CORRELATION

The relationship between two are more more variables is called “correlation “ and the variables are said to be correlated. The relationship between two variables is also known as “ covariation”. The term relationship can be used in two different senses, viz, mutual dependence and cause and effect relationship.

Mutual Dependence

Consider the two variables, rate of oxygen consumption and metabolism in organisms. When the oxygen consumption increases, there is increase in the metabolism as well. Similarly when the organism increases its activity (metabolism) it consumes more oxygen. On the other hand, when the oxygen consumption decreases, the activity, i.e., the metabolism decreases. When the organism becomes less active, its oxygen consumption also becomes lesser. A relationship between two variables in which a change in the value of one of the two variables brings about a change in the value of the other variable is said to be ‘mutually dependent’.

Cause and Effect Relationship

A relationship between two variables in which changes in the values of one variable is the cause of the changes in the values of the other variable is said to be ‘cause and effect relationship’ between the two variables. For example, consider the two variables, environmental temperature and the body temperature in the environment. When there is increase in the environmental temperature there is an increase in the body temperature. Here the increase in the environmental temperature is the ‘cause’ and the increase in the body temperature is the ‘effect’ .Such a relationship between two variables is known as ‘cause and effect’ relationship.

The cause and effect relationship between two variables may be either direct or indirect.

The nature of correlation between two variables need not be same at all times. For example, the relationship between height and weight of humans or the length and weight of organisms may not be same. Generally, with increase in the height or length, there is increase in the weight. However, it is common to see people who are tall weighing less and those who are short weighing heavier.

Another important thing to be remembered is about " non- sense" or " spurious" correlation between variables. Any two variables, which do not have any logical or biological basis but show a statistical correlation, are said to have non-sense or spurious correlation. For instance, you may find a correlation between the number of cellular phones and the number of mosquitoes in an area. Unless we are able to establish a logical or biological background for a relationship between the above variables, it is a spurious correlation.

TYPES OF CORRELATION

Correlation between variables may be simple or multiple. A simple correlation deals with only two variables where as a multiple correlation deals with more than two variables. We shall be discussing only the simple correlation. Correlation between two variables may be a positive correlation or a negative correlation. Whether it is positive or negative, it may be linear or non- linear.

Positive Correlation

A correlation between two variables in which, with an increase in the values of one variable the values of the other variable also increases, and with a decrease in the value of the one variable the value of the other variable also decreases, is said to be a positive correlation. In other words, in a positive correlation between two variables the values of both the variables move in the same direction. For example, the correlation between the environmental temperature and the body temperature of poikilotherms is a positive correlation

Negative Correlation

A correlation between two variables in which when there is an increase in the values of one variable, the values of the other variable decreases, and when there is a decrease in the values of one variable the other variable increases, is said to be a negative correlation the values of the two variables move in opposite direction. For example, the correlation between environmental temperature and bacterial growth, having a cause and effect relationship, is negative one. With an increase in the temperature the bacterial growth declines and with a decrease in the temperature the bacterial growth increases.

Linear Correlation

When the values of two variables vary in a constant ratio, the correlation between the two variables is said to be linear. The correlation between the optical density and the intensity of the colour of a solution is an example of linear correlation. The relationship between two variables could be classified either as linear or curvilinear. Relationship between two variables x and y is said to be linear if graph between x and y is represented in the form of straight line whereas if graph is represented by a curve, it is curvilinear. Sine we shall confine our discussion in this book to linear relation only hence we can take a liberty to use the correlation term for linear relation.

PEARSON PRODUCT MOMENT CORRELATION

In order to measure the magnitude of linear relation between two variables, a coefficient known as product moment correlation coefficient is defined. It is denoted by r. For convenience we shall use the term correlation coefficient for product moment correlation coefficient.

Product moment correlation coefficient is defined as an index which measures the linear relation between two variables and is denoted by r.

The formula for r is

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

Limits of Correlation Coefficient

The limit pf r is -1 to + 1 i.e., $-1 \leq r \leq 1$. The value of r = +1 indicates the perfect positive linear relation between the two variables x and y. In such situation increase (decrease) in x by any amount is followed by the increase (decrease) in y by the same amount and vice versa. The relationship between x and y is represented by a straight line making a 45 ^{o} angle with x axis and passing through origin.

