UNIT 1 PRODUCT MOMENT COEFFICIENT OF CORRELATION

Structure

1.0 INTRODUCTION

We measure psychological attributes of people by using tests and scales in order to describe individuals. There are times when you realise that increment in one of the characteristics is associated with increment in other characteristic as well. For example, individuals who are more optimistic about the future are more likely to be happy. On the other hand, those who are less optimistic about future (i.e., pessimistic about it) are less likely to be happy. You would realise that as one variable is increasing, the other is also increasing and as the one is decreasing the other is also decreasing. In the statistical language it is referred to as correlation. It is a description of "relationship" or "association" between two variables (more than two variables can also be correlated, we will see it in multiple correlation).

In this unit you will be learning about direction of Correlation that is, Positive and Negative and zero correlation. You will also learn about the strength of correlation and how to measure correlation. Specifically you will be learning Pearson's Product Moment Coefficient of Correlation and how to interpret this correlation coefficient. You will also learn about the ramifications of the Pearson's r. You will also learn the coefficient of correlation equations with numerical examples.

1.1 OBJECTIVES

After reading and doing exercises in this unit, you will be able to:

- describe and explain concept of correlation;
- plot the scatter diagram;
- explain the concept of direction, and strength of relationship;
- differentiate between various measures of correlations;
- analyse conceptual issues in correlation and causality;
- describe problems suitable for correlation analysis;
- describe and explain concept of Pearson's Product Moment Correlation;
- compute and interpret Pearson's correlation by deviation score method and raw score method; and
- test the significance and apply the correlation to the real data.

1.2 CORRELATION: MEANING AND INTERPRETATION

Correlation is a measure of association between to variables. Typically, one variable is denoted as X and the other variable is denoted as Y. The relationship between these variables is assessed by correlation coefficient. Look at the earlier example of optimism and happiness. It states the relationship between one variable, optimism (X) and other variable, happiness (Y). Similarly, following statements are example of correlations:

As the *intelligence* (IQ) increases the *marks* obtained increases.

As the *introversion* increases *number of friends* decreases.

More the *anxiety* a person experiences, weaker the *adjustment* with the stress.

As the score on *openness to experience* increases, scores on *creativity* test also increase.

More the *income*, more the *expenditure*.

On a reasoning task, as the *accuracy* increases, the *speed* decreases.

As the *cost* increases the *sales* decrease.

Those who are good at *mathematics* are likely to be good at *science*.

As the age of the child increases, the *problems solving capacity* increase.

More the *practice*, better the *performance*.

All the above statements exemplify the correlation between two variables. The variables are shown in *italics*. In this first section, we shall introduce ourselves to the concept of correlation.

1.2.1 Scatter Diagram: Graphical Presentation of Relationship

Scatter diagram (also called as *scatterplot, scattergram,* or *scatter*) is one way to study the relationship between two variables. Scatter diagram is to plot pairs of values of subjects (observations) on a graph. Let's look at the following data of five subject, A to E (Table 1.1). Their scores on intelligence and scores on reasoning task are provided. The same data is used to plot a scatter diagram shown in Figure 1.1. Now, I shall quickly explain 'how to draw the scatter diagram'.

Table 1: Data of five subjects on intelligence and scores of reasoning

Subject	Intelligence	Scores on reasoning task
A	104	12
B	127	25
C	109	18
D	135	31
E	116	19

Step 1. Plotting the Axes

Draw the x and y axis on the graph and plot one variable on x-axis and another on y-axis.

(Although, correlation analyses do not restrict you from plotting any variable on any axis, plot the causal variable on x-axis in case of implicitly assumed cause-effect relationship.)

Also note that correlation does not necessarily imply causality.

Step 2. Range of Values

Decide the range of values depending on your data.

Begin from higher or lower value than zero.

Conventionally, the scatterplot is square.

So plot x and y values about the same length.

Step 3. Identify the pairs of values

Identify the pairs of values.

A pair of value is obtained from a data.

A pair of values is created by taking a one value on first variable and corresponding value on second variable.

Step 4. Plotting the graph

Now, locate these pairs in the graph.

Find an intersection point of x and y in the graph for each pair.

Mark it by a clear dot (or any symbol you like for example, star).

Then take second pair and so on.

The scatterplot shown below is based on the data given in table 1. (Refer to Figure 1).

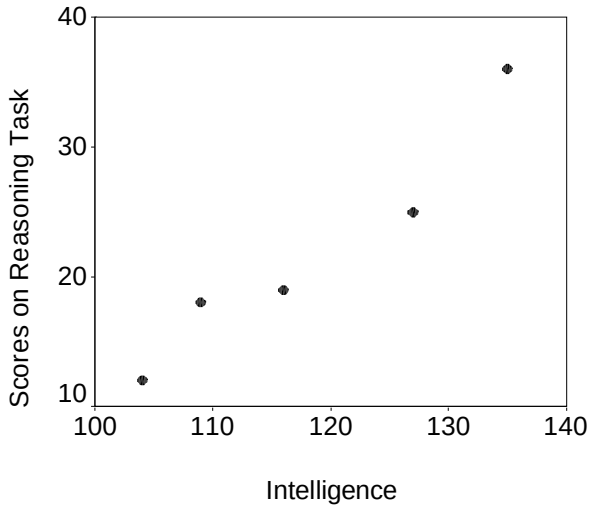


Fig. 1: Scatter diagram depicting relationship between intelligence and score on reasoning task

The graph shown above is scatterplot representing the relationship between intelligence and the scores on reasoning task. We have plotted intelligence on x-axis because it is a cause of the performance on the reasoning task. The scores on reasoning have started from 100 instead of zero simply because the smallest score on intelligence is 104 which is far away from zero. We have also started the range of reasoning scores from 10 since the lowest score on reasoning is 12. Then we have plotted the pair of scores. For example, subject A has score of 104 on intelligence and 12 on reasoning so we get x,y pair of 104,12. We have plotted this pair on the point of intersection between these two scores in the graph by a dot. This is the lowest dot at the left side of the graph. You can try to practice the scatter by using the data given in the practice.

1.2.2 Correlation: Linear and Non-Linear Relationship

The relationship between two variables can be of various types. Broadly, they can be classified as linear and nonlinear relationships. In this section we shall try to understand the linear and nonlinear relationships.

Linear Relationship

One of the basic forms of relationship is linear relationship. *Linear* relationship can be expressed as a relationship between two variables that can be plotted as a *straight* line. The linear relationship can be expressed in the following equation (eq. 1.1):

$$Y = á + â\ X \qquad \text{(eq. 1.1)}$$

In the equation 1.1,

- Y is a dependent variable (variable on y-axis),

- á (alpha) is a constant or Y intercept of straight line,
- â (beta) is slope of the line and
- X is independent variable (variable on x-axis).

We again plot scatter with the line that best fits for the data shown in table 1. So you can understand the linearity of the relationship. Figure 2 shows the scatter of the same data. In addition, it shows the line which is best fit line for the data. This line is plotted by using the method of least squares. We will learn more about it later (Unit 4). Figure 2 shows that there is a linear relationship between two variables, intelligence and Scores on Reasoning Task. The graph also shows the straight line relationship indicating linear relation.

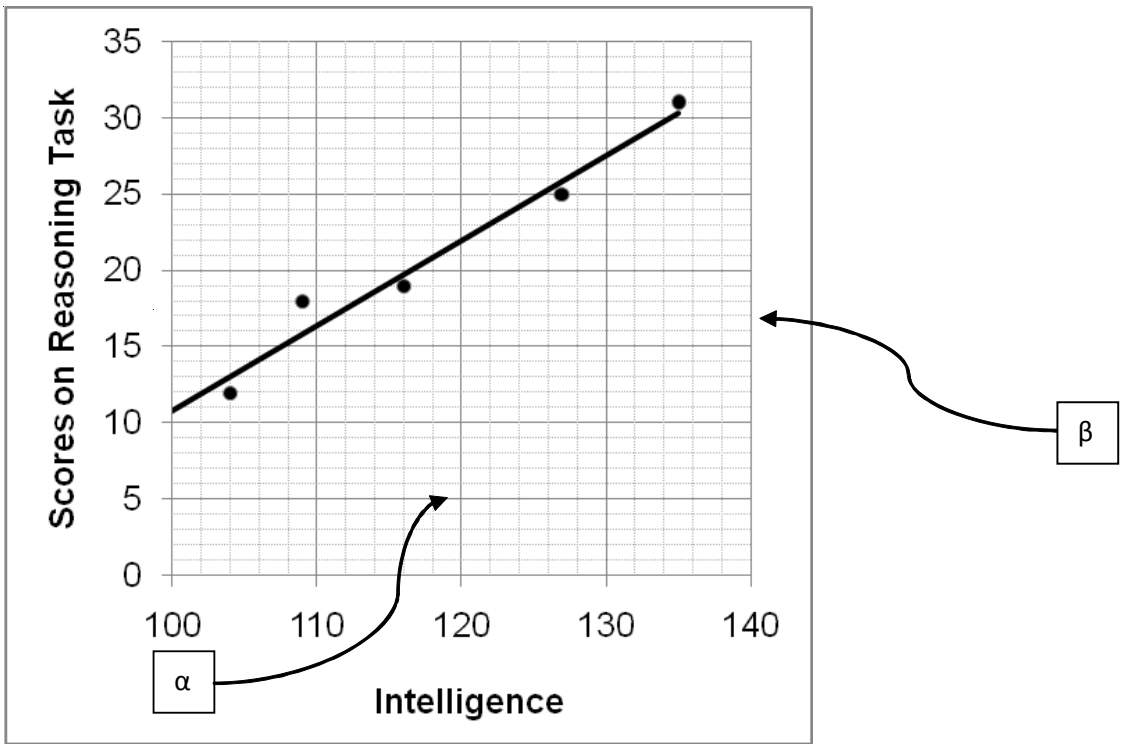


Fig. 2: Scatter showing linearity of the relationship between Intelligence and Scores on Reasoning Task

Non-linear Relationship

There are other forms of relationships as well. They are called as curvilinear or non-linear relationships. The Yorkes-Dodson Law, Steven's Power Law in Psychophysics, etc. are good examples of non-linear relationships. The relationship between stress and performance is popularly known as Yorkes-Dodson Law. It suggests that the performance is poor when the stress is too little or too much. It improves when the stress is moderate. Figure 3 shows this relationship. The *non-linear* relationships, cannot be plotted as a *straight line*.

The performance is poor at extremes and improves with moderate stress. This is one type of curvilinear relationship.

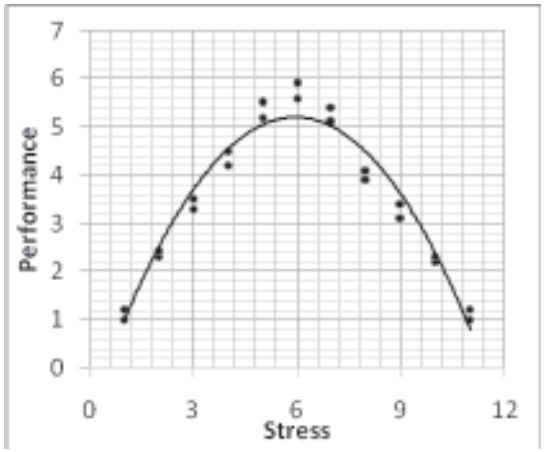


Fig. 3: Typical relationship between stress and performance

The curvilinear relationships are of various types (cubic, quadratic, polynomial, exponential, etc.). The point we need to note is that *relationships can be of various types*. This block discussed only *linear* relationships. Other forms of relationship are *not* discussed. The types of correlation presented in this block represent linear relationships. Pearson's product-moment correlation, Spearman's rho, etc. are linear correlations.

The Stevens' Power Law states that $r = cs^b$ where, r is sensation, s is stimulus, c and b are constants and coefficients, respectively. This is obviously a non-linear relationship between stimulus and sensation. Although, a reader who can recall some basic mathematics of 10^{th} grade can easily understand that by taking the log of both sides, the equation can be converted into linear equation.

1.2.3 Direction of Correlation: Positive and Negative

The direction of the relationship is an important aspect of the description of relationship. If the two variables are correlated then the relationship is either positive or negative. The absence of relationship indicates "zero correlation". Let's look at the positive, negative and zero correlation.

Positive Correlation

The positive correlation indicates that as the values of one variable increases the values of other variable also increase. Consequently, as the values of one variable decreases, the values of other variable also decrease. This means that both the variables move in the samer direction. For example,

a) As the *intelligence* (IQ) increases the *marks* obtained increases.

b) As *income* increases, the *expenditure* increases.

The figure 4 shows *scatterplot* of the positive relationship. You will realise that the higher scores on X axis are associated with higher score on Y axis and lower scores on X axis are generally associated with lower score on Y axis. In the 'a' example, higher scores on *intelligence* are associated with the higher score on *marks obtained*. Similarly, as the scores on *intelligence* drops down, the *marks obtained* has also dropped down.

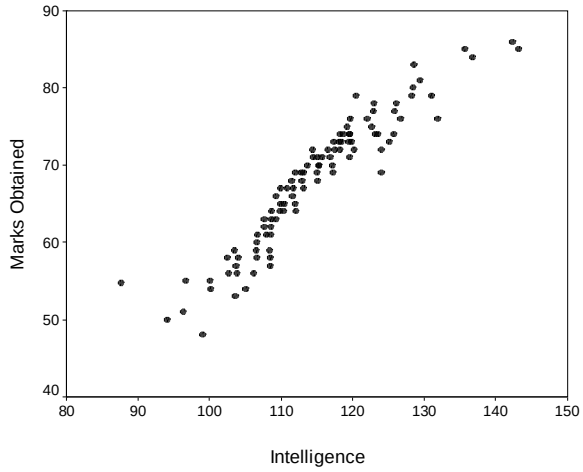


Fig. 4: Positive correlation: Scatter showing the positive correlation between *intelligence* and *marks obtained.*

Negative Correlation

The Negative correlation indicates that as the values of one variable increases, the values of the other variable decrease. Consequently, as the values of one variable decreases, the values of the other variable increase. This means that two variables move in the opposite direction. For example,

a) As the *intelligence* (IQ) increases the *errors on reasoning task* decreases.

b) As *hope* increases, *depression* decreases.

Figure 5 shows *scatterplot* of the negative relationship. You will realise that the higher scores on X axis are associated with lower scores on Y axis and lower scores on X axis are generally associated with higher score on Y axis.

In the 'a' example, higher scores on *intelligence* are associated with the lower score on *errors on reasoning task*. Similarly, as the scores on *intelligence* drops down, the *errors on reasoning task* have gone up.

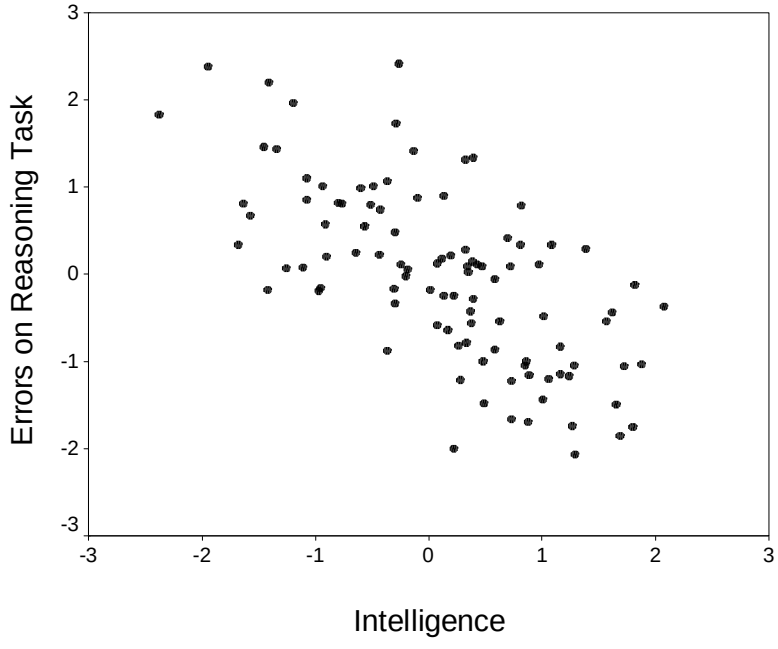


Fig. 5: Negative correlation: Scatter showing the negative correlation between *intelligence* and *errors on reasoning task*

No Relationship

Until now you have learned about the positive and negative correlations. Apart from positive and negative correlations, it is also possible that there is no relationship between *x* and *y*. That is the two variables do not share any relationship. If they do not share any relationship (that is, technically the correlation coefficient is zero), then, obviously, the direction of the correlation is neither positive nor negative. It is often called as zero correlation or no correlation.

(Please note that 'zero order correlation' is a different term than 'zero correlation' which we will discuss afterwards).

For example, guess the relationship between shoe size and intelligence?

This sounds an erratic question because there is no reason for any relationship between them. So there is no relationship between these two variables.

The data of one hundred individuals is plotted in Figure 6. It shows the scatterplot for no relationship.

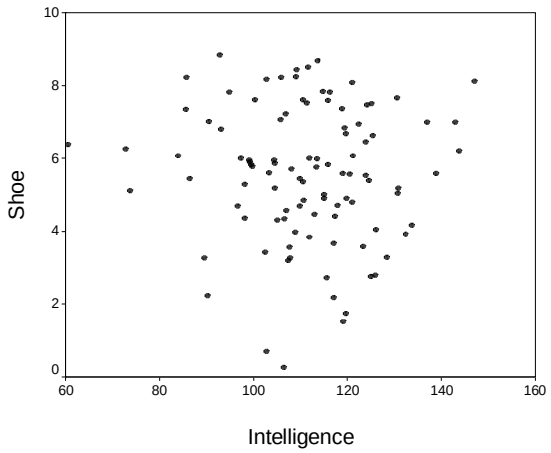


Fig. 6: Scatter between shoe size and intelligence of the individual

1.2.4 Correlation: The Strength of Relationship

You have so far learnt the direction of relationship between two variables. Any curious reader will ask a question "how strong is the relationship between the two variables?" For example, if you correlate intelligence with scores on reasoning, and creativity, what kind of relationship will you expect?

Obviously, the relationship between intelligence and reasoning as well as the relationship between intelligence and creativity are positive. At the same time the correlation coefficient (described in the following section) is higher for intelligence and reasoning than for intelligence and creativity, and therefore we realise that the relationship between intelligence and reasoning is stronger than relationship between intelligence and creativity. The strength of relationship between the two variables is an important information to interpret the relationship.

Correlation Coefficient

The correlation between any two variables is expressed in terms of a number, usually called as correlation coefficient. The correlation coefficient is denoted by various symbols depending on the type of correlation. The most common is '*r*' (small 'r') indicating the Pearson's product-moment correlation coefficient.

The representation of correlation between X and Y is r_{xy}.

The range of the correlation coefficient is from –1.00 to + 1.00.

It may take any value between these numbers including, for example, – 0.72, – 0.61, – 0.35, + 0.02, + 0.31, + 0.98, etc.

If the correlation coefficient is 1, then relationship between the two variables is perfect.

This will happen if the correlation coefficient is – 1 or + 1.

As the correlation coefficient moves nearer to + 1 or – 1, the strength of relationship between the two variables increases.

If the correlation coefficient moves away from the + 1 or – 1, then the strength of relationship between two variables decreases (that is, it becomes weak).

So correlation coefficient of + 0.87 (and similarly – 0.82, – 0.87, etc.) shows strong association between the two variables. Whereas, correlation coefficient of + 0.24 or – 0.24 will indicate weak relationship. Figure 7 indicates the range of correlation coefficient.

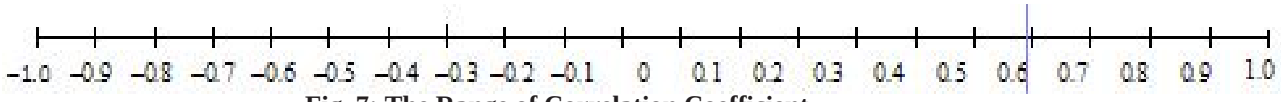


Fig. 7: The Range of Correlation Coefficient.

You can understand the strength of association as the common variance between two correlated variables. The correlation coefficient is NOT percentage.

So correlation of 0.30 does NOT mean it is 30% variance.

The shared variance between two correlated variables can be calculated. Let me explain this point. See, every variable has variance. We denote it as S_x^2 (variance of X). Similarly, Y also has its own variance (S_y^2). In the previous block you have learned to compute them. From the complete variance of X, it shares some variance with Y. It is called covariance.

The Figure 8 shown below explains the concept of shared variance. The circle X indicates the variance of X. Similarly, the circle Y indicates the variance of Y. The overlapping part of X and Y, indicated by shaded lines, shows the shared variance between X and Y. One can compute the shared variance.

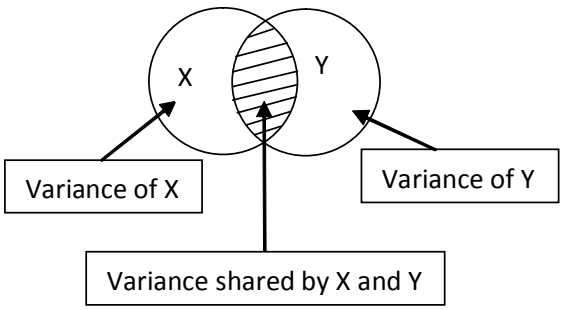


Fig. 8: Covariance indicates the degree to which X shares variance with Y

To calculate the percentage of shared variance between X and Y (common variance), one needs to square the correlation coefficient (*r*). The formula is given below:

Percentage of common variance between X and Y = $r_{xy}^2 \times 100$ (eq. 1.2)

For instance, if the correlation between X and Y is 0.50 then the percent of variation shared by X and Y can be calculated by using equation 1.2 as follows.

Percentage of common variance between X and Y = $r_{xy}^2 \times 100 = 0.50^2 \times 100 = 0.25 \times 100 = 25\%$

It indicates that, if the correlation between X and Y is 0.50 then 25% of the variance is shared by the two variables, X and Y. You would note that this formula is applicable to negative correlations as well. For instance, if $r_{xy} = -0.81$, then shared variance is:

Percentage of common variance between X and Y = $\times 100 = -0.81^2 \times 100 = 0.6561 \times 100 = 65.61\%$

1.2.5 Measurements of Correlation

Correlation coefficient can be calculated by various ways. The correlation coefficient is a description of association between two variables in the sample. So it is a descriptive statistics. Various ways to compute correlation simply indicate the degree of association between variables *in the sample*. The distributional assumptions are *not* required to compute correlation as a descriptive statistics. So it is not a parametric or nonparametric statistics.

The calculated sample correlation coefficient can be used to estimate population correlation coefficient.

The sample correlation coefficient is usually denoted by symbol '*r*'.

The population correlation coefficient is denoted by symbol '$\tilde{n}$'.

It is Greek letter *rho($\tilde{n}$)*, pronounced as *row* (Spearman's correlation coefficient is also symbolised as *rho*.

This may create some confusion among the readers. Therefore, I shall use symbol r_s for Spearman's *rho* as a sample statistics and $\tilde{n}_s$ to indicate the population value of the Spearman's *rho*.

Henceforth, I shall also clearly mention the meaning with which $\tilde{n}$ is used in this block.

- When the population correlation coefficient is estimated from sample correlation coefficient.
- then the correlation coefficient becomes an inferential statistic.
- Inference about population correlation ($\tilde{n}$) is drawn from sample statistics (*r*).
- The population correlation ($\tilde{n}$) is always unknown.
- What is known is sample correlation (*r*).
- The population indices are called as parameters and the sample indices are called as statistics.
- So $\tilde{n}$ is a parameter and *r* is a statistics.

While inferring a parameter from sample, certain distributional assumptions are required. From this, you can understand that the descriptive use of the correlation coefficient does not require any distributional assumptions.

The most popular way to compute correlation is 'Pearson's Product Moment Correlation (*r*)'. This correlation coefficient can be computed when the data on both the variables is on at least equal interval scale or ratio scale.

Apart from Pearson's correlation there are various other ways to compute correlation. Spearman's Rank Order Correlation or Spearman's *rho* (r_s) is useful correlation coefficient when the data is in rank order.

Similarly, Kendall's *tau* (ô) is a useful correlation coefficient for rank-order data.

Biserial, Point Biserial, Tetrachoric, and Phi coefficient, are the correlations that are useful under special circumstances.

Apart from these, multiple correlations, part correlation and partial correlation are useful ways to understand the associations (Please note that the last three require more than two variables).

1.2.6 Correlation and Causality

The correlation does not necessarily imply causality. But, if the correlation between two variables is high then it might indicate the causality. If X and Y are correlated, then there are three different ways in which the relationship between two variables can be understood in terms of causality.

1) X is a cause of Y.

2) Y is cause of X.

3) Both, X and Y are caused by another variable Z.

However, causality can be inferred from the correlations.

Regression analysis, path analysis, structural equation modeling, are some examples where correlations are employed in order to understand causality.

1.3 PEARSON'S PRODUCT MOMENT COEFFICIENT OF CORRELATION

The Person's correlation coefficient was developed by Karl Pearson in 1886. Person was a editor of "Biometrica" which is a leading journal in statistics. Pearson was a close associate of psychologist Sir Francis Galton. The Pearson's correlation coefficient is usually calculated for two continuous variables. If either or both the variables are not continuous, then other statistical procedures are to be used. Some of them are equivalent to Pearson's correlation and others are not. We shall learn about these procedure after learning the Pearson's Correlation coefficient.

1.3.1 Variance and Covariance: Building Blocks of Correlations

Understanding product moment correlation coefficient requires understanding of mean, variance and covariance. We shall understand them once again in order to understand correlation.

Mean : Mean of variable X (symbolised as $\overline{X}$) is sum of scores ($\sum_{i=1}^{n} X_i$) divided by number of observations (n). The mean is calculated in following way.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{(eq. 1.3)}$$

You have learned this in the first block. We will need to use this as a basic element to compute correlation.

Variance

The variance of a variable X (symbolised as S_X^2) is the sum of squares of the deviations of each X score from the mean of X ($\sum (X - \overline{X})^2$)divided by number of observations (*n*).

$$S_x^2 = \frac{\sum (X - \overline{X})^2}{n} \qquad \text{(eq. 1.4)}$$

You have already learned that standard deviation of variable X , symbolised as , is square root of variance of X, symbolised as .

Covariance

The covariance between X and Y (or S_{XY}) can be stated as

$$Cov_{XY} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{n} \qquad \text{(eq. 1.5)}$$

Covariance is a number that indicates the association between two variables. To compute covariance, deviation of each score on X from its mean ( ) and deviation of each score on Y from its mean ($\overline{y}$) is initially calculated.

Then products of these deviations are obtained.

Then, these products are summated.

This sum gives us the numerator for covariance.

Divide this sum by number of observations (*n*). The resulting number is covariance.

1.3.2 Equations for Pearson's Product Moment Coefficient of Correlation

Having revised the concepts, we shall now learn to compute the Pearson's Correlation Coefficient.

Formula

Since we have already learned to compute the covariance, the simplest way to define Pearson's correlation is…

$$r = \frac{Cov_{XY}}{S_X S_Y} \qquad \text{(eq. 1.6)}$$

Where,

the Cov_{XY} is covariance between X and Y,

S_X is standard deviation of X

S_Y is standard deviation of Y.

Since, it can be shown that Cov_{XY} is always smaller than or equal to $S_X S_Y$, the maximum value of correlation coefficient is bound to be 1.

The sign of Pearson's *r* depends on the sign of Cov_{XY}.

If the Cov_{XY} is negative, then *r* will be negative and

if Cov_{XY} is positive then *r* will be a positive value.

The denominator of this formula ($S_X S_Y$) is always positive. This is the reason for a – 1 to + 1 range of correlation coefficient. By substituting covariance equation (eq. 1.5) for covariance we can rewrite equation 1.6 as

$$r = \frac{\frac{\sum (X - \bar{X})(Y - \bar{Y})}{n}}{S_X S_Y} \qquad \text{(eq. 1.7)}$$

By following a simple rule, a÷ b ÷ c = a ÷ (b × c), we can rewrite equation 1.7 as follows.

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n S_X S_Y} \qquad \text{(eq. 1.8)}$$

1.3.3 Numerical Example

Now we shall use this formula to compute Pearson's correlation coefficient. For this purpose we will use the following data. The cognitive theory of depression argues that hopelessness is associated with depression. Aron Beck developed instruments to measure depression and hopelessness. The BHS (Beck Hopelessness Scale) and the BDI (Beck Depression Inventory) are measures of hopelessness and depression, respectively.

Let's take a hypothetical data of 10 individuals on whom these scales were administered. (In reality, such a small data is not sufficient to make sense of correlation; roughly, at least a data of 50 to 100 observations is required). We can hypothesize that the correlation between hopelessness and depression will be positive. This hypothetical data is given below in table 2.

Table 2: Hypothetical data of 10 subjects on BHS and BDI

Subject	BHS (X)	BDI (Y)	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
1	11	13	0	1	0	1	0
2	13	16	2	4	4	16	8
3	16	14	5	2	25	4	10
4	9	10	-2	-2	4	4	4
5	6	8	-5	-4	25	16	20
6	17	16	6	4	36	16	24
7	7	9	-4	-3	16	9	12
8	12	12	1	0	1	0	0
9	5	7	-6	-5	36	25	30
10	14	15	3	3	9	9	9
n = 10	$\sum X$ =110	$\sum Y$ =120			$\sum(X-\bar{X})^2$ = 156	$\sum(Y-\bar{Y})^2$ = 100	$\sum(X-\bar{X})(Y-\bar{Y})$ = 117
	$\bar{X}=11$	$\bar{Y}=12$					

$$S_X = \sqrt{\sum(X-\bar{X})^2/n} = 4.16$$

$$S_Y = \sqrt{\sum(Y-\bar{Y})^2/n} = 3.33$$

$$r = \sum(X-\bar{X})(Y-\bar{Y})/nS_XS_Y = 117/(10)(4.16)(3.33) = +0.937$$

Step 1. You need scores of subjects on two variables. We have scores on ten subjects on two variables, BHS and BDI.

Step 2. Then list the pairs of scores on two variables in two columns. The order will not make any difference. Remember, same individuals' two scores should be kept together. Label one variable as X and other as Y. We label BHS as X and BDI as Y.

Step 3. Compute the mean of variable X and variable Y. It was found to be 11 and 12 respectively.

Step 4. Compute the deviation of each X score from its mean ( ) and each Y score from its own mean ($\bar{Y}$). This is shown in the column labeled as $X-\bar{X}$ and $Y-\bar{Y}$. As you have learned earlier, the sum of these columns has to be zero.

Step 5. Compute the square of $x-\bar{x}$ and $y-\bar{y}$.

This is shown in next two columns labelled as $(x-\bar{x})^2$ and $(y-\bar{y})^2$.

Step 6. Then compute the sum of these squared deviations of X and Y. The sum of squared deviations for X is 156 and for Y it is 100.

Step 7. Divide them by *n* to obtain the standard deviations for X and Y. The S_x was found to be 4.16. Similarly, the S_y was found to be 3.33.

Step 8. Compute the cross-product of the deviations of X and Y. These cross-products are shown in the last column labeled as $(x - \bar{x})(y - \bar{y})$.

Step 9. Then obtain the sum of these cross-products. It was found to be 117. Now, we have all the elements required for computing *r*.

Step 10. Use the formula of *r* to compute correlation. The sum of the cross-product of deviations is numerator and n, S_x, S_y, are denominators. Compute *r*. the value of *r* is 0.937 in this example.

1.3.4 Significance Testing of Pearson's Correlation Coefficient

Statistical significance testing is testing the hypothesis about the population parameter from sample statistics. When the Pearson's Correlation coefficient is computed as an index of description of relationship between two variables in the sample, the significance testing is not required. The interpretation of correlation from the value and direction is enough.

However, when correlation is computed as an estimate of population correlation, obviously, statistical significance testing is required.

"Whether the obtained sample value of Pearson's correlation coefficient is greater than the value that can be obtained by chance?" is the question answered by statistical significance testing about correlation coefficient.

Different values of correlation can be obtained between any two variables, X and Y, for different samples of different sizes belonging to the same population.

The researcher is not merely interested in knowing the finding in the specific sample on which the data are obtained. But they are interested in estimating the population value of the correlation.

Testing the significance of correlation coefficient is a complex issue. It is because of the distribution of the correlation coefficient. The *t*-distribution and z-distribution are used to test statistical significance of *r*.

The population correlation between X and Y is denoted by $\tilde{n}_{xy}$. The sample correlation is r_{xy}.

As you have learned, we need to write a null hypothesis (H_O) and alternative hypothesis (H_A) for this purpose.

The typical null hypothesis states that population correlation coefficient between X and Y ($\tilde{n}_{xy}$) is zero.

$H_O : \tilde{n}_{xy} = 0$

$H_A : \tilde{n}_{xy} \neq 0$

If we reject the H_O then we accept the alternative (H_A) that the population correlation coefficient is other than zero. It implies that the finding obtained on the data is not a sample-specific error.

Sir Ronald Fisher has developed a method of using *t*-distribution for testing this null hypothesis.

The degrees of freedom (*df*) for this purpose are $n - 2$. Here *n* refers to number of observations.

We can use Appendix C in a statistic book for testing the significance of correlation coefficient. Appendix C provides critical values of correlation coefficients for various degrees of freedom. Let's learn how to use the Appendix C. We shall continue with the example of BHS and BDI.

The correlation between BHS and BDI is +.937 obtained on 10 individuals. We decide to do statistical significance testing at 0.05 level of significance, so our á = .05.

We also decided to apply two-tailed test.

The two-tailed test is used if alternative hypothesis is non-directional, i.e. it does not indicate the direction of correlation coefficient (meaning, it can be positive or negative) and one-tail test is used when alternative is directional (it states that correlation is either positive or negative).

Let us write the null hypothesis and alternative hypothesis:

Null hypothesis

$$H_O : \rho_{\text{BHS BDI}} = 0$$
$$H_A : \rho_{\text{BHS BDI}} \neq 0$$

Now we will calculate the degree of freedom for this example.

$df = n - 2 = 10 - 2 = 8$ (eq. 1.9)

So the *df* for this example are 8. Now look at Appendix C. Look down the leftmost *df* column till you reach *df* = 8. Then look across to find correlation coefficient from column of two-tailed test at level of significance of 0.05. You will reach the critical value of *r*:

$r_{\text{critical}} = 0.632$

Because the obtained (i.e., calculated) correlation value of + 0.937 is greater than critical (i.e., tabled) value, we reject the null hypothesis that there is no correlation between BHS and BDI in the population.

So we accept that there is correlation between BHS and BDI in the population. This method is used regardless of the sign of the correlation coefficient.

We use the absolute value (ignore the sign) of correlation while doing a two-tailed test of significance. The sign is considered while testing one-tailed hypothesis.

For example, if the H_A: $\tilde{n} > 0$, which is a directional hypothesis, then any correlation that is negative will be considered as insignificant.

1.3.5 Adjusted *r*

The Pearson's correlation coefficient (*r*) calculated on the sample is not an unbiased estimate of population coefficient ($\tilde{n}$). When the number of observations (sample size) are small the sample correlation is a biased estimate of population correlation. In order to reduce this bias, the calculated correlation coefficient is adjusted. This is called as adjusted correlation coefficient (r_{adj}).

$$r_{\text{adj}} = \sqrt{1 - \frac{(1 - r^2)(n - 1)}{n - 2}}$$

Where,

r_{adj}= adjusted *r*

r^{2-} = the square of Pearson's Correlation Coefficient obtained on sample,

n = sample size

In case of our data, presented in table 1.2, the correlation between BHS and BDI is +.937 obtained on the sample of 10. The adjusted r can be calculated as follows

$$r_{adj} = \sqrt{1 - \frac{(1 - .937^2)(10 - 1)}{10 - 2}} = \sqrt{1 - \frac{(.1220)(9)}{8}} = \sqrt{1 - 0.1373} = .929$$

The r_{adj} is found to be 0.929. This coefficient is unbiased estimate of population correlation coefficient.

1.3.6 Assumptions for Significance Testing

One may recall that simple descriptive use of correlation coefficient does not involve any assumption about the distribution of either of the variables. However, using correlation as an inferential statistics requires assumptions about X and Y. These assumptions are as follows. Since we are using *t*-distribution, the assumptions would be similar to *t*.

Assumptions:

Independence among the pairs of score

This assumption implies that the scores of any two observations (subjects in case of most of psychological data) are not influenced by each other. Each pair of observation is independent. This is assured when different subjects provides different pairs of observation.

The population of X and the population of Y follow normal distribution and the population pair of scores of X and Y has a normal bivariate distribution.

This assumption states that the population distribution of both the variables (X and Y) is normal. This also means that the pair of scores follows bivariate normal distribution. This assumption can be tested by using statistical tests for normality.

It should be remembered that the *r* is a robust statistics. It implies that some violation of assumption would not influence the distributional properties of *t* and the probability judgments associated with the population correlation.

1.3.7 Ramifications in the Interpretation of Pearson's r

The interpretation of the correlation coefficient depends primarily on two things: direction and strength of relationship. We have already discussed them in detail and hence repetition is avoided.

Direction

If the correlation is positive, then the relationship between two variables is positive. It means that as there is an increase in one there is an increase in another, and as there is a decrement in one there is a decrease in another. When the direction of correlation is negative then the interpretation is vice-versa.

Strength

The strength can be calculated in terms of percentage. We have already learned this formula. So we can convert the correlation coefficient into percentage of common

variance explained and accordingly interpret. For example, if the correlation between X and Y is 0.78, then the common variance shared by X and Y is 60.84 percent.

Usually distinct psychological variables do not share much of the common variance. In fact, the reliability of psychological variables is an issue while interpreting the correlations. Cohen and Cohen have suggested that considering the unreliability of psychological variables, the smaller correlations should also be considered significant.

Although, direction and strength are key pointers while interpreting the correlation, there are finer aspects of interpretation to correlations.

They are :

- range,
- outliers,
- reliability of variables, and
- linearity

The above are some of the important aspects which all obscure the interpretation of correlation coefficient. Let's discuss them one by one.

1.3.8 Restricted Range

It is expected the variables in correlation analysis are measured with full range. For example, suppose we want to study the correlation between hours spent in studies and marks. We are suppose to take students who have varying degree of hours of studies, that is, we need to select students who have spent very little time in studies to the once who have spent great deal of time in studies. Then we will be able to obtain true value of the correlation coefficient.

But suppose we take a very restricted range then the value of the correlation is likely to reduce. Look at the following examples the figure 1.9a and 1.9b.

The figure 1.9a is based on a complete range.

The figure 1.9b is based on the data of students who have studied for longer durations.

The scatter shows that when the range was full, the correlation coefficient was showing positive and high correlation. When the range was restricted, the correlation has reduced drastically.

You can think of some such examples. Suppose, a sports teacher selects 10 students from a group of 100 students on basis of selection criterion, that is their athletic performance.

The actual performance of these ten selected students in the game was correlated with the selection criterion. A very low correlation was obtained between selection criterion and actual game performance. This would naturally mean that the selection criterion is not related with actual game performance. Is it true..? Why so…?

If you look at the edata, you will realise that the range of the scores on selection criterion is extremely restricted (because these ten students were only high scorers) and hence the relationship is weak. So note that whenever you interpret correlations, the range of the variables is large. Otherwise the interpretations will not be valid.

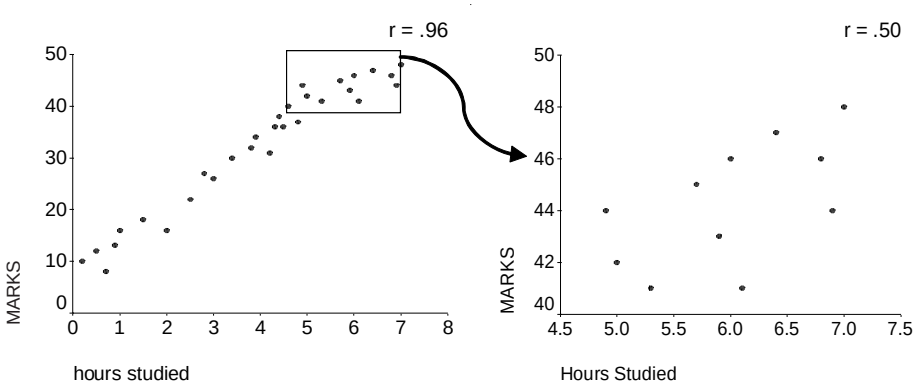


Fig. 1.9a: Scatter showing full range on both variables

Fig. 1.9b: Scatter with restricted range on hours studied

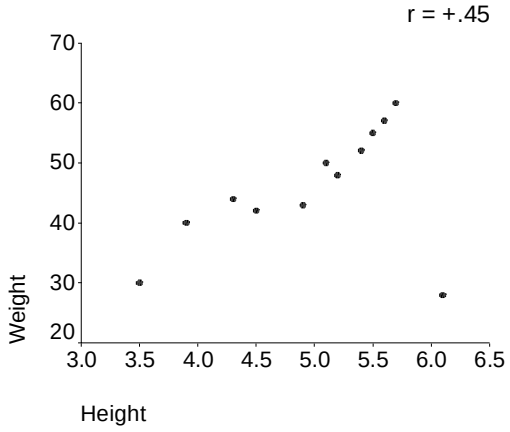
1.4 UNRELIABILITY OF MEASUREMENT

Psychological research involves the use of scales and tests. One of the psychometric property psychological instruments should poses is reliability. Reliability refers to consistency of a measurement. If the instrument is consistent, then the test has high reliability. But at times one of the variable or both the variables may have lower reliability. In this case, the correlation between two less reliable variable reduces. Generally, while interpreting the correlation, the reliability is assumed to be high. The general interpretations of correlations are not valid if the reliability is low. This reduction in the correlation can be adjusted for the reliability of the psychological test. More advanced procedures are available in the books of psychological testing and statistics. They involve calculating disattenuated correlations. