Correlation

The tendency of simultaneous variation between two variables is called correlation or covariation.

In statistics, correlation is a method of determining the correspondence or proportionality between two series of measures (or scores).

For example, there may exist a relationship between heights and weights of a group of students

the scores of students in two different subjects are expected to have an interdependence or relationship between them.

**Types of Correlation:**

**In a bivariate distribution, the correlation may be:**

1. Positive, Negative and Zero Correlation; and

2. Linear or Curvilinear (Non-linear).

**1. Positive, Negative or Zero Correlation:**

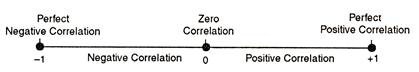
When the increase in one variable (X) is followed by a corresponding increase in the other variable (Y); the correlation is said to be positive correlation. The positive correlations range from 0 to +1; the upper limit i.e. +1 is the perfect positive coefficient of correlation.

The perfect positive correlation specifies that, for every unit increase in one variable, there is proportional increase in the other. For example “Heat” and “Temperature” have a perfect positive correlation.

If, on the other hand, the increase in one variable (X) results in a corresponding decrease in the other variable (Y), the correlation is said to be negative correlation.

The negative correlation ranges from 0 to – 1; the lower limit giving the perfect negative correlation. The perfect negative correlation indicates that for every unit increase in one variable, there is proportional unit decrease in the other.

Zero correlation means no relationship between the two variables X and Y; i.e. the change in one variable (X) is not associated with the change in the other variable (Y). For example, body weight and intelligence, shoe size and monthly salary; etc. The zero correlation is the mid-point of the range – 1 to + 1.

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image004-18.jpg)**

**2. Linear or Curvilinear Correlation:**

Linear correlation is the ratio of change between the two variables either in the same direction or opposite direction and the graphical representation of the one variable with respect to other variable is straight line.

Consider another situation. First, with increase of one variable, the second variable increases proportionately upto some point; after that with an increase in the first variable the second variable starts decreasing.

The graphical representation of the two variables will be a curved line. Such a relationship between the two variables is termed as the curvilinear correlation.

**Methods of Computing Co-Efficient of Correlation:**

**In ease of ungrouped data of bivariate distribution, the following three methods are used to compute the value of co-efficient of correlation:**

1. Scatter diagram method.

2. Pearson’s Product Moment Co-efficient of Correlation.

3. Spearman’s Rank Order Co-efficient of Correlation.

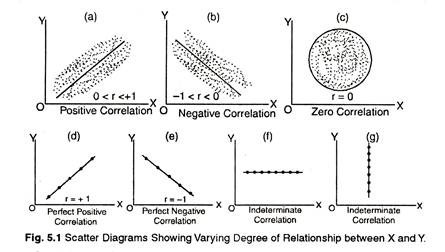
**1. Scatter Diagram Method:**

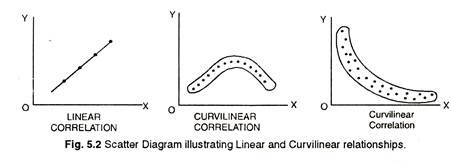
Scatter diagram or dot diagram is a graphic device for drawing certain conclusions about the correlation between two variables.

In preparing a scatter diagram, the observed pairs of observations are plotted by dots on a graph paper in a two dimensional space by taking the measurements on variable X along the horizontal axis and that on variable Y along the vertical axis.

The placement of these dots on the graph reveals the change in the variable as to whether they change in the same or in the opposite directions. It is a very easy, simple but rough method of computing correlation.

The frequencies or points are plotted on a graph by taking convenient scales for the two series. The plotted points will tend to concentrate in a band of greater or smaller width according to its degree. ‘The line of best fit’ is drawn with a free hand and its direction indicates the nature of correlation. Scatter diagrams, as an example, showing various degrees of correlation are shown in Fig. 5.1 and Fig. 5.2.

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image006-12.jpg)**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image008-11.jpg)**

If the line goes upward and this upward movement is from left to right it will show positive correlation. Similarly, if the lines move downward and its direction is from left to right, it will show negative correlation.

The degree of slope will indicate the degree of correlation. If the plotted points are scattered widely it will show absence of correlation. This method simply describes the ‘fact’ that correlation is positive or negative.

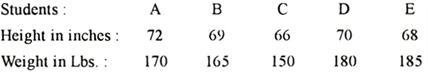
**2. Pearson’s Product Moment Co-efficient of Correlation:**

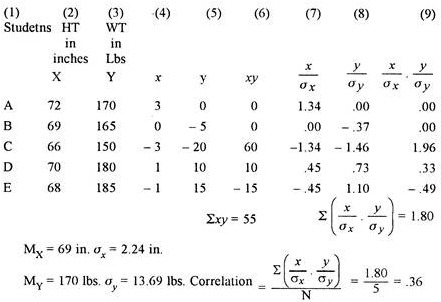
The coefficient of correlation, r, is often called the “Pearson r” after Professor Karl Pearson who developed the product-moment method, following the earlier work of Gallon and Bravais.

**Coefficient of correlation as ratio:**

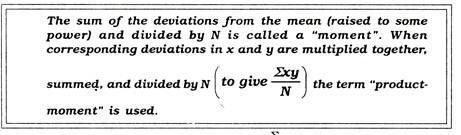
The product-moment coefficient of correlation may be thought of essentially as that ratio which expresses the extent to which changes in one variable are accompanied by—or dependent upon-changes in a second variable.