The value of r = -1 indicates the perfect negative linear relation between the two variables x and y. Here any amount of increase (decrease) in x is followed by the decrease (increase) in y by the same amount and vice versa. The graph between x and y is a straight line making an angle of 45^{o} with both the axis.

r = 0 represents the absence of linear relation between x and y . Here increase or decrease in x does not effect y at all and vice versa. Here the graph) between x and y are straight lines either parallel to x or y.

Example :

Compute the product moment correlation coefficient between the variables.

X	X^2	Y	Y^2	XY
27	729	20	400	540
23	529	24	576	552

20	400	19	361	380
25	625	18	324	450
26	676	23	529	598
19	361	20	400	380
21	441	19	361	399
20	400	21	441	420
27	729	22	484	594
22	484	18	324	396

$\sum x = 230 \quad \sum x^2 = 5374 \quad \sum y = 204 \quad \sum y^2 = 4200 \quad \sum xy = 4709$

Solution

Here N=10

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

$$= \frac{10X4709 - 230X204}{\sqrt{\left(10X5374 - (230)^2\right)\left(10X4200 - (204)^2\right)}}$$

= 0.2993

SPEARMAN RANK ORDER CORRELATION

Rank correlation is used to find the linear correlation between the two variables where scores are in rank order. It is represented by a greek letter ρ (rho). The formula ρ for is $\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$

where D: difference between the ranks. N: number of paired scores.

Rank correlation ρ is a non parametric statistic. Ranks assume only position of subjects or items in the series: no weightage is given for gaps or differences between adjacent scores. For example, individuals with scores 58, 56, 34 and 16 on a test would be ranked 1,2,3 and 4, although the difference between the first and second, second and third, and third and fourth scores are 2,22 and 18 respectively.

Use of Rank Correlation

Often we come across a situation where sport performance can not be measured objectively due to lack of any existing objective criteria, in such a situation evaluation of athlete is done by means of grading. For example in evaluating the dribbling performance in basketball or judging the passing accuracy in soccer can be done only through ranks. In a situation like this correlation between such variables could be measured only through rank correlation. If scores on one variable are ranks and that of other's are actual measurements, the rank correlation is obtained by converting the actual measurements into their ranks. Further if both the variables are in score form, p can be obtained by converting these scores into their corresponding ranks.

Example

Consider an experiment in which judge's marks were obtained. Calculate rank correlation between the two parameters.

X	15	13	12	14	16	11	10	12
Y	16	10	8	12	10	14	13	10

Solution: Computation of rank correlation between the two parameters is shown in the following table.

X Scores	Y Scores	R_x	R_y	$D= R_x- R_y$	D^2
15	16	2	1	1	1
13	10	4	6	-2	4
12	8	5.5	8	-2.5	6.25
14	12	3	4	-1	1
16	10	1	6	-5	25
11	14	7	2	5	25
10	13	8	3	5	25
12	10	5.5	6	-0.5	0.25
				--	------

$$\sum D^2 = 87.50$$

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

$$\rho = 1 - \frac{6X87.5}{8(8^2 - 1)}$$

= 1-1.04

= -0.04

Remarks

In the above illustration a tie occurs in the X score 12, so both the scores were ranked by means of averaging their ranks, giving each a rank of 5.5. In Y scores too, three scores i.e. 10 had been tied and thus the ranks were averaged and this average rank of 6 was assigned to each score.

Example

Compute the rank order correlation.

X	3.4	4.1	4.2	3.5	4.3	3.3	2.1	4.8	3.4	2.5
Y	175	180	170	164	166	172	150	163	165	145

Solution

X	Y	R_x	R_y	$D= R_x- R_y$	D^2
3.4	175	6.5	2	4.5	20.25
4.1	180	4	1	3	9
4.2	170	3	4	-1	1
3.5	164	5	7	-2	4
4.3	166	2	5	-3	9
3.3	172	8	3	5	25

2.1	150	10	9	1	1
4.8	163	1	8	-7	49
3.4	165	6.5	6	0.5	0.25
2.5	145	9	10	-1	1

$$\sum D^2 = 119.50$$

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

$$\rho = 1 - \frac{6X119.5}{10(10^2 - 1)}$$

= 0.28

Remark Significance of rank correlation can be tested in a manner similar to product moment correlation coefficient.

PHI CORRELATION

The phi correlation coefficient (phi) is one of a number of correlation statistics developed to measure the strength of association between two variables. The phi is a nonparametric statistic used in cross-tabulated table data where both variables are dichotomous. *Dichotomous* means that there are only two possible values for a variable.

For a 2×2 contingency table where a, b, c, and d represent the observation frequencies (the cell count). The formula for phi is:

$$\Phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Example:

Find phi for the following contingency table:

		Politicans	
		Truthful	Not Truthful
Scientists	Truthful	14	10
	Not Truthful	6	13

Solution:

Insert the counts into the formula and solve.

Φ = ad – bc / √((a + b)(c + d)(a + c)(b + d))

Φ = 14*13 – 10*6 / √((14 + 10)(6 + 13)(14 + 6)(10 + 13))

Φ = 182 – 60 / √((24)(19)(20)(23))

Φ = 122/ √((24)(19)(20)(23))

Φ = 122/ 458

Φ = 0.266.

Phi Coefficient

The phi coefficient is a **symmetrical statistic**, which means the independent variable and dependent variables are interchangeable. The interpretation for the phi coefficient is similar to the Pearson Correlation Coefficient. The range is from -1 to 1, where:

- 0 is no relationship.
- 1 is a perfect positive relationship: most of your data falls along the diagonal cells.
- -1 is a perfect negative relationship: most of your data is *not* on the diagonal.

BISERIAL CORRELATION

A Biserial correlation is used to measure the strength and direction of the association that exists between one continuous variable and one dichotomous variable. It is a special case of the Pearson's

product-moment correlation, which is applied when you have two continuous variables, whereas in this case one of the variables is measured on a dichotomous scale.

For example, use a Biserial correlation to determine whether there is an association between salaries, measured in US dollars, and gender (i.e., your continuous variable would be "salary" and your dichotomous variable would be "gender", which has two categories: "males" and "females"). Alternately Biserial correlation to determine whether there is an association between cholesterol concentration, measured in mmol/L, and smoking status (i.e., your continuous variable would be "cholesterol concentration", a marker of heart disease, and your dichotomous variable would be "smoking status", which has two categories: "smoker" and "non-smoker").

PARTIAL CORRELATION

Correlation involved two variables only. There are situation in which more than two variables are related with each other.

Definition

Partial correlation is defined as linear relation between two variables after partialling out the effect of other variables. In most of the situations an investigator wishes to find the extent of actual relation between two variables. This could be obtained by partialling out the effect of other variables. Mathematically it cloud be done by means of partial correlation.

Order of Partial Correlation

Order of partial correlation depends upon the number of variables whose effects have to be partial out. Product moment correlation coefficient is known as zero order correlation as in this case none of the variable's effect is eliminated. For each additional variable in the correlation the order is increased accordingly. Thus a first order correlation has three variables and a second order correlation has four variables. Further first order correlation is represented by $r_{12..3}$, Here correlation is obtained between first and second variable after eliminating the effect of third variable. Similarly second order partial correlation is represented by $r_{12..34}$ and in the same manner partial correlation of order n-2 is given by

$r_{12..34567.....n}$

Computation of First Order Partial Correlation

In first order partial correlation three variables are involved, out of which one variable is held constant. Let us consider that the variables are represented by 1.2 and 3 and if the third variable is held constant. First order partial correlation is given by

$$r_{12..3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Similarly $r_{23.1}$ and $r_{13.2}$ represent the first order partial correlation where first and second variables have been partialled out respectively.

Thus

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{1 - r_{21}^2}\sqrt{1 - r_{31}^2}} \text{ and } \quad r_{13.2} = \frac{r_{13} - r_{21}r_{32}}{\sqrt{1 - r_{21}^2}\sqrt{1 - r_{32}^2}}$$

Example

In an experiment conducted on fifteen school volleyball players, playing ability, height and arm length were measured. Product moment correlation among variables are shown in the compute $r_{12.3}$ and $r_{13.2}$

Correlation matrix

	x_1	x_2	x_3
x_1	1.00	0.67	0.75
x_2		1.00	0.94
x_3			1.00

x_1 : Playing ability x_2 : Height x_3 : Arm length

Solution :

(i) Since $r_{12..3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$

$$r_{12..3} = \frac{0.67 - 0.75X0.94}{\sqrt{1-0.75^2}\sqrt{1-0.94^2}}$$

= -0.15

(ii) $r_{13.2} = \frac{r_{13} - r_{21}r_{32}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{32}^2}}$

$$r_{13.2} = \frac{0.75 - 067X0.94}{\sqrt{1-0.67^2}\sqrt{1-0.94^2}}$$

= 0.47

General formula for Partial Correlation

In case of n variables, partial correlation will be order n-2 and would be represented by $r_{12.345\ldots\ldots n}$ A General formula is given by

$$r_{12..345\ldots\ldots n} = \frac{r_{12.34\ldots\ldots(n-1)} - r_{1n.34\ldots\ldots(n-1)}r_{2n.34\ldots\ldots(n-1)}}{\sqrt{1-r^2{}_{1n.34\ldots\ldots(n-1)}}\sqrt{1-r^2{}_{2n.34\ldots\ldots(n-1)}}}$$

Remark Like product moment correlation coefficient, partial correlation also lies is between -1 to + 1.

Limitations of Partial Correlation

Like product moment correlation coefficient, partial correlation too measure linear relation only. Thus when we refer true relation between any two variables after partialling out the effects of other variables

by means of partial correlation, we mean the linear relation only and it does not explain other type of relationship,

Even after partialling out the effect of variables in partial correlation we may not be sure that it measures the true linear relation between the two variables as there may be many other causative factors which might be affecting the correlation and were not considered in partialling out the effects.

Thus while partialling out the effect of variables one must explore the possibilities of all such variables which might be affecting the correlation coefficient. For instance partial correlation between height and weight after parialling out the effect of age may not explain the true linear relation between the height and weight of an individual as there may be other factors like height of one's parent, socio economic status etc which might be affecting the relation between height and weight of an individual, whose effects have not been partialled out.

Further large sample must be taken to calculate the partial correlation. This will give more reliable assessment about the true relationship between the two variables.

Utilities of Partial Correlation

Partial correlation has a special significance in research. We may be interested in knowing the fact that what are all parameters which affect the variables? Naturally there may be many parameters By computing zero order correlation coefficient if endurance is found to be highly associated with the variable

Multiple Correlation

If performance on 100 meter distance is to be estimated on the basis of few parameters like reaction time, acceleration speed maintenance level and deacceleration phase, it is essential to know how this group of parameters is related with it . Multiple correlation is used as a yard stick in this regard

Multiple correlation could be defined as the correlation between a group of variables and a single variable not included in that group. Multiple correlation is essentially a correlation between dependent variable and its expected values. These expected values are obtained through estimation by means of regression equation using the group of independent variables.

Order of Multiple Correlation

Multiple correlation is represented by the symbol R along with subscripts like R_{123} These subscripts represent the number of variables involved. The first subscript represents the dependent variable while subsequent subscripts represent independent variables. The first two subscript represent the zero order correlation and other orders start after second subscript. If n is the total number of variables in the multiple correlation., its order would be (n-2) .Thus first order multiple correlation is $R_{1.23}$ whereas $R_{1.234}$ denotes second order.

First order multiple correlation $R_{1.23}$ is computed by the following formula

$$R_{1.23} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} \text{ (or) } R_{1.23} = \sqrt{1 - \left[\left(1 - r^2{}_{12}\right)\left(1 - r^2{}_{13.2}\right)\right]}$$

Limits of Multiple Correlation

Limits of multiple correlation are from 0 to +1. Since the expression inside the square root should be positive in order to get the value of square root to be real, implies that $R_{1.23}$ can not be negative.Thus $0 \le R \le 1$.

Remarks

1. Multiple correlation is high if the correlation between independent variables are low.
2. A multiple correlation will not be less than the highest zero order correlation with the dependent variable.
3. Since multiple correlation is computed with product moment correlations, hence it also measures only linear relation.

Use of Multiple Correlation

As per the definition multiple correlation is a correlation between dependent variable and its expected values obtained by estimating from independent variables. Higher multiple correlation indicates that more accurate estimation of dependent variable is possible from independent variables. Thus multiple correlation provides an index of efficiency in estimation procedure.

The above mentioned concept is very useful in research. Multiple correlation is helpful in selecting the most valid battery of test for forecasting a criterion. Thus this method may be used to show the importance of height, weight and shoulder strength in estimating the performance in shot put.

If $R_{1.234}$ is the multiple correlation, where shot put performance is dependent variable and height, weight and shoulder strength are independent variables, then efficiency of estimating the shot put performance on the basis of these three variables.