} f_{ij}\right) = \sum_{i=1}^{m} n_i = N, \qquad \text{(say)}.$$

If $\bar{y}_i$ and $\bar{y}$ denote the means of the $i$th array and the overall mean respectively,

then $$\bar{y}_i = \frac{\sum_j f_{ij}\, y_{ij}}{\sum_j f_{ij}} = \frac{\sum_j f_{ij}\, y_{ij}}{n_i} = \frac{T_i}{n_i} \quad \text{and} \quad \bar{y} = \frac{\sum_i \sum_j f_{ij}\, y_{ij}}{\sum_i \sum_j f_{ij}} = \frac{\sum_i n_i \bar{y}_i}{\sum_i n_i} = \frac{T}{N}$$

In other words $\bar{y}$ is the weighted mean of all the array means, the weights being the array frequencies.

***Def.*** *The correlation ratio of $Y$ on $X$, usually denoted by $\eta_{YX}$ is given by :*

$$\eta_{YX}^2 = 1 - \frac{\sigma_{eY}^2}{\sigma_Y^2}, \qquad \ldots (12\cdot1)$$

where $$\sigma_{eY}^2 = \frac{1}{N}\sum_i \sum_j f_{ij}\,(y_{ij} - \bar{y}_i)^2 \text{ and } \sigma_Y^2 = \frac{1}{N}\sum_i \sum_j f_{ij}\,(y_{ij} - \bar{y})^2 \qquad \ldots (12\cdot1a)$$

A convenient expression for $\eta_{YX}$ can be obtained in terms of standard deviation $\sigma_{mY}$ of the means of vertical arrays, each mean being weighted by the array frequency.

We have

$$N\sigma_Y^2 = \sum\sum f_{ij}\,(y_{ij} - \bar{y})^2 = \sum_i \sum_j f_{ij}\,\{(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})\}^2$$

$$= \sum_i \sum_j f_{ij}\,(y_{ij} - \bar{y}_i)^2 + \sum_i \sum_j f_{ij}\,(\bar{y}_i - \bar{y})^2 + 2\sum_i \sum_j f_{ij}\,(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y})$$

The term $2\left[\left(\sum_i (\bar{y}_i - \bar{y})\left\{\sum_j f_{ij}(y_{ij} - \bar{y}_i)\right\}\right]\right.$ vanishes since $\sum_j f_{ij}(y_{ij} - \bar{y}_i) = 0$, being the algebraic sum of the deviations from mean.

$$\therefore \quad N\sigma_Y^2 = \sum_i \sum_j f_{ij}(y_{ij} - \bar{y}_i)^2 + \sum_i n_i(\bar{y}_i - \bar{y})^2$$

$$\Rightarrow \quad N\sigma_Y^2 = N\sigma_{eY}^2 + N\sigma_{mY}^2 \quad \Rightarrow \quad \sigma_Y^2 = \sigma_{eY}^2 + \sigma_{mY}^2$$

where $\sigma_{mY}^2 = \frac{1}{N}\sum n_i(\bar{y}_i - \bar{y})^2$, is variance of means of the vertical arrays.

$$\Rightarrow \quad 1 - \frac{\sigma_{eY}^2}{\sigma_Y^2} = \frac{\sigma_{mY}^2}{\sigma_Y^2}, \text{ which on comparison with (12·1), gives}$$

$$\eta_{YX}^2 = \frac{\sigma_{mY}^2}{\sigma_Y^2} = \frac{\sum_i n_i(\bar{y}_i - \bar{y})^2}{\sum_i \sum_j f_{ij}(y_{ij} - \bar{y})^2} \qquad \ldots (12\cdot2)$$

We have: $N\sigma_{mY}^2 = \sum_i n_i(\bar{y}_i - \bar{y})^2 = \sum_i n_i \bar{y}_i^2 - N\bar{y}^2 = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N}$

$$\therefore \quad \eta_{YX}^2 = \left[\sum_i \left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right] \Big/ N\sigma_Y^2, \qquad \ldots (12\cdot3)$$

a formula, much more convenient for computational purposes.

***Remarks*** **1.** (12·1) implies that : $\sigma_{eY}^2 = \sigma_Y^2\ (1 - \eta_{YX}^2)$

Since $\sigma_{eY}^2$ and $\sigma_Y^2$ are non-negative, we have

$$1 - \eta_{YX}^2 \geq 0 \quad \Rightarrow \quad \eta_{YX}^2 \leq 1 \quad \text{or} \quad |\eta_{YX}| \leq 1 \qquad \ldots (12\cdot3a)$$

**2.** Since the sum of squares of deviations in any array is minimum when measured from its mean, we have :

$$\sum_i \sum_j f_{ij}(y_{ij} - \bar{y}_i)^2 \leq \sum_i \sum_j f_{ij}(y_{ij} - \hat{y}_{ij})^2, \qquad \ldots (*)$$

where $\hat{y}_{ij}$ is the estimate of $y_{ij}$ for given value of $X = x_i$, say, as given by the line of regression of $Y$ on $X$, *i.e.*, $\hat{y}_{ij} = a + bx_i$, $(j = 1, 2, \ldots, n)$.

But $\sum_i \sum_j f_{ij}(y_{ij} - \bar{y}_i)^2 = N\sigma_{eY}^2 = N\sigma_Y^2(1 - \eta_{YX}^2)$ and $\sum_i \sum_j f_{ij}(y_{ij} - a - bx_i)^2 = N\sigma_Y^2(1 - r^2)$

$\ldots (12\cdot3b)$

$$\therefore (*) \quad \Rightarrow \quad 1 - \eta_{YX}^2 \leq 1 - r^2 \quad \Rightarrow \quad \eta_{YX}^2 \geq r^2 \quad \Rightarrow \quad |\eta_{YX}| \geq |r|$$

Thus the absolute value of the correlation ratio can never be less than the absolute of $r$, the correlation coefficient.

When the regression of $Y$ on $X$ is linear, straight line of means of arrays coincides with the line of regression and $\eta_{YX}^2 = r^2$. Thus $\eta_{YX}^2 - r^2$ is the departure of regression from linearity. It is also clear (from Remark 1) that the more nearly $\eta_{YX}^2$ approaches unity, the smaller is $\sigma_{eY}^2$ and, therefore, closer are the points to the curve of means of vertical arrays.

When $\eta_{YX}^2 = 1$, $\sigma_{eY}^2 = 0 \quad \Rightarrow \quad \sum\sum f_{ij}(y_{ij} - \bar{y}_i)^2 = 0$ or $y_{ij} = \bar{y}_i$, $\forall\, j = 1, 2, \ldots, n$, *i.e.*, all the points lie on the curve of means. This implies that there is a functional relationship between $X$ and $Y$. $\eta_{YX}$ is, therefore, the measure of the degree to which the association between the variables approaches a functional relationship of the form $Y = \phi(X)$, where $\phi(X)$ is a single valued function of $X$, $[\phi(X) = a + bX]$.

**3.** It is worth noting that the value of $\eta_{YX}$ is not independent of the classification of the data. As the class intervals become narrower $\eta_{YX}$ approaches unity, since in that case $\sigma_{mY}^2$ gets nearer

to $\sigma_Y^2$. If the grouping is so fine that only one item appears in each row (related to each $x$-class), that item will constitute the mean of that column and thus in this case $\sigma_{mY}^2$ and $\sigma_Y^2$ become equal so that $\eta_{YX}^2 = 1$. On the other hand, a very coarse grouping tends to make the value of $\eta_{YX}$ approach $r$. "Student" has given a formula for 'the correction' to be made in the correlation ratio 'for grouping' in Biometrika (Vol IX page 316-320)

4. It can be easily proved that $\eta_{YX}^2$ is independent of change of origin and scale of measurements.

5. $\eta_{XY}^2$, the second correlation ratio of $X$ on $Y$ depends upon the scatter of observations about the line of column means.

6. $r_{XY}$ and $r_{YX}$ are same but $\eta_{YX}$ is, in general, different from $\eta_{XY}$.

7. In terms of expectation, correlation ratio is defined as follows :

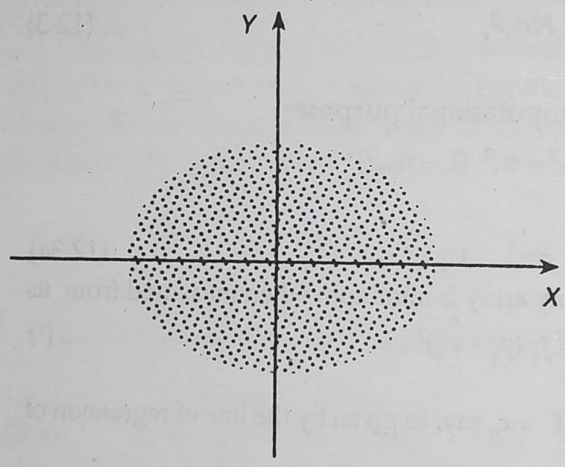
$$\eta_{YX}^2 = \frac{E_X[E(Y \mid X) - E(Y)]^2}{E[Y - E(Y)]^2} = \frac{E_X[E(Y \mid X) - E(Y)]^2}{\sigma_Y^2}$$

and

$$\eta_{XY}^2 = \frac{E_Y[E(X \mid Y) - E(X)]^2}{E[X - E(X)]^2} = \frac{E_Y[E(X \mid Y) - E(X)]^2}{\sigma_X^2}$$

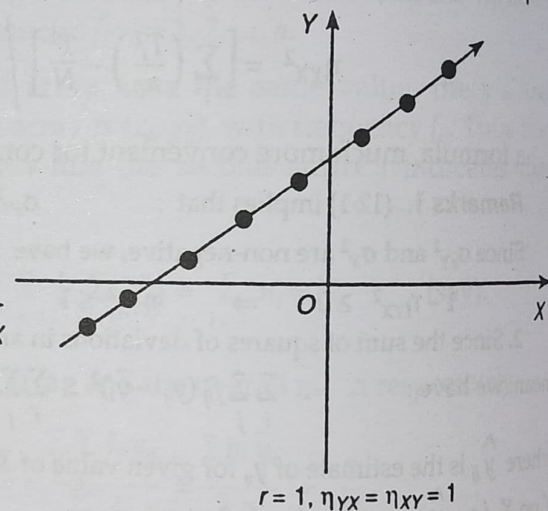
8. We give below some diagrams, exhibiting the relationship between $r$ and $\eta_{YX}$.

(*i*) For completely random scattering of the dots with no trend, both $\rho$ and $n$ are zero



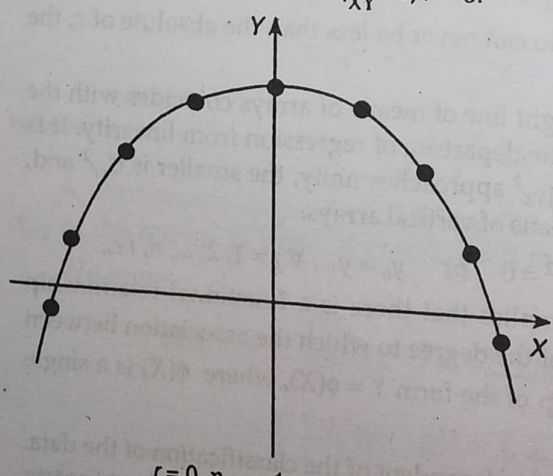

$r = 0, \eta_{YX} = \eta_{XY} = 0$

(*ii*) If dots lie precisely on a line, $r = 1$ and $\eta = 1$



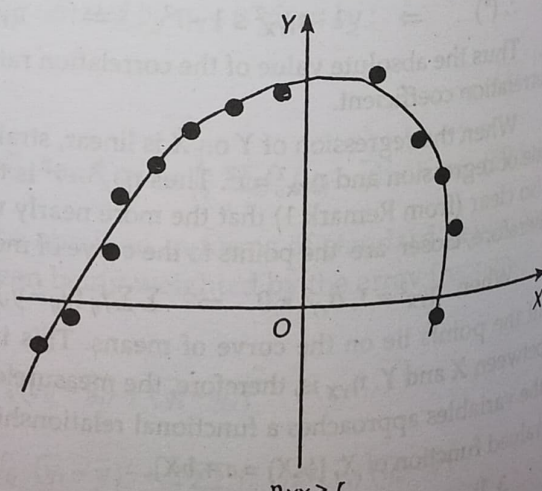

$r = 1, \eta_{YX} = \eta_{XY} = 1$

(*iii*) If dots lie on a curve, such that no ordinate cuts it more than once, $\eta_{YX} = 1$ and if furthermore, the dots are symmetrically placed about $Y$ - axis, then $\eta_{XY} = 0, r = 0$.




$r = 0, \eta_{YX} = 1, \eta_{XY} = 0$

(*iv*) If $\eta_{YX} > r$, the dots are scattered around a definitely curved trend line.




$\eta_{YX} > r$

## 12·2. INTRA-CLASS CORRELATION

Intra-class correlation means within class correlation. It is distinguishable from product moment correlation in as much as here both the variables measure the same characteristics. Sometimes specially in biological and agricultural study, it is of interest to know how the members of a family or group are correlated among themselves with respect to some one of their common characteristic. For example, we may require the correlation between the heights of brothers of a family or between yields of plots of an experimental block. In such cases both the variables measure the same characteristic, *e.g.*, height and height or weight and weight. There is nothing to distinguish one from the other so that one may be treated as $X$- variable and the other as the $Y$-variable.

Suppose we have $A_1, A_2, \ldots, A_n$ families with $k_1, k_2, \ldots, k_n$ members, each of which may be represented as shown below :

$$\begin{array}{ccccccc} x_{11} & x_{21} & \cdots & x_{i1} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{i2} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{1j} & x_{2j} & \cdots & x_{ij} & \cdots & x_{nj} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{1k_1} & x_{2k_2} & \cdots & x_{ik_i} & \cdots & x_{nk_n} \end{array}$$

and let $x_{ij}$ ($i = 1, 2, \ldots, n\ ;\ j = 1, 2, \ldots, k_i$) denote the measurement on the $j$th member in the $i$th family.

We shall have $k_i(k_i - 1)$ pairs for the $i$th family or group like $(x_{ij}, x_{il})$, $j \neq l$. There will be $\sum_{i=1}^{n} k_i (k_i - 1) = N$, pairs for all the $n$ families or groups. If we prepare a correlation table there will be $k_i\ (k_i - 1)$ entries for the $i$th group or family and $\sum_i k_i(k_i - 1) = N$ entries for all the $n$ families or groups. The table is symmetrical about the principal diagonal. Such a table is called an *intra-class correlation table* and the correlation is called *intra-class correlation.*

In the bivariate table $x_{i1}$ occurs $(k_i - 1)$ times, $x_{i2}$ occurs $(k_i - 1)$ times, ..., $x_{ik_i}$ occurs $(k_i - 1)$ times, *i.e.*, from the $i$th family we have $(k_i - 1)\sum_j x_{ij}$ and hence for all the $n$ families we have $\sum_i \left[ (k_i - 1) \sum_j x_{ij} \right]$ as the marginal frequency, the table being symmetrical about principal diagonal.

$$\therefore \qquad \bar{x} = \bar{y} = \frac{1}{N} \left[ \sum_i \{(k_i - 1) \sum_j x_{ij}\} \right]$$

$$\text{Similarly, } \sigma_X^2 = \sigma_Y^2 = \frac{1}{N} \left[ \sum_i \{(k_i - 1) \sum_j (x_{ij} - \bar{x})^2 \} \right]$$

$$\text{Further Cov } (X, Y) = \frac{1}{N} \sum_i \left[ \sum_{j,l} (x_{ij} - \bar{x})(x_{il} - \bar{x}) \right], j \neq l$$

$$= \frac{1}{N} \sum_i \left[ \sum_{j=1}^{k_i} \sum_{l=1}^{k_i} (x_{il} - \bar{x})(x_{ij} - \bar{x}) - \sum_{j=1}^{k_i} (x_{ij} - \bar{x})^2 \right]$$

If we write $\bar{x}_i = \sum_j x_{ij}/k_i$, then

$$\sum_i \left[ \sum_{j=1}^{k_i} \sum_{l=1}^{k_i} (x_{ij} - \bar{x})(x_{il} - \bar{x}) \right] = \sum_i \left[ \sum_j (x_{ij} - \bar{x}) \sum_i (x_{il} - \bar{x}) \right] = \sum_i \left[ k_i (\bar{x}_i - \bar{x})\, k_i (\bar{x}_i - \bar{x}) \right]$$

$$= \sum_i k_i^2 (\bar{x}_i - \bar{x})^2$$

Therefore intra-class correlation coefficient is given by :

$$r(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{V(X)\, V(Y)}} = \frac{\sum_i k_i^2 (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{\sum_i \sum_j (k_i - 1)(x_{ij} - \bar{x})^2} \quad \ldots (12\cdot4)$$

If we put $k_i = k$, *i.e.*, if all families have equal members then :

$$r = \frac{k^2 \sum_i (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{(k-1) \sum_i \sum_j (x_{ij} - \bar{x})^2} = \frac{nk^2 \sigma_m^2 - nk\sigma^2}{(k-1)nk\sigma^2}$$

$$= \frac{1}{(k-1)} \left\{ \frac{k\sigma_m^2}{\sigma^2} - 1 \right\} \quad \ldots (12\cdot4a)$$

where $\sigma^2$ denotes the variance of $X$ and $\sigma^2_m$ the variance of means of families.

***Limits.*** We have from (12·4*a*),

$$1 + (k-1)r = \frac{k\sigma_m^2}{\sigma^2} \geq 0 \quad \Rightarrow \quad r \geq -\frac{1}{(k-1)}$$

Also $1 + (k-1)r \leq k$ as the ratio $\dfrac{\sigma_m^2}{\sigma^2} \leq 1 \quad \Rightarrow \quad r \leq 1$

$\therefore$ $$-\frac{1}{(k-1)} \leq r \leq 1. \quad \ldots (12\cdot4b)$$

***Interpretation.*** Intra-class correlation cannot be less than $-1/(k-1)$, though it may attain the value +1 on the positive side, so that it is a skew coefficient and a negative value has not the same significance as a departure from independence as an equivalent positive value.

# Partial and Multiple Correlation

## NEED AND IMPORTANCE OF PARTIAL CORRELATION

While conducting studies in the field of education and psychology, we often find that the relationship between two variables is greatly influenced by a third variable or an additional variable. In such a situation, it becomes quite difficult to have a reliable and an independent estimate of correlation between these two variables unless there is some possibility of nullifying the effects of a third variable (or a number of other variables) on the variables in question. It is the partial correlation which helps in such a situation for nullifying the undesired influence of a third or any additional variable on the relationship of the two variables.

Let us illustrate such a possibility with the help of an example. Suppose in a study, we want to know the effect of participation in co-curricular activities upon the academic achievement. For this, we take the sample of students studying in various schools. We collect two types of scores regarding their performance in co-curricular activities and academic achievement and then try to find a measure of correlation between these two variables. A close analysis of the factors affecting the academic achievement or participation in co-curricular activities may reveal that both these variables are certainly influenced by so many other factors or variables like intelligence, socio-economic status, environmental differences, age, health and physique and other similar factors.

However, in our study, we are only concerned with the evaluation of the correlation between participation in co-curricular activities and academic achievement. Our aim is to have an independent and a reliable measure of correlation between these two variables. It can only happen when we first adopt some measures to nullify the effect of the intervening variables like intelligence, socio-economic status, age, and health, on both the variables being correlated. In other words, there is an urgent need for exercising control over all other variables and factors except the two whose relationship we have to measure.

There are two ways of controlling or ruling out the influence of undesirable or intervening variables on the two variables being correlated. One calls for experimental technique in which we can select students of the same intelligence, socio-economic status, age, health and physique, and thus apply the matching pair technique. However, to select and make use of such a sample is quite impracticable. Here we may have to reduce drastically the size of our sample. Otherwise, matching them for so many factors will be a cumbersome task; also, it will neither be feasible nor appropriate.

Another, and the most practicable and convenient way, is to exercise statistical control. In this, we hold the undesirable or the intervening variables constant through the partial correlation method. Here, we can make use of the whole data without sacrificing any information as needed in the experimental method, for making equal and matching pairs.

Partial correlation can thus be described as a special correlation technique, which is helpful in estimating independent and reliable relationship between any two variables by eliminating and ruling out any undesirable influence or interference of a third or an additional variable on the variables being correlated.

*Assumption.* The partial correlation technique is based on the following important assumption:

> *Control the main variables (two or three) and they will automatically control so many other variables since other variables are related to these controlled variables.*

Working on this assumption in the following example, we can attempt at eliminating a few intervening variables like intelligence and socio-economic status to study the correlation between academic achievement and participation in co-curricular activities. The other intervening variables such as age, health and physique, environmental condition, education and health of the parents, living habits and temperament will be automatically controlled.

## Computation of Partial Correlation

We now give some formulae for the computation of partial correlation.

1. **Formulae for the Computation of First Order Partial Correlation:**

$$r_{123} = \frac{r_{12} - r_{13} \times r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}$$

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}}$$

First order partial correlations are those in which the relationship between two variables is estimated while making the third variable constant or partialled out.

2. **Formula for the Computation of Second Order Partial Correlation:**

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}\, r_{24.3}}{\sqrt{1-r_{14.3}^2}\sqrt{1-r_{24.3}^2}}$$

3. **Formula for the Computation of Third Order Partial Correlation:**

$$r_{12.345} = \frac{r_{12.34} - r_{15.34}\, r_{25.34}}{\sqrt{1-r_{15.34}^2}\sqrt{1-r_{25.34}^2}}$$

The second or the third order partial correlations are those correlations in which the relationship between two variables is estimated while making two or three other variables constant or partialled out.

Let us now clarify the concept and working of these formulae with the help of an example. Let

1 = Achievement scores
2 = Participation in co-curricular activities scores
3 = IQ scores
4 = Socio-economic status scores
5 = Age score

Hence in the first order partial correlation, $r_{12.3}$ means the correlation between 1 and 2 (achievement and participation in co-curricular activities) while making the third variable "intelligence" as constant or partialled out.

In the second order partial correlation, $r_{12.34}$ means correlation between 1 and 2 (achievement and participation) while making the third and fourth ("intelligence" and "socio-economic status") as constant or partialled out.

In the third order partial correlation, $r_{12.345}$ means the correlation between 1 and 2 while making the third, fourth and fifth variables (i.e. "intelligence", "socio-economic status" and "age") constant or partialled out.

In this way, the number to the right of the decimal point (here in first order 3, second order 34, third order 345) represents variables

whose influence is ruled out and the number to the left (here 12) represents those two variables whose relationship needs to be estimated.

It is also clear from the given formulae that, for computation of second order correlations, all the necessary first order correlations have to be computed, and for the third order correlations, the second order correlations must be known.

Let us now illustrate the process of computation of partial correlation with the help of examples.

***Example 13.1:*** From a certain number of schools in Delhi, a sample of 500 students studying in classes IX and X was taken. These students were evaluated in terms of their academic achievement and participation in co-curricular activities. Their IQ's were also tested. The correlation among these three variables was obtained and recorded as follows:

$$r_{12} = 0.80, \quad r_{23} = 0.70, \quad r_{13} = 0.60$$

Find out the independent correlation between the main (first two) variables—academic achievement and participation—in co-curricular activities.

***Solution.*** The independent and reliable correlation between academic achievement and participation in co-curricular activities—the main two variables—can be found by computing partial correlation between these two variables, i.e. by computing $r_{12.3}$ (keeping constant the third variable)

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1-r_{13}^2}\,\sqrt{1-r_{23}^2}}$$

Substituting the given value of correlation in the foregoing formula, we obtain

$$r_{12.3} = \frac{0.80 - 0.60 \times 0.70}{\sqrt{1 - 0.60 \times 0.60}\,\sqrt{1 - 0.70 \times 0.70}}$$

$$= \frac{0.80 - 0.42}{\sqrt{1 - 0.36}\,\sqrt{1 - 0.49}}$$

$$= \frac{0.38}{\sqrt{0.64}\,\sqrt{0.51}} = \frac{0.38}{0.8 \times 0.714}$$

$$= \frac{0.38}{0.5712} = \frac{3800}{5712} = \frac{1900}{2856} = 0.67$$

[**Ans.** Partial correlation = 0.67.]

***Example 13.2:*** While pursuing further with the experiment in the sample cited in Example 13.1, the researcher also collected data regarding socio-economic status (fourth variable) of all the 500 students and then recorded the computed correlations as follows:

$$r_{12} = .80, \quad r_{23} = .70, \quad r_{13} = .60, \quad r_{14} = .50$$
$$r_{24} = .40 \text{ and } r_{34} = .30$$

Compute the independent correlation between academic achievement and participation in co-curricular activities of the main variables.

***Solution.*** Here, the researcher has to exercise control or partial out the two intervening variables, viz. variable 3 (i.e. intelligence), and variable 4 (i.e. socio-economic status). Hence the problem requires the computation of second order partial correlation.
The formula required for such a correlation is

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}\, r_{24.3}}{\sqrt{1 - r^2_{14.3}}\sqrt{1 - r^2_{24.3}}}$$

Since we have already computed the value of $r_{12.3}$ in Example 13.1 as $r_{12.3} = 0.67$, now we have to compute only the values of $r_{14.3}$ and $r_{24.3}$. Thus,

$$r_{14.3} = \frac{r_{14} - r_{13}\, r_{34}}{\sqrt{1 - r^2_{13}}\sqrt{1 - r^2_{34}}}$$

$$= \frac{.50 - (.60 \times .30)}{\sqrt{1 - .60 \times .60}\sqrt{1 - .30 \times .30}}$$

$$= \frac{.50 - .18}{\sqrt{1 - .36}\sqrt{1 - .09}} = \frac{.32}{\sqrt{.64 \times .91}}$$

$$= \frac{3200}{\sqrt{64 \times 91}} - \frac{50}{91} = 0.42$$

$$r_{24.3} = \frac{r_{24} - r_{23}\, r_{34}}{\sqrt{1 - r^2_{23}}\sqrt{1 - r^2_{34}}}$$

$$= \frac{.40 - .70 \times .30}{\sqrt{1 - .70 \times .70}\sqrt{1 - .30 \times .30}}$$

$$= \frac{.40 - .21}{\sqrt{1 - .49}\sqrt{1 - .09}} = \frac{.19}{\sqrt{.51}\sqrt{.91}}$$

$$= \frac{.19}{.714 \times .953} = \frac{190{,}000}{714 \times 953} = 0.28$$

$$r_{12.34} = \frac{.67 - (.55 \times .28)}{\sqrt{1 - (.55 \times .55)}\sqrt{1 - (.28 \times .28)}}$$

$$= \frac{.5160}{\sqrt{1 - .3205}\sqrt{1 - .0784}}$$

$$= \frac{.5160}{\sqrt{.6975}\sqrt{.9216}} = \frac{.5160}{.835 \times .96}$$

$$= \frac{.5160}{.8016} = \frac{5160}{8016} = .64$$

[**Ans.** Partial correlation = 0.64.]

## Application of Partial Correlation

Partial correlation can be used as a special statistical technique for eliminating the effects of one or more variables on the two main variables, for which we want to compute an independent and a reliable measure of correlation. Besides its major advantage lies in the fact that it enables us to set up a multiple regression equation (see Chapter 14) of two or more variables, by means of which we can predict another variable or criterion.

## Significance of Partial Correlation Coefficient

The significance of the first and the second order partial correlation '*r*' can be tested easily by using the '*t*' distribution

$$t = r\sqrt{\frac{N - 2 - K}{1 - r^2}}$$

where

$K$ = Order of partial $r$
$r$ = Value of partial correlation
$N$ = Total frequencies in the sample study

Therefore,

$$\text{Degree of freedom} = N - 2 - K$$

# NEED AND IMPORTANCE OF MULTIPLE CORRELATION

In many studies related to education and psychology, we find that a variable is dependent on a number of other variables called *independent variables*. For example, if we take the case of one's academic achievement,

it may be found associated with or dependent on variables like intelligence, socio-economic status, education of the parents, the methods of teaching, the quality of teachers, aptitude, interest, environmental set-up, number of hours spent on studies and so on. All these independent variables affect the achievement scores obtained by the students. If we want to study the combined effect or influence of these variables on a single dependent variable, then we have to compute a special statistic called *coefficient of multiple correlation*.

The coefficient of multiple correlation signifies a statistic used for denoting the strength of relationship between one variable, called *dependent variable,* and two or more variables, called *independent variables.* In simple terms, therefore, by multiple correlation, we mean the relationship between one variable and a combination of two or more variables.

Let us make the meaning of the term clearer by taking the simple case of one dependent variable and two independent variables, known for exercising their influence on the dependent variable. Once again, let this dependent variable be "academic success" (to be obtained from achievement scores in an examination). We can have two independent variables which are well known for their impact on academic success, namely, general intelligence (to be known through a battery of intelligence tests) and socio-economic status (known through a scale). We can name the dependent variable as $X_1$, and the two independent variables as $X_2$ and $X_3$. In order to compute multiple correlation here, we try to find out the measure of correlation between $X_1$ and the combined effects of $X_2$ and $X_3$. In such a case, we designate the required multiple correlation coefficient as $R_{1.23}$, meaning thereby that we are computing a correlation between the dependent variable 1, and the combination of the two independent variables 2 and 3.

## Computation of Multiple Correlation

**Formula for Computation of Multiple Correlation Coefficient:**

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

where, $R_{1.23}$ denotes the coefficient of multiple correlation between the dependent variable $X_1$ and the combination of two independent variables $X_2$ and $X_3$, in

$r_{12}$ = Correlation between $X_1$ and $X_2$
$r_{13}$ = Correlation between $X_1$ and $X_3$
$r_{23}$ = Correlation between $X_2$ and $X_3$

Let us now illustrate the use of this formula.

***Example 13.3:*** In a study, a researcher wanted to know the impact of a person's intelligence and his socio-economic status on his academic success. For computing the coefficient of multiple correlation, he collected the required data and computed the following inter-correlations:

$$r_{12} = 0.60, \qquad r_{13} = 0.40, \qquad r_{23} = 0.50$$

where 1, 2, 3 represent the variables "academic success", "intelligence" and "socio-economic status", respectively. In this case, find out the required multiple correlation coefficient.

***Solution.*** The multiple correlation coefficient

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

Substituting the respective values of $r_{12}$, $r_{13}$ and $r_{23}$ in the above formula, we get

$$R_{1.23} = \sqrt{\frac{0.60^2 + 0.40^2 - 2 \times 0.60 \times 0.40 \times 0.50}{1 - (0.50)^2}}$$

$$= \sqrt{\frac{0.36 + 0.16 - 0.24}{1 - 0.25}} = \sqrt{\frac{0.28}{0.75}} = \sqrt{0.3733}$$

[**Ans**. Multiple correlation coefficient = 0.61.]

## Other Methods of Computing Coefficient of Multiple Correlation

Multiple correlation coefficient can also be computed by other means like below.

**First**, it can be done with the help of partial correlation. Thus,

$$R_{1.23} = \sqrt{1 - \left(1 - r_{12}^2\right)\left(1 - r_{13.2}^2\right)}$$

In this formula, to understand the relationship between variable 1 and the combined effect of variables 2 and 3, we compute the multiple correlation coefficient. The formula requires the values of $r_{12}$ (correlation between 1 and 2) and the partial correlation $r_{13.2}$, which can be computed by using the formula

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}$$

However, in case if there are four variables instead of three, then we can compute the multiple correlation coefficient as

$$R_{1.234} = \sqrt{1-\left(1-r_{12}^2\right)\left(1-r_{13.2}^2\right)\left(1-r_{14.23}^2\right)}$$

where

$r_{12}$ = Correlation between 1 and 2
$r_{13.2}$ = First order partial correlation
$r_{14.23}$ = Second order partial correlation

The value of $r_{14.23}$ can be computed by the following formula:

$$r_{14.23} = \frac{r_{14.2} - r_{13.2}\, r_{34.2}}{\sqrt{1-r_{13.2}^2}\sqrt{1-r_{34.2}^2}}$$

**Secondly,** multiple correlation coefficient can also be computed with the help of standard partial regression coefficient called betas, which are used in multiple regression equation (see Chapter 14). Now,

$$R_{1.23} = \sqrt{\beta_{12.3}\, r_{12} + \beta_{13.2}\, r_{13}}$$

where

$$\beta_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{1-r_{23}^2}$$

$$\beta_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{1-r_{23}^2}$$

*Note:* These two methods are used only when we have the required partial correlations or the values of $\beta_1$, $\beta_2$, .... In most of the cases, it is quite economical to make use of the general formula

$$R_{12.3} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\cdot r_{13}\cdot r_{23}}{1-r_{23}^2}}$$

Let us make use of this formula to find solutions to some more problems.

***Example 13.4:*** A researcher was interested in studying the relationship between success in a job and the training received. He collected data regarding these and treated them as the two main variables and added a third variable, "interest" (measured by interest inventory). The correlations among these three variables were

$$r_{12} = .41, \qquad r_{13} = .50, \qquad r_{23} = .16.$$

This data enabled him to compute the multiple correlation coefficient and thus to find out the relationship between success in the job and the combined effect of the two independent variables—"training" and "interest". Find the value of the correlation coefficient in his study.

***Solution.*** The multiple correlation coefficient is given by

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13} r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{0.41^2 + 0.50^2 - 2 \times 0.41 \times 0.50 \times 0.16}{1 - (0.16)^2}}$$

$$= \sqrt{\frac{0.1681 + 0.2500 - 0.0656}{1 - 0.0256}}$$

$$= \sqrt{\frac{0.4181 - 0.0656}{0.9744}} = \sqrt{\frac{0.3525}{0.9744}} = \sqrt{0.3618}$$

$$= 0.601$$

[**Ans.** Multiple correlation coefficient = 0.601.]

***Example 13.5:*** One thousand candidates appeared for an entrance test. The test had some sub-tests, namely, general intelligence test, professional awareness test, general knowledge test and aptitude test. A researcher got interested in knowing the impact or the strength of the association of any two sub-tests on the total entrance test scores ($X_1$). Initially, he took two sub-tests scores—intelligence test scores ($X_2$) and professional awareness scores ($X_3$)—and derived the necessary correlations. Compute the multiple correlation coefficient for measuring the strength of relationship between $X_1$ and $(X_2 + X_3)$, if $r_{12} = .80$, $r_{13} = .70$ and $r_{23} = .60$.

***Solution.*** The multiple correlation coefficient

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{0.80^2 + 0.70^2 - 2 \times 0.80 \times 0.70 \times 0.60}{1 - 0.60^2}}$$

$$= \sqrt{\frac{0.640 + 0.490 - 0.672}{1 - 0.36}} = \sqrt{\frac{0.458}{0.64}}$$

$$= \sqrt{\frac{458}{640}} = \sqrt{0.7156} = 0.845$$

[**Ans.** Multiple correlation coefficient = 0.845.]

## Characteristics of Multiple Correlation

- Multiple correlation is the correlation that helps one to measure the strength of association of a dependent variable with two or more independent variables.
- It is related to the correlations of independent variables as well as to the correlations of these variables with the dependent variable.
- It helps us estimate the combined effect or influence of the independent variables on the dependent variable.
- It helps in the selection and rejection of the sub-tests for a particular test or tests.
- For proper computation of multiple correlation it is desirable that the number of cases and, especially, the number of variables, be large.
- The multiple correlation coefficient $R$ is always positive, less than 1.00, and is greater than the zero order correlation coefficients $r_{12}$, $r_{13}$, ....

## Significance of Multiple Correlation Coefficient $R$

The significance of multiple correlation $R$ can be easily tested with the help of its standard error. The standard error of multiple $R$ can be computed by using the following formula:

$$SE_R = \frac{1 - R^2}{\sqrt{N - m}}$$

where

$m$ = Number of variables being correlated
$N$ = Size of the sample
$N - m$ = Degree of freedom

## EXERCISES

1. What do you mean by partial correlation? Describe its characteristics and applications. Discuss the situations where it is used in educational and psychological studies by citing specific examples.
2. What is multiple correlation? Describe its characteristics and applications. Discuss where you would like to use it in the educational and psychological investigations by citing specific examples.
3. An investigator, during one of his studies, collected some data and arrived at the following conclusions:

   (a) The correlation between height and weight = 0.80
   (b) The correlation between weight and age = 0.50
   (c) The correlation between height and age = 0.60

   Compute the net correlation between height and weight by partialling out the third variable, viz age.
4. (a) Compute the partial correlation coefficients ($r_{23.1}$) from the data given in Example 3 for computing correlation between weight and age by partialling out the third variable, viz. height.

   (b) Compute $r_{13.2}$ the net correlation between height and age, by partialling out the third variable, namely, weight.
5. During test construction, an investigator obtained the following results:

   (a) Correlation between total test score and a sub-test = 0.72

(b) Correlation between total test score and another sub-test = 0.36

(c) Correlation between both the sub-tests = 0.54

Compute the multiple correlation between the total test ($X_1$) and the combination of the sub-tests $X_2$ and $X_3$.

**6.** In an experimental study, a researcher obtained the following results:

(a) The correlation between learning ($X$) and motivation ($Y$) = .67

(b) The correlation between learning ($X$) and hours per week devoted to study ($Z$) = 0.75

(c) The correlation between motivation ($Y$) and hours per week devoted to study ($Z$) = 0.63

Find out the multiple correlation between $X$ (the dependent variable) and the combination of $Y$ and $Z$ (the independent variables).