CHAPTER 5

# DATA ANALYSIS: COMPUTING RELATIONSHIP BETWEEN VARIABLES

Measures of relationship are used when the interest is in the degree of relationship between pairs of two or more variables, or the extent to which scores on one test are associated with scores on another test. For example, a researcher might want to see the relationship between reading speed and comprehension scores of a specific group of students. To give another example, a study could be conducted to see the relationship between one teacher's rating of a set of compositions and another teacher's rating. In both cases, the researcher would like to identify whether a student scoring high on one measure also scores high on the other and whether a student who scores low on one measure also scores low on the other. The researcher is interested in the extent to which the two sets of scores tend to cover.

To represent two sets of scores graphically we use a scatterplot or scattergram to plot the scores. By convention the left axis is the independent variable and the right axis is the dependent variable. To illustrate the steps to measure this type of relationship, the scores (out of 20) of imaginary ten students from their reading and listening tests are listed below. When these scores are entered into a graphic program, the desired graph can be obtained (see Figure 5.1 for the graph and Chapter 4 for other graphic representations).

| STUDENTS | LISTENING SCORES | READING SCORES |
|---|---|---|
| S1 | 5 | 10 |
| S2 | 7 | 13 |
| S3 | 9 | 20 |
| S4 | 10 | 16 |
| S5 | 15 | 18 |
| S6 | 7 | 10 |
| S7 | 7 | 13 |
| S8 | 9 | 15 |
| S9 | 17 | 20 |
| S10 | 12 | 17 |

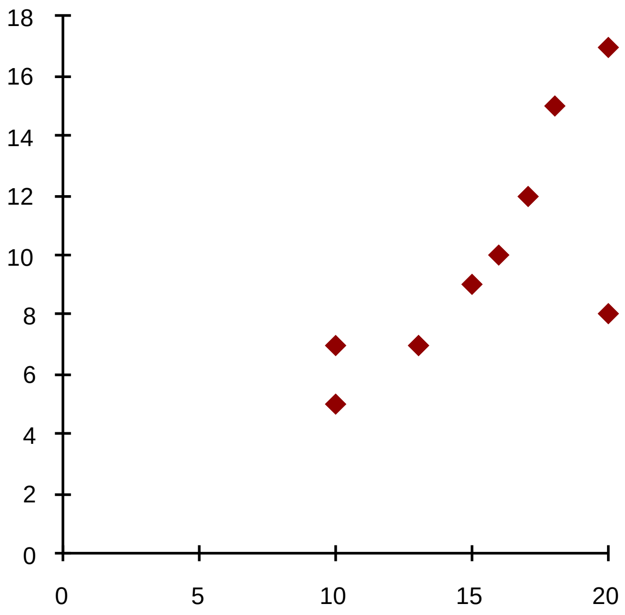

Figure 5.1 The scatterplot graph

The scatterplot graph in Figure 5.1 indicates the correlation pattern between the listening scores (Variable 1) and the reading scores (Variable 2) of the ten students. As indicated in the scatterplot graph, there is a positive correlation between these two variables.

There are three basic correlation patterns indicating the relationship between the two variables that are under study: 1) positive, 2) negative, or 3) no (zero) relationship (see Figures 5.1 -.5.4). In a positive or a negative relationship, the scores for two different variables fall along the same line. For that reason , these relations are called linear.

If the two sets of scores show a perfect linear relationship, the dots on a graph will fall on a straight line, which can be interpreted as a perfect positive (see Figure 5.2), or negative (see Figure 5.3) correlation depending on the incline of the dots. If there is no relationship, the dots will be scattered all over as in Figures 5.4.

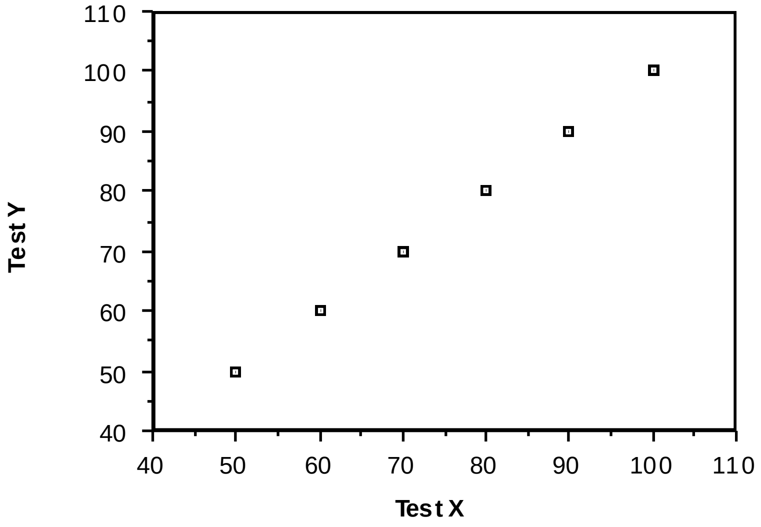

Figure 5.2 A scatterplot graph showing a positive relationship

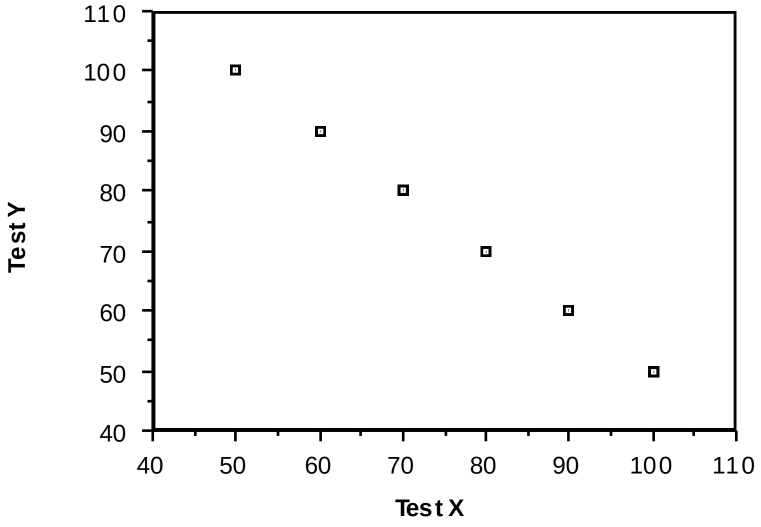

Figure 5.3 A scatterplot graph showing a negative relationship

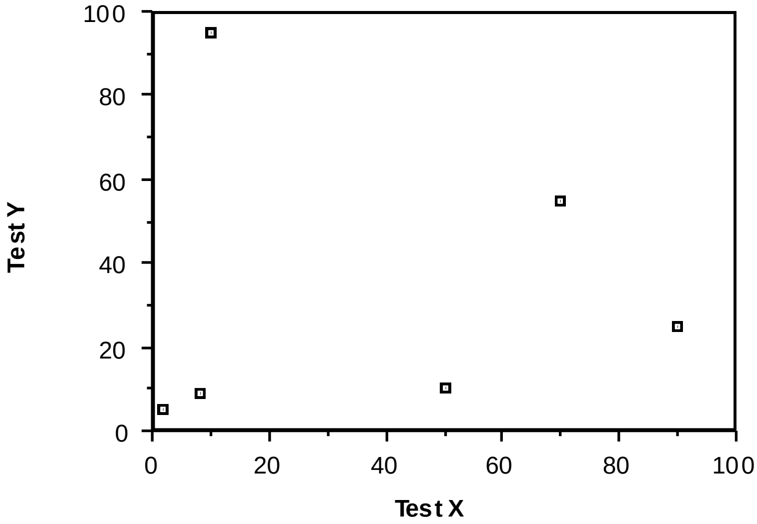

Figure 5.4 A scatterplot graph showing a zero relationship

We can also observe the degree of a positive or a negative correlation by looking at these graphs. If the scores fall very close to the positive line, the degree of association is considered high (see Figure 5.5). If the scores fall around the positive line, the association is considered moderate. If the scores spread apart from one another , then there is zero association among the scores. In other words, the scores do not fall around a line in any direction. The imaginary line around which the points cluster is called the slope.

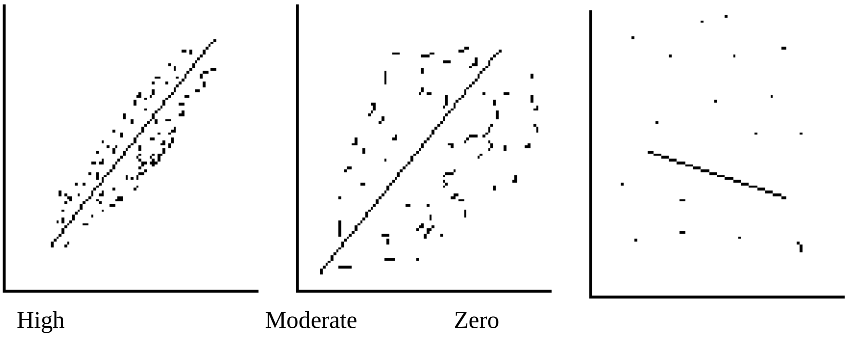

Figure 5.5 Scatterplot graphs indicating the degree of association between the two sets of scores

With large sampling, scatter plotting does not give us any quantitative measure of the degree of relationship between two variables. The statistic that describes the relationship between two variables is called a correlation coefficient, which indicates the degree of relationship between two variables (see Vol. 1, Chapter 5; Gay, 1987; Moore, 1983; Hatch & Lazaraton, 1991; Hatch & Farhady, 1982; Brown 1988). While discussing correlational research the values of the correlation coefficient are indicated in three terms:

| | |
|---|---|
| 1. For perfect positive relationship: | r= +1.00 |
| 2. For perfect negative relationship: | r= -1.00 |
| 3. For perfect zero relationships: | r= 0.0 |

If scores (X) on one distribution increase or decrease based on the degree of increase or decrease of another set of scores (Y), the relationship between these scores is considered to be positive. This is because, either the tests applied are considered to be measuring the same type of behavior, or the two variables that are measured are very related. If the value of r is closer to +1.00, the positive relationship is estimated to be stronger. For instance, let us suppose that our variables were *level of IQ of children* and *their language aptitude*. After administering two different kinds of tests to subjects of the same age and educational background, if we see a positive relationship between these scores, we make a generalization that language aptitude is highly correlated with the level of IQ of the subject. Thus, subjects of high IQs are expected to learn a second language much better.

If the relation proves to be working conversely, the relationship is said to be negative. If the value is closer to -1.00, the negative relationship is stronger. Then we can say that children with high IQs have low language aptitude or vice versa.

If the scores are around 0.0, this indicates that there is a perfect zero relationship between the two measures. In this case, we say that there is no correlation between the level of IQ of a subject and his/her language aptitude.

## Pearson Product-moment<br>Correlation Coefficient (r)

Pearson Product-moment Correlation Coefficient is a statistical device used in calculating the relationship between the two measures. The letter r is the symbol for the correlation coefficient and can be defined as the mean cross product of the z-scores.

## Computation of (r)

In calculating correlation coefficient, the researcher starts with the following assumptions:

- The two variables are continuous.
- Scores on X and Y are independent of one another.
- The relationship between X an Y is linear.

The measure or the size of the correlation coefficient indicates how well two sets of scores go together. There is no cause and effect relationship between the variables. The correlation coefficient is used only to show the degree of relationship.

The correlation coefficient can be computed using one of the following:

1) z-scores
2) raw scores
3) covariance

Computing r using z-scores

1. We first convert the two sets of scores (X scores and Y scores) into z-scores (see also Chapter 3). For instance, if we want to calculate the correlation coefficient related to students' Reading and Listening scores, we convert the raw scores to z-scores by subtracting the mean from each score and then dividing it by the total number of students who take the test.

| | READING SCORES | | LISTENING SCORES | |
|---|---|---|---|---|
| Stds. | Raw | z-scores | Raw | z-scores |
| S1 | 52 | - 1.8 | 45 | - 2.0 |
| S2 | 60 | - 1.0 | 50 | - 1.5 |
| S3 | 67 | - .3 | 55 | - 1.5 |
| S4 | 69 | - .1 | 60 | - .5 |
| S5 | 69 | - .1 | 66 | + .1 |
| S6 | 71 | + .1 | 70 | + .5 |
| S7 | 73 | + .3 | 70 | + .5 |
| S8 | 76 | + .6 | 75 | + 1.0 |
| S9 | 80 | + 1.0 | 79 | + 1.4 |
| S10 | 83 | + 1.3 | 80 | + 1.5 |

2. In the next step, we obtain the product of these z-scores. Product is obtained by multiplying each S's pair of z-scores.

| Sts. | READING z-scores | LISTENING z-scores | PRODUCT |
|---|---|---|---|
| S1 | - 1.8 | - 2.0 | 3.60 |
| S2 | - 1.0 | - 1.5 | 1.50 |
| S3 | - .3 | - 1.5 | .45 |
| S4 | - .1 | - .5 | .05 |
| S5 | - .1 | + .1 | .01 |
| S6 | + .1 | + .5 | .05 |
| S7 | + .3 | + .5 | .15 |
| S8 | + .6 | + 1.0 | .60 |
| S9 | + 1.0 | + 1.4 | 1.40 |
| S10 | + 1.3 | + 1.5 | 1.95 |

3. Finally, if we add these cross-products and divide the total by the number of pairs, we get a correlation coefficient of +1.0.

9.76 divided by 10 = .976

Computing r from raw scores

It might be easier to demonstrate the computation of r with an example. Suppose there are two raters scoring the speaking skill of the students. Each rater scores what they hear directly from the students or from the tape recordings with a possible score range of 0 to 20. If we want to find out how well the two raters agree regarding the students' speaking skills, we pursue the following steps and indicate the results as in Table 5.1 below.

1. List the two different scores
   (X= first rater's score, Y= second rater's score)
   for each speaker in parallel columns.
2. Square each score and enter into $X^2$ and $Y^2$ columns respectively.
3. Multiply the X and Y scores together and enter in the XY column.
4. Add up each column.

Table 5.1 Application of the four steps

| Sts. | X | Y | $X^2$ | $Y^2$ | XY |
|---|---|---|---|---|---|
| 1 | 13 | 7 | 169 | 49 | 91 |
| 2 | 12 | 11 | 144 | 121 | 132 |
| 3 | 10 | 3 | 100 | 9 | 30 |
| 4 | 8 | 7 | 64 | 49 | 56 |
| 5 | 7 | 2 | 49 | 4 | 14 |
| 6 | 6 | 12 | 36 | 144 | 72 |
| 7 | 6 | 6 | 36 | 36 | 36 |
| 8 | 4 | 2 | 16 | 4 | 8 |
| 9 | 3 | 9 | 9 | 81 | 27 |
| 10 | 1 | 6 | 1 | 36 | 6 |
| Totals | X=70 | Y= 65 | $X^2$= 624 | $Y^2$= 533 | XY = 472 |

5 Finally apply the following formula using the obtained results indicated in the table:

$$r_{xy} = \frac{N (\sum XY) - (\sum X) ( \sum Y)}{[ N \sum X^2 - (\sum X)^2 ] \; [ N \sum Y^2 - (\sum Y)^2 ]}$$

$$= \frac{(10) (472) - (70) (65)}{[(10) (624) - (70)^2] \; [(10) (533) - (65)^2]}$$

$$= \frac{(4720 -4550)}{[ (6240 - 4900)] \; [(10) (533) - (65)^{2]}}$$

$$= \frac{170}{(1240) (1105)}$$

$$= \frac{170}{1480700}$$

$$= \frac{170}{1218.84}$$

$$= +.14$$

Computing r using covariance

Covariance is another way of computing correlation. Covariance is concerned with how two variables covary. Covariance is defined as the cross-product of the deviation scores from X and Y.

$$\mathrm{Cov}_{r\ xy} = \frac{\sum (X-X)\ (Y-Y)}{N - 1}$$

Since the two tests do not have equal standard deviations, the covariance value must be adjusted for the amount of variance in both X and Y. This is attained by dividing covariance by the cross-product of standard deviations of X and Y. The adjusted covariance gives us the Pearson product moment correlation coefficient.

$$r_{xy} = \frac{\mathrm{Cov}_{xy}}{S_x\ S_y}$$

The Pearson r is a more appropriate measure of correlation if the data represent *interval* or *ratio* scales. Since most educational measures represent interval scales, the Pearson r is the most appropriate one to be utilized as long as the relationship between the variables to be correlated is linear.

## Interpretation of Correlation Coefficient

Correlation coefficient can be interpreted easily if the degree of variance overlapping between the two measures is calculated. This allows us to see how much variance in one measure can be accounted for by the other. To do this, we simply square the correlation coefficient to obtain the common variance between the two tests. The total standardized variance in any test is 1. The two measures share variance depending on how much they correlate (see Figures 5.6 and 5.7).

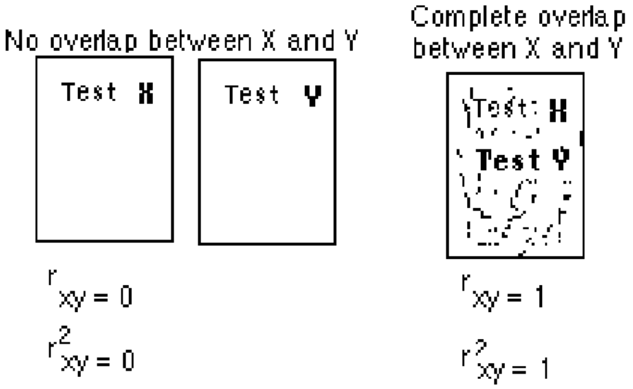

Figure 5.6 No variance means no correlation and perfect overlap means perfect correlation

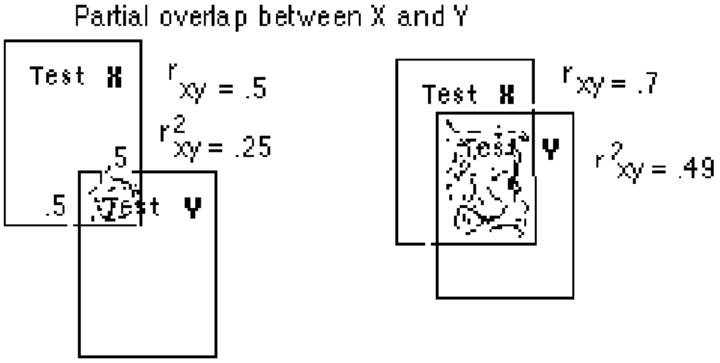

Fig. 5.7 Partial correlation

To the extent that the correlation between two sets of scores deviates from the perfect overlap of 1, there is the indication of less space shared by the two measures. The following are the degrees starting from complete correlation ending with almost none.

| $^{r}$xy =1.0 | $^{r}$xy = 1.00 |
|---|---|
| .9 | .81 |
| .8 | .64 |
| .7 | .49 |
| .6 | .36 |
| .5 | .25 |
| .4 | .16 |
| .3 | .09 |
| .2 | .04 |

.1 .01

## Testing the Pearson Correlation Coefficient

In order to see the applicability of the results obtained from a sample, to a whole population, the population correlation coefficient (p= *rho*) is computed. At the initial stage, both variables (x and y) are assumed to be normally distributed. Therefore, the researcher starts with a null hypothesis ($H_o$) claiming that the given variables are not correlated (see also Vol. 1. Chapter 3). Thus, $H_o$ means x and y are not correlated, so p = 0

The expected results to be obtained from the research are formulated in the alternative hypothesis ($H_1$). There are three ways of formulating $H_{1,}$ the choice of this alternative hypothesis depends on the expectation of the results being *positive*, *negative*, or simply *zero*. The assumption, the formulation of $H_{1,}$ and what type of hypothesis to use for each case is given in Table 5.2.

Table 5.2 Alternate Hypotheses

| Assumption | Formulation of Hypotheses | Type of Tests Used |
|---|---|---|
| p > O | $H_0$: p = 0<br>$H_1$: p > 0 | Right-tailed test |
| p < O | $H_0$: p = 0<br>$H_1$: p < 0 | Left-tailed test |
| p = O | $H_0$: p = 0<br>$H_1$: p = 0 | Two-tailed test |

The right-tailed test, left-tailed test, and the two-tailed test are related with the values indicated by r and consequently with the interpretation of the values. For instance, if the r value for the sample is 0.87, which is close to 1, it is interpreted as that x and y have a positive relationship. Accordingly, an alternative hypothesis is formulated in the same direction for the whole population ($H_1$: p > 0), and right-tailed test is used since this positive value falls to the right of the normal curve.

____________________________ r
-1 0 + 1

The significance of this positive relationship between x and y is determined by referring to the statistical figures listed in statistical books (see Appendix C: Table C1). In order to find the appropriate figure, the following information is needed:

- the number of data points ( the number of the sample)
- the level of significance (either 0.01 or 0.05)

The number of our sample is 10, so we find the critical value for number 10 from the list, which is .72 at 0.01 level; and .54 at 0.05 level for one-tailed tests. This means that within the normal curve, the region to the right of .72 is the critical, or rejection region. Since r obtained for this sample is 0.87, it falls into this critical region at a significance level of 0.01. Therefore, the null hypothesis ($H_0$: p = 0 ), which claims that there is no correlation, is rejected; the alternate hypothesis ($H_1$: p > 0), which claims that there is a positive relationship between the variables, is accepted as true. In other words, the correlation coefficient of the population is concluded to be positive.

If the r value calculated were 0.67 instead of 0.87, it would not be within the boundaries of the critical value (0.72 and 1) indicated for the level of 0.01 but within the boundaries of 0.54 and 1). This means that correlation coefficient being 0.67 would still indicate significance but at a level of 0.05 rather than at a level of 0.01.

If the calculated r value is close to -1 indicating a negative correlation, then the alternative hypothesis is formulated as $H_1$: p < 0. For that reason, a left-tailored test is used since the negative values are indicated to the left of the normal curve. Again depending on the number of the sample, the critical value is set up. Since it is a left-tailed test, the critical value corresponding to the number (the number comprising the sample) is indicated with a negative sign. For instance, if the r value were -0.87, then the critical value for 10 would be -0.72 at the significance level of 0.01 and 0.54 at the significance level of 0.05.

If the calculated r value is close to 0.0, then the assumption would be that there is no correlation. Here the null hypothesis is formulated as $H_0$: p = 0. Under these circumstances, in figuring out the critical value of the obtained value, figures under the two-tailed tests need to be taken into consideration because neither a positive nor a negative correlation is expected. Suppose the correlation coefficient is 0.25, this value does not fall in the critical region. The null hypothesis is not rejected, and it is concluded that r being 0.25 is not significant even at 0.05 level.

## Spearman Rank-order Correlation
## (Rho)

The Spearman rho, which is another type of correlation, is appropriately applied if the data for one of the variables are given in ranks instead of scores. In other words, when the data represent an ordinal scale, and when the median is used in calculation, Spearman rho is much more appropriate. If one of the variables is given in the rank order, the other variable to be correlated should be expressed in terms of ranks as well. Another important issue that needs to be considered is that if there is more than one subject with the same score, the average of these students' ranks is taken.

For example, in comparing the performance of three students on a writing test, the student who is thought to have performed the best is placed in the highest order, and accordingly the others are ranked from high to low according to their comparable performance. Therefore, in the computation of rho, the scores on the two variables are arranged in rank order from high to low. The obtained coefficient indicates how the ranking of scores on the two variables are related. The interpretation of the results obtained from rho is similar to that for Pearson r. The coefficients range from +1.0 to -1.0.

Generally, rho is applied to small samples (of less than 30). For example, by means of this type of measure, the degree of correlation between the two learners on the order of morpheme accuracy is computed. In this case, the aim is to see whether there is a correlation between the learning speeds and priorities regarding morphemes. To give another example, let us suppose that there are five candidates from the department for a chairperson position. In order to assess these candidates, they are ranked by two different groups such as the instructors and the students. In order to see the reliability of the obtained scores, the correlation on how these candidates are ranked by different groups is measured by rho. By means of rho computation, answers to the following questions are obtained:

- Is there a relationship between instructor ranking and student ranking? In other words, does a candidate who is ranked higher by the instructors tend to ranked higher by the students as well?
- Is it just the contrary? In other words, while a candidate is ranked high by the students, is he/she ranked low by the instructors?

## Computation of Spearman Rank-order Correlation

Taking the above example, we can compute rho by assigning imaginary means for student and instructor ranks for the five candidates.

| | Instructor Rank | Student Rank | | |
|---|---|---|---|---|
| Candidates | x | y | d = x-y | $d^2$ |
| 1 | 3 | 2 | 1 | 1 |
| 2 | 5 | 4 | 1 | 1 |
| 3 | 1 | 1 | 0 | 0 |
| 4 | 4 | 5 | - 1 | 1 |
| 5 | 2 | 3 | - 1 | 1 |
| | | | | $\sum d^2 = 4$ |

Since the sample size is 5, and the sum of $d^2$ is 4, the Spearman rank-order correlation coefficient is

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

$$= 1 - \frac{6(4)}{5(25 - 1)}$$

$$= .829$$

## Interpretation of Spearman Rank-order Correlation Coefficient:

The Spearman rank-order correlation coefficient (rho) indicates a value between -1 and 1. If rho is closer to -1, it means that there is a strong negative relationship between variable x and variable y. If rho is 0, there is no recognized relationship between x and y. If rho is closer to +1, then a claim is made that there is a strong position relationship between the two variables. Thus, values of rho, close to -1 or +1 are indications of a strong tendency for x and y to have a monotone relationship either in the negative or the positive direction. Values close to 0 indicate a very weak monotone relationship.

The probability distribution of rho for the whole population depends on the sample size. The Spearman rank-order correlation coefficient is the sample estimate for ps, the population Spearman rank-order correlation coefficient.

A test of significance for the Spearman rank-order correlation coefficient is constructed in much the same way it is constructed for Pearson correlation coefficient. In statistical terms, the hypothesis of a research study is expressed in the null form $H_0$ p = 0 (see Vol. 1, Chapter 3 and also the section on Pearson product-moment correlation coefficient), which does not claim any relationship between the variables. The alternate hypothesis, formulating the assumed result of the research, depends on the type of test to be used (see Table 3.1).

The interpretation of the results is similar to the one explained in relation to the Pearson product-moment correlation. The critical values and critical regions of the rho distribution are determined from statistical tables taking into consideration the number of the sample (see Appendix C: Table C2). Any rho value that falls within the given critical region is considered statistically significant.

## Simple Linear Regression

In previous sections, linear correlation with related statistical assessments was discussed. As indicated in these sections, in *correlation* problems, two or more variables are studied simultaneously *to investigate the relationship* between these variables.

In some cases, only one variable of interest is taken into consideration, and other variables are used *to predict the behavior* of this specific variable under the given conditions. Statistical problems investigating this type of issues are called *regression* problems. As expressed by Brase & Brase (1995) " the methods of regression literally predict the value of one variable by going back to (or regressing to) the values of another related variable" (p. 665).

The study of regression, as in the case of correlation, begins with a table and/or a graph illustrating the values of the paired data. For instance, if we want to know if there is a connection between the writing and reading scores in foreign language classes, students from these classes are chosen randomly from different levels of proficiency, and their scores for both of these language skills are calculated at different intervals to find out the correlation between the two sets of scores for each student.

If the correlation were always perfect, in order to predict one score from another score, the researcher would only locate the score of the individual on one axis (X or Y) and then find where the line from that score hits the slope line and check across (see Figure 5.8) .

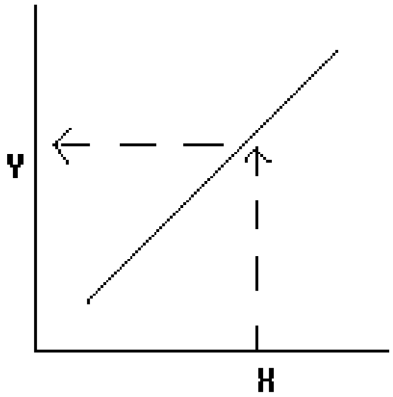

Figure 5.8 Hypothetical prediction based on the linear correlation

This, however, is not possible in the real situation. Thus, to make an accurate prediction about a score on Y given the information on the score for X, there is a need for the following information:

- the mean of Y (Y)
- the score on X
- the slope of the best fitting straight line of the joint distribution.

The statistics used for such predictions is regression analysis. By regressing Y on X, it is possible to predict Y from X. The first step is to find the best-fitting straight line, the slope, which is symbolized as a$^{s}$ b). In order to determine the slope, there is a need to know the correlation coefficient between X and Y and the standard deviation for X and Y. Accordingly, the formula to determine the slope can be indicated in the following manner:

Slope (b) = correlation coefficient $\frac{Sy}{Sx}$

For example, if the correlation coefficient is .40 and the standard deviation is 10 for Y and 9 for X, b is calculated as .44.

$$b = .40 \quad \frac{10}{9}$$

b = .44

As for the regression formula, which is sometimes called the prediction formula) is as follows:

Y = Y + b (X -X)

For example, X was a reading test and Y was a grammar test, and the mean of X was 80. A student who got 55 on the reading test (X) would be expected to get 75 on the grammar test (Y). Accordingly:

$X = 80$
$b = .60$
$X = 55$
$Y = 75$

$$\begin{aligned} Y &= Y + b\,(X - X) \\ &= 75 + .60\,(55 - 80) \\ &= 75 + .60\,(-25) \\ &= 75 - 15 \\ &= 60 \end{aligned}$$

The accuracy of prediction depends entirely on the correlation. If the correlation is weak, the prediction is weak. The standard error of estimate (SEE) tells us how great the amount of error in prediction is likely to be. The larger the SEE, the greater the amount of error in prediction. If it is very large, the mean can be used as the best estimate.

Predicting X from Y

$$\begin{aligned} X &= X + b\,(Y - Y) \\ b &= r\,\frac{sx}{sy} \end{aligned}$$

Formulas to calculate SEE:

$$SEE = sx\ V\ 1 - r^2$$

$$S_e = V\frac{SSy - bSSxy}{n - 2}$$ (Braser & Braser, 1995, p. 676)

The values of X and Y are exchanged in all the formulas. As for the multiple regression analysis, it is done on the computer using a statistical program.

In Chapter 3, we tried to discuss the simple measures serving as a base for descriptive statistics. In this chapter, we demonstrated the use of statistics in computing or assessing the relationship between variables.

## EXERCISES

A. Below are the freshman GPA's (X) and t the entrance exam scores(Y) of the same group of students. Calculate the Product-Moment Correlation Coefficient (Pearson r) for these two sets of test scores.

| X | Y |
|---|---|
| 24 | 325 |
| 35 | 310 |
| 50 | 400 |
| 80 | 450 |
| 30 | 315 |
| 46 | 320 |
| 75 | 425 |
| 82 | 430 |
| 65 | 450 |
| 69 | 330 |

B. Interpret the meaning of the correlations presented for these exams measuring the four different language skills.

| | Listening | Speaking | Reading | Writing |
|---|---|---|---|---|
| Listening | 1.000 | .874 | .919 | .774 |
| Speaking | | 1.000 | .796 | .838 |
| Reading | | | 1.000 | .880 |
| Writing | | | | 1.000 |

C. Suppose you wanted to conduct a study to find the relationship between the oral performance in Second Language and the amount of alcohol taken before starting the oral performance. Can you use Pearson Product-moment Correlation Coefficient (r) to measure the relationship between the given variables?
If not, why?