**As an illustration, consider the following simple example which gives the paired heights and weights of five college students:**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image010.png)**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image012-9.jpg)**

The mean height is 69 inches, the mean weight 170 pounds, and the o is 2.24 inches and o is 13.69 pounds, respectively. In the column (4) the deviation (x) of each student’s height from the mean height, and in column (5) the deviation, (y) of each student’s weight from the mean weight are given. The product of these paired deviations (xy) in column (6) is a measure of the agreement between individual heights and weights. The larger the sum of xy column the higher the degree of correspondence. In above example the value of ∑xy/N is 55/5 or 11. Where perfect agreement, i.e. r = ± 1.00, the value of ∑*xy*/N exceeds maximum limit.

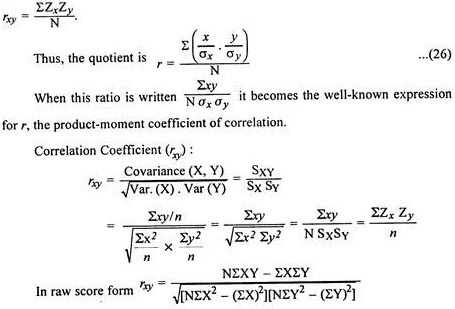
**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image014-6.jpg)**

Thus, ∑*xy*/N would not yield a suitable measure of relationship between x and y. The reason is that such an average is not a stable measure, as it is not independent of the units in which height and weight have been expressed.

In consequence, this ratio will vary if centimeters and kilograms are employed instead of inches and pounds. One way to avoid the trouble-some matter of differences in units is to express each deviation as a σ score or standard score or Z score, i.e. to divide each x and y by its own σ.

Each x and y deviation is then expressed as a ratio, and is a pure number, independent of the test units. The sum of the products of the σ scores column (9) divided by N yields a ratio which is a stable expression of relationship. This ratio is the “product-moment” coefficient of correlation. In our example, its value of .36 indicates a fairly high positive correlation between height and weight in this small sample.

The student should note that our ratio or coefficient is simply the average product of the σ scores of corresponding X and Y measures i.e.

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image016-4.jpg)**

**Nature of rxy:**

(i) rxy is a product moment r **[http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image018_thumb-3.jpg](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image018-3.jpg)**

(ii) rxy is a ratio, = rxy.

(iii) rxy can be + ve or – ve bound by limits – 1.00 to + 1.00.

(iv) rxy may be regarded as an arithmetic mean (rxy is the mean of standard score products).

(v) rxy is not affected by any linear transformation of scores on either X or Y or both.

(vi) When variables are in the standard score form, r gives a measure of the average amount of change in one variable associated with the change of one unit the other variable.

(vii) rxy = √byx bxy where byx = regression coefficient of Y on X, bxy = regression coefficient of X on Y. rxy = square root of the slopes of the regression lines.

(viii)rxy is not influenced by the magnitude of means (scores are always relative).

(ix)rxy cannot be computed if one of the variables has no variance S2x or S2Y = 0

(x) rxy of 60 implies the same magnitude of relationship as rxy = – .60. The sign tells about the direction of relationship, and the magnitude about the strength of the relationship.

(xi) df for rxy is N – 2, which is used for testing significance of rxy. Testing significance of r is testing significance of regression. Regression line involves slope and intercept, hence 2 *df* is lost. So when N = 2, rxyis either + 1.00 or – 1.00 as there is no freedom for sampling variation in the numerical value of r.

**A. Computation of rxy (Ungrouped Data)**:

Here, using the formula for computation of r depends on “where from the deviations are taken”. In different situations deviations can be taken either from actual mean or from zero or from A.M. Type of Formula conveniently applied for the calculation of coefficient correlation depends upon mean value (either in fraction or whole).

(i) The Formula of r when Deviations are taken from Means of the Two Distributions X and Y**.**

**[http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image020_thumb-3.jpg](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image020-3.jpg)**

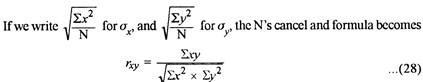
where rxy = Correlation between X and Y

x = deviation of any X score from the mean in the test X

y = deviation of corresponding Y score from the mean in test Y.

∑xy = Sum of all the products of deviations (X and Y)

σx and σy = Standard deviations of the distribution of X and Y score.

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image022-2.jpg)**

in which x and y are deviations from the actual means and ∑x2 and ∑y2are the sums of squared deviations in x and y taken from the two means.

**This formula is preferred:**

i. When mean values of both the variables are not in fraction.

ii. When to find out correlation between short, ungrouped series (say, twenty- five cases or so).

iii. When deviations are to be taken from actual means of the two distributions.

**The steps necessary are illustrated in Table 5.1. They are enumerated here:**

**Step 1:**

List in parallel columns the paired X and Y scores, making sure that corresponding scores are together.

**Step 2:**

Determine the two means Mx and My. In table 5.1, these are 7.5 and 8.0, respectively.

**Step 3:**

Determine for every pair of scores the two deviations x and y. Check them by finding algebraic sums, which should be zero.

**Step 4:**

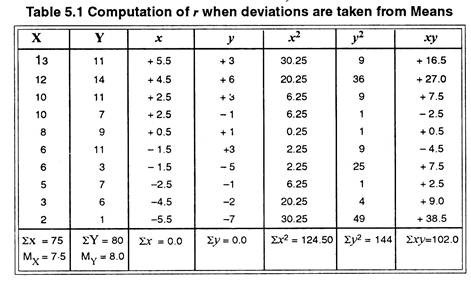
Square all the deviations, and list in two columns. This is for the purpose of computing σxand σy.

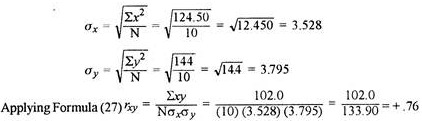
**Step 5:**

Sum the squares of the deviations to obtain ∑x2and ∑y2 Find xy product and sum these for ∑xy.

**Step 6:**

From these values compute σx and σy.

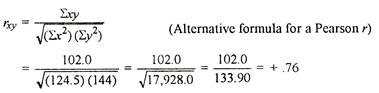
**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image024-3.jpg)**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image026-2.jpg)**

**An alternative and shorter solution:**

There is an alternative and shorter route that omits the computation of σx and σy, should they not be needed for any other purpose.

**Applying Formula (28):**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image028-2.jpg)**

**(ii) The Calculation of rxy from Original scores or Raw scores:**

It is an another procedure with ungrouped data, which does not require the use of deviations. It deals entirely with original scores. The formula may look forbidding but is really easy to apply.

**[http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image030_thumb-1.jpg](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image030-1.jpg)**

**This formula is preferred:**

i. When to compute r from direct raw scores.

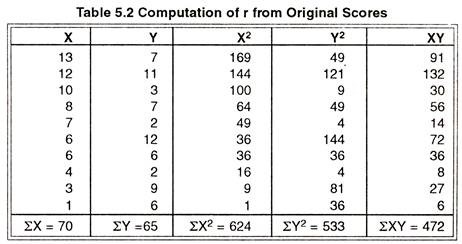
ii. Original scores ft. when data are small ungrouped.

iii. When mean values are in fractions.

iv. When good calculating machine is available.

X and Y are original scores in variables X and Y. Other symbols tell what is done with them.

**We follow the steps that are illustrated in Table 5.2:**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image034-1.jpg)**

**Step 1:**

Square all X and Y measurements.

**Step 2:**

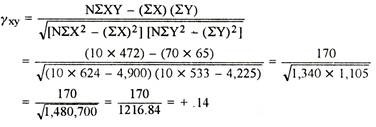
Find the XY product for every pair of scores.

**Step 3:**

Sum the X’s, the Y’s, the X2, the Y2, and the XY.

**Step 4:**

**Apply formula (29):**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image032-1.jpg)**

**(ii) Computation of rxy when deviations are taken from Assumed Mean:**

Formula (28) is useful in calculating r directly from two ungrouped series of scores, but it has the disadvantages as it requires “long method” of calculating means and *σ*’s. The deviations x and y when taken from actual means are usually decimals and the multiplication and squaring of these values is often a tedious task.

For this reason—even when working with short ungrouped series—it is often easier to assume means, calculate deviations from these A.M.’s and apply the formula (30).

**[http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image036_thumb-1.jpg](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image036-1.jpg)**

**This formula is preferred:**

i. When actual means are usually decimals and the multiplication and squaring of these values is often a tedious task.

ii. When deviations are taken from A.M.’s.

iii. When we are to avoid fractions.

**The steps in computing r may be outlined as follows:**

**Step 1:**

Find the mean of Test 1 (X) and the mean of Test 2 (Y). The means as shown in Table 5.3 MX= 62.5 and MY = 30.4 respectively.

**Step 2:**

Choose A.M.’s of both X and Y i.e. A.M.X as 60.0 and A.M.Y as 30.0.

**Step 3:**

Find the deviation of each score on Test 1 from its A.M., 60.0, and enter it in column x’. Next find the deviation of each score in Test 2 from its A.M., 30.0, and enter it in column y’.

**Step 4:**

Square all of the x’ and all of they’ and enter these squares in column x’2 and y’2, respectively. Total these columns to obtain ∑x’2 and ∑y’2.

**Step 5:**

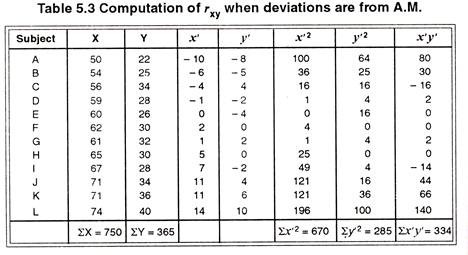
Multiply x’ and y’, and enter these products (with due regard for sign) in the x’y’ column. Total x’y’ column, taking account of signs, to get ∑x’y’.

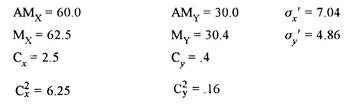
**Step 6:**

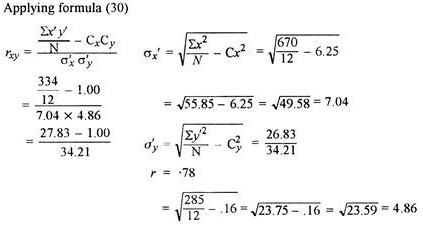
The corrections, Cx and Cy, are found by subtracting AMX from Mxand AMy from My. Then, Cx found as 2.5 (62.5 – 60.0) and Cy as .4 (30.4 – 30.0).

**Step 7:**

Substitute for ∑x’y’ , 334, for ∑x’2, 670 and for ∑y’2, 285 in formula (30), as shown in Table 5.3, and solve for rxy.

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image038.jpg)**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image040-1.jpg)**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image042-1.jpg)**

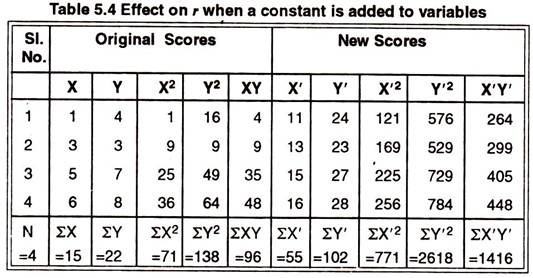
**Properties of*r*:**

**1. The value of the coefficient of correlation r remains unchanged when a constant is added to one or both variables:**

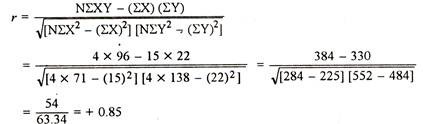
In order to observe the effect on the coefficient correlation r when a constant is added to one or both the variables, we consider an example.

Now, we add a score of 10 to each score in X and 20 to each score of Y and represent these scores by X’ and Y’ respectively.

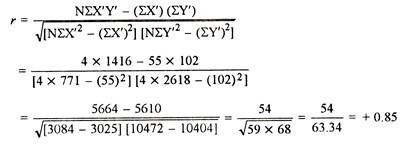
**The calculations for computing r for original and new pairs of observations are given in Table 5.4:**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image044-2.jpg)**

**By using formula (29), the coefficient of correlation of original score will be:**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image046-1.jpg)**

**The same formula for new scores can be written as:**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image048-1.jpg)**

Thus, we observe that the value of the coefficient of correlation r remains unchanged when a constant is added to one or both variables.

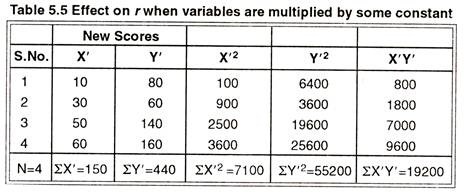
**2. The value of the coefficient of correlation r remains unchanged when a constant is subtracted from one or both variables:**

Students can examine this by taking an example. When each score of one or both variables are subtracted by a constant the value of coefficient of correlation r also remains unchanged.

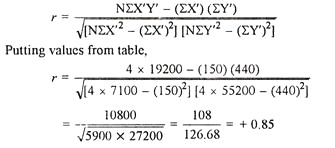
**3. The value of the coefficient of correlation r remains unaltered when one or both sets of variate values are multiplied by some constant:**

In order to observe the effect of multiplying the variables by some constant on the value of r, we arbitrarily multiply that original scores of first and second sets in the previous example by 10 and 20 respectively.

**The r between X’ and Y’ may then be calculated as under:**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image050-1.jpg)**

**The correlation of coefficient between X’ and Y’ will be:**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image052-2.jpg)**

Thus, we observe that the value of the coefficient of correlation r remains unchanged when a constant is multiplied with one or both sets of variate values.

**4. The value of r will remain unchanged even when one or both sets of variate values are divided by some constant:**

Students can examine this by taking an example.

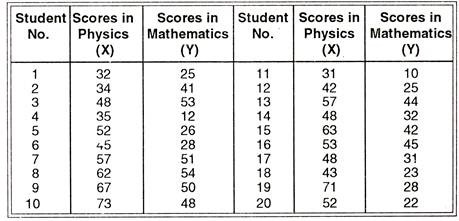
**B. Coefficient of Correlation in Grouped Data**:

When the number of pairs of measurements (N) on two variables X and Y are large, even moderate in size, and when no calculating machine is available, the customary procedure is to group data in both X and Y and to form a scatter diagram or correlation diagram which is also called two-way frequency distribution or bivariate frequency distribution.

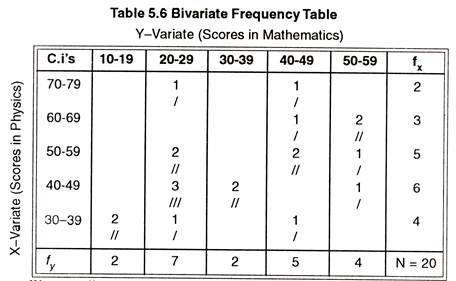
The choice of size of class interval and limits of intervals follows much the same rules as were given previously. To clarify the idea, we consider a bivariate data concerned with the scores earned by a class of 20 students in Physics and Mathematics examination.

**Preparing a Scatter diagram:**

In setting up a double grouping of data, a table is prepared with columns and rows. Here, we classify each pair of variates simultaneously in the two classes, one representing score in Physics (X) and the other in Mathematics (Y) as shown in Table 5.6.

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image054-1.jpg)**

**The scores of 20 students in both Physics (X) and Mathematics (Y) are shown in Table below:**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image056-1.jpg)**

We can easily prepare a bivariate frequency distribution table by putting tallies for each pair of scores. The construction of a scattergram is quite simple. We have to prepare a table as shown in the diagram above.

Along the left hand margin the class intervals of X-distribution are laid off from bottom to top (in ascending order). Along the top of the diagram the c.i’s of Y-distribution are laid off from left to right (in ascending order).

Each pair of scores (both in X and Y) is represented through a tally in the respective cell. No. 1 student has secured 32 in Physics (X) and 25 in Mathematics (Y). His score of 32 in (X) places him in the last row and 25 in (Y) places him in the second column. So, for the pair of scores (32, 25) a tally will be marked in the second column of 5th row.

In a similar way, in case of No. 2 student, for scores (34, 41), we shall put a tally in the 4th column of the 5th row. Likewise, 20 tallies will be put in the respective rows and columns. (The rows will represent the X-scores and the columns will represent the Y-scores).

Along the right-hand margin the *fx* column, the number of cases in each c.i., of X-distribution are tabulated and along the bottom of the diagram in the*fy* row the number of cases in each c.i., of Y-distribution are tabulated.

The total of *f*xcolumn is 20 and the total of *fy* row is also 20. It is in fact a bi-variate distribution because it represents the joint distribution of two variables. The scattergram is then a “correlation table.”

**Calculation of r from a correlation table:**

**The following outline of the steps to be followed in calculating r will be best understood if the student will constantly refer to Table 5.7 as he reads through each step:**

**Step 1:**

Construct a scattergram for the two variables to be correlated, and from it draw up a correlation table.

**Step 2:**

Count the frequencies of each c.i. of distribution – X and write it in the *fx* column. Count the frequencies for each c.i. of distribution – Y and fill up the *fy* row.

**Step 3:**

Assume a mean for the X-distribution and mark off the c.i. in double lines. In the given correlation table, let us assume the mean at the c.i., 40 – 49 and put double lines as shown in the table. The deviations above the line of A.M. will be (+ ve) and the deviations below it will be (- ve).

The deviation against the line of A.M., i.e., against the c.i. where we assumed the mean is marked 0 (zero) and above it the *d*‘s are noted as +1, +2. 13 and below it*d* is noted to be – 1. Now dx column is filled up. Then multiply *fx*. and *dx*of each row to get *fdx*. Multiply *dx* and *fdx* of each row to get *fdx*2.

[Note: While computing the S.D. in the assumed mean method we were assuming a mean, marking the d’s and computing *fd*and *fd*2. Here also same procedure is followed.]

**Step 4:**

Adopt the same procedure as in step 3 and compute *dy*, *fdy*and *fdy*2. For the distribution-Y, let us assume the mean in the c.i. 20-29 and put double lines to mark off the column as shown in the table. The deviations to the left of this column will be negative and right be positive.

Thus, d for the column where mean is assumed is marked 0 (zero) and the d to its left is marked – 1 and *d’*s to its right are marked +1, +2 and +3. Now *dy* column is filled up. Multiply the values of *fy* and *dy* of each column to get *fdy*. Multiply the values of *dy*and *fdy* to each column to get *fdy*2.

**Step 5:**

As this phase is an important one, we are to mark carefully for the computation of *dy* for different c.i.’s of distribution X and *dx* for different c.i.’s of distribution -Y.

*dy*for different c. i. ‘s of distribution-X: In the first row, 1*f* is under the column, 20-29 whose *dy*is 0 (Look to the bottom. The *dy* entry of this row is 0). Again 1*f* is under the column, 40- 49 whose *dy* is + 2. So *dy* for the first row = (1 x 0) + (1 x 2) = + 2.

**In the second row we find that:**

1 *f*is under the column, 40-49 whose *dy* is + 2 and

2*f*s are under the column, 50-59 whose *dy*’s are + 3 each.

So *dy*for 2nd row = (1 x 2) + (2 X 3) = 8.

In the third row,

2*f*s are under the column, 20-29 whose *dy*‘s are 0 each,

2*f*s are under the column, 40-49 whose *dy*‘s are +2 each, and 1 *f* is under the column, 50-59 whose *dy* is +3.

So dy for the 3rd row = (2 x 0) + (2 x 2) + (1 X 3) = 7.

In the 4th row,

3*f*s are under the column, 20-29 whose *dy*‘s are 0 each,

2*f*s are under the column, 30-39 whose *dy*‘s are +1 each, and 1*f* is under the column, 50-59 whose *dy*is + 3,

So *dy* for the 4th row = (3 X 0) + (2 X 1) + (1 x 3) = 5.

Likewise in the 5th row

*dy*for the 5th row = (2 x – 1) + (1 x 0) + (1 x 2) = 0

*dx*for different c.i. ,’v of distribution – Y :

In the first column,

2*f*s are against the row, 30-39 whose *dx* is – 1.

So *dx* of the 1st column = (2 x – 1) = – 2

In the second column,

1*f* is against the c.i., 70-79 whose *dx*is +3,

2*f*s are against the c.i., 50-59 whose *dx*‘s are +1 each,

3*f*s are against the c.i., 40-49 whose *dx*‘s are 0 each,

1 *f* is against the c.i., 30-39 whose *dx* is – 1.

So *dx* for the 2nd column = (1 x 3) + (2 X 1) + (3 X 0) + (1 x – 1) = 4. In the third column,

*dx* for the 3rd column = 2×0 = 0

In the fourth column,

*dx* for the 4th column = (1 x 3) + (1 x 2) + (2 x 1) + (1 x – 1) = 6.

In the fifth column,

*dx* for the 5th column = (2 x 2) + (1 x 1) + (1 X 0) = 5.

**Step 6:**

Now, calculate *dx.dy* each row of distribution – X by multiplying the *dx* entries of each row by *dy*entries of each row. Then calculate *dx.dy* for each column of distribution – Y by multiplying *dy*entries of each column by the *dx*entries of each column.

**Step 7:**

Now, take the algebraic sum of the values of the columns fdx, *fdx*2, *dy* and *dx.dy* (for distribution – X). Take the algebraic sum of the values of the rows *fdy,* *fdy*2, *dx*and *dx.dy* (for distribution – Y)

**Step 8:**

∑.*dx.dy* of X-distribution = ∑*dx.dy* of Y-distribution

∑*fdx* = total of *dx* row (i.e. ∑*dx*)

∑*fdy* = total of *dy* column (i.e. ∑*dy*)

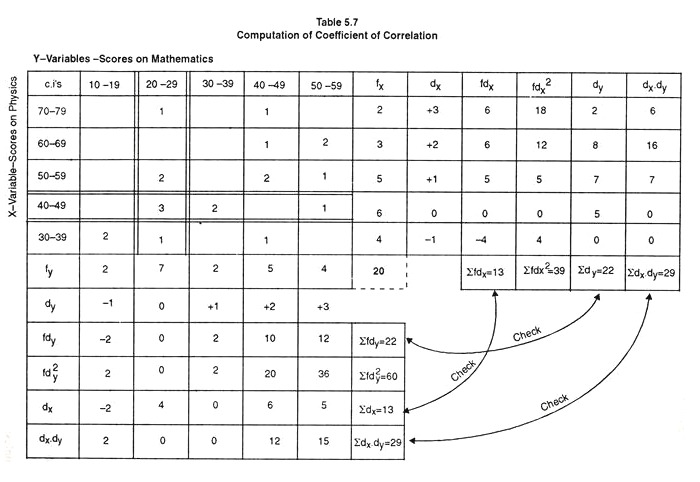
**Step 9:**

The values of the symbols as found

∑*fdx* = 13, ∑*fd*2x = 39

∑*fdy* =22, ∑*fd*2*y* = 60

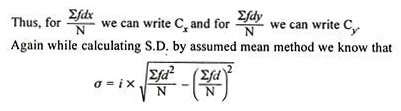
∑*dx.dy* = 29 and N = 20.

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image058-1.jpg)**

**In order to compute coefficient of correlation in a correlation table following formula can be applied:**

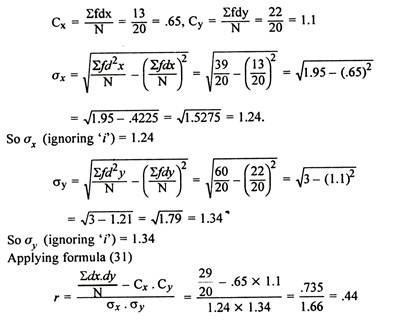
**[http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image060_thumb-1.jpg](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image060-1.jpg)**

**[http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image060_thumb-2.jpg](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image060-2.jpg)**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image060-3.jpg)**

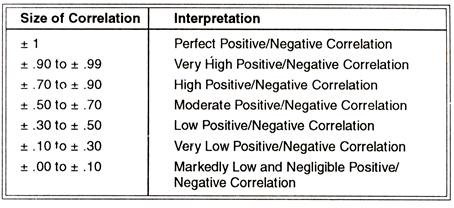
We may mark that in the denominator of formula (31) we apply the formula for ax and aywith the exception of no i’s. We may note here that Cx, Cy, σx, σv are all expressed in units of class intervals (i.e., in unit of i). Thus, while computing σxand σy, no i’s are used. This is desirable because all the product deviations i.e., ∑*dx.dy’*s are in interval units.

**Thus, we compute:**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image062-1.jpg)**

**Interpretation of the Coefficient of Correlation:**

Merely computation of correlation does not have any significance until and unless we determine how large must the coefficient be in order to be significant, and what does correlation tell us about the data? What do we mean by the obtained value of coefficient of correlation?

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image064-1.jpg)**

**Misinterpretation of the Coefficient of Correlation:**

Sometimes, we misinterpret the value of coefficient of correlation and establish the cause and effect relationship, i.e. one variable causing the variation in the other variable. Actually we cannot interpret in this way unless we have sound logical base.

Correlation coefficient gives us, a quantitative determination of the degree of relationship between two variables X and Y, not information as to the nature of association between the two variables. Causation implies an invariable sequence— A always leads to B, whereas correlation is simply a measure of mutual association between two variables.

**For example there may be a high correlation between maladjustment and anxiety:**

But on the basis of high correlation we cannot say maladjustment causes anxiety. It may be possible that high anxiety is the cause of maladjustment. This shows that maladjustment and anxiety are mutually associated variables. Consider another example.

There is a high correlation between aptitude in a subject at school and the achievement in the subject. At the end of the school examinations will this reflect causal relationship? It may or may not.

Aptitude in the study of subject definitely causes variation in the achievement of the subject, but high achievement of the student in the subject is not the result of the high aptitude only; it may be due to the other variables also.

Thus, when interpreting the size of the correlation co-efficient in terms of cause and effect it is appropriate, if and only if the variables under investigation provide a logical base for such interpretation.

**Factors influencing the size of the Correlation Coefficient:**

**We should also be aware of the following factors which influence the size of the coefficient of correlation and can lead to misinterpretation:**

1. The size of “r” is very much dependent upon the variability of measured values in the correlated sample. The greater the variability, the higher will be the correlation, everything else being equal.

2. The size of ‘r’ is altered, when an investigator selects an extreme group of subjects in order to compare these groups with respect to certain behavior. “r” obtained from the combined data of extreme groups would be larger than the “r” obtained from a random sample of the same group.

3. Addition or dropping the extreme cases from the group can lead to change on the size of “r”. Addition of the extreme case may increase the size of correlation, while dropping the extreme cases will lower the value of “r”.

**Uses of Product moment r:**

**Correlation is one of the most widely used analytic procedures in the field of Educational and Psychological Measurement and Evaluation. It is useful in:**

i. Describing the degree of correspondence (or relationship) between two variables.

ii. Prediction of one variable—the dependent variable on the basis of independent variable.

iii. Validating a test; e.g., a group intelligence test.

iv. Determining the degree of objectivity of a test.

v. Educational and vocational guidance and in decision-making.

vi. Determining the reliability and validity of the test.

vii. Determining the role of various correlates to a certain ability.

viii. Factor analysis technique for determining the factor loading of the underlying variables in human abilities.

**Assumptions of Product moment r**:

**1. Normal distribution:**

The variables from which we want to calculate the correlation should be normally distributed. The assumption can be laid from random sampling.

**2. Linearity:**

The product-moment correlation can be shown in straight line which is known as linear correlation.

**3. Continuous series:**

Measurement of variables on continuous series.

**4. Homoscedasticity:**

It must satisfy the condition of homoscedasticity (equal variability).

**3. Spearman’s Rank Correlation Coefficient:**

There are some situations in Education and Psychology where the objects or individuals may be ranked and arranged in order of merit or proficiency on two variables and when these 2 sets of ranks covary or have agreement between them, we measure the degrees of relationship by rank correlation.

Again, there are problems in which the relationship among the measurements made is non-linear, and cannot be described by the product-moment r.

For example, the evaluation of a group of students on the basis of leadership ability, the ordering of women in a beauty contest, students ranked in order of preference or the pictures may be ranked according to their aesthetic values. Employees may be rank-ordered by supervisors on job performance.

School children may be ranked by teachers on social adjustment. In such cases objects or individuals may be ranked and arranged in order of merit or proficiency on two variables. Spearman has developed a formula called Rank Correlation Coefficient to measure the extent or degree of correlation between 2 sets of ranks.

**This coefficient of correlation is denoted by Greek letter ρ (called Rho) and is given as:**

**[http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image066_thumb-1.jpg](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image066-1.jpg)**

where, ρ = rho = Spearman’s Rank Correlation Coefficient

D = Difference between paired ranks (in each case)

N = Total number of items/individuals ranked.

**Characteristics of Rho (ρ):**

1. In Rank Correlation Coefficient the observations or measurements of the bivariate variable is based on the ordinal scale in the form of ranks.

2. The size of the coefficient is directly affected by the size of the rank differences.

(a)If the ranks are the same for both tests, each rank difference will be zero and ultimately D2will be zero. This means that the correlation is perfect; i.e. 1.00.

(b)If the rank differences are very large, and the fraction is greater than one, then the correlation will be negative.

**Assumptions of Rho (ρ):**

i. N is small or the data are badly skewed.

ii. They are free, or independent, of some characteristics of the population distribution.

iii. In many situations Ranking methods are used, where quantitative measurements are not available.

iv. Though quantitative measurements are available, ranks are substituted to reduce arithmetical labour.

v. Such tests are described as non-parametric.

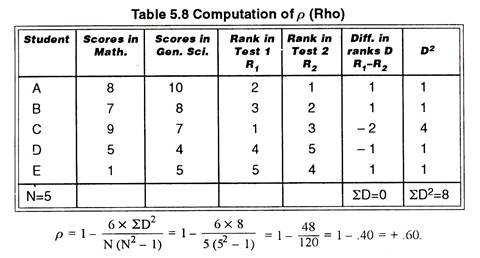
vi. In such cases the data are comprised of sets of ordinal numbers, 1st, 2nd, 3rd….Nth. These are replaced by the cardinal numbers 1, 2, 3,………, N for purposes of calculation. The substitution of cardinal numbers for ordinal numbers always assumes equality of intervals.

**I. Calculating ρ from Test Scores:**

**Example 1:**

**The following data give the scores of 5 students in Mathematics and General Science respectively:**

Compute the correlation between the two series of test scores by Rank Difference Method.

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image068-1.jpg)**

The value of coefficient of correlation between scores in Mathematics and General Science is positive and moderate.

**Steps of Calculation of Spearman’s Co-efficient of Correlation:**

**Step 1:**

List the students, names or their serial numbers in column 1.

**Step 2:**

In column 2 and 3 write scores of each student or individual in test I and II.

**Step 3:**

Take one set of score of column 2 and assign a rank of 1 to the highest score, which is 9, a rank of 2 to the next highest score which is 8 and so on, till the lowest score get a rank equal to N; which is 5.

**Step 4:**

Take the II set of scores of column 3, and assign the rank 1 to highest score. In the second set the highest score is 10; hence obtain rank 1. The next highest score of B student is 8; hence his rank is 2. The rank of student C is 3, the rank of E is 4, and the rank of D is 5.

**Step 5:**

Calculate the difference of ranks of each student (column 6).

**Step 6:**

Check the sum of the differences recorded in column 6. It is always zero.

**Step 7:**

Each difference of ranks of column 6 is squared and recorded in column 7. Get the sum ∑D2.

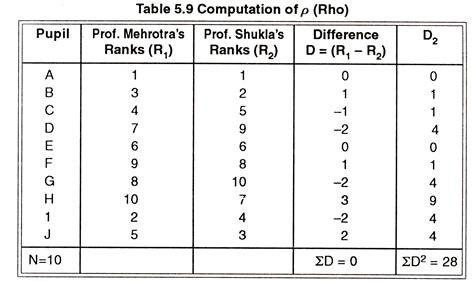
**Step 8:**

Put the value of N and 2D2 in the formula of Spearman’s co-efficient of correlation.

**2. Calculating from Ranked Data:**

**Example 2:**

In a speech contest Prof. Mehrotra and Prof. Shukla, judged 10 pupils. Their judgements were in ranks, which are presented below. Determine the extent to which their judgements were in agreement.

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image070-1.jpg)**

**[http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image072_thumb-1.jpg](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image072-1.jpg)**

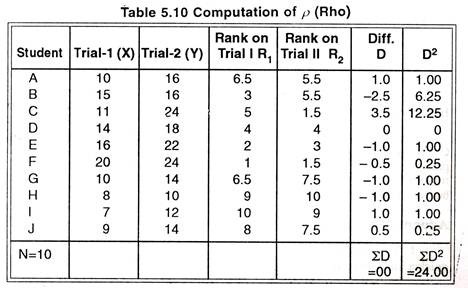
The value of co-efficient of correlation is + .83. This shows a high degree of agreement between the two judges.

**3. Calculating ρ (Rho) for tied Ranks:**

**Example 3:**

**The following data give the scores of 10 students on two trials of test with a gap of 2 weeks in Trial I and Trial II.**

**Compute the correlation between the scores of two trials by rank difference method:**

**[](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image074-1.jpg)**

**[http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image076_thumb-1.jpg](http://cdn.yourarticlelibrary.com/wp-content/uploads/2016/10/clip_image076-1.jpg)**

The correlation between Trial I and II is positive and very high. Look carefully at the scores obtained by the 10 students on Trial I and II of the test.

Do you find any special feature in the scores obtained by the 10 students? Probably, your answer will be “yes”.

In the above table in column 2 and 3 you will find that more than one students are getting the same scores. In column 2 students A and G are getting the same score viz. 10. In column 3, the students A and B, C and F and G and J are also getting the same scores, which are 16, 24 and 14 respectively.

Definitely these pairs will have the same ranks; known as Tied Ranks. The procedure of assigning the ranks to the repeated scores is somewhat different from the non-repeated scores.

Look at column 4. Student A and G have similar scores of 10 each and they possess 6th and 7th rank in the group. Instead of assigning the 6th and 7th rank, the average of the two rank i.e. 6.5 (6 + 7/2 = 13/2) has been assigned to each of them.

The same procedure has been followed in respect of scores on Trial II. In this case, ties occur at three places. Students C and F have the same score and hence obtain the average rank of (1 + 2/2 = 1.5). Student A and B have rank position 5 and 6; hence are assigned 5.5 (5 + 6/2) rank each. Similarly student G and J have been assigned 7.5 (7 + 8/2) rank each.

**If the values are repeated more than twice, the same procedure can be followed to assign the ranks:**

**For example:**

if three students get a score of 10, at 5th, 6th and 7th ranks, each one of them will be assigned a rank of 5 + 6 + 7/3= 6.

The rest of the steps of procedure followed for calculation of ρ (rho) are the same as explained earlier.

**Interpretation:**

The value of ρ can also be interpreted in the same way as Karl Pearson’s Coefficient of Correlation. It varies between -1 and + 1. The value + 1 stands for a perfect positive agreement or relationship between two sets of ranks while ρ = – 1 implies a perfect negative relationship. In case of no relationship or agreement between ranks, the value of ρ = 0.

**Advantages of Rank Difference Method:**

1. The Spearman’s Rank Order Coefficient of Correlation computation is quicker and easier than (r) computed by the Pearson’s Product Moment Method.

2. It is an acceptable method if data are available only in ordinal form or number of paired variable is more than 5 and not greater than 30 with minimum or a few ties in ranks.

3. It is quite easy to interpret p.

**Limitations:**

1. When the interval data are converted into rank-ordered data the information about the size of the score differences is lost; e.g. in the Table 5.10, if D in Trial II gets scores from 18 up to 21, his rank remains only 4.

2. If the number of cases are more, giving ranks to them becomes a tedious job.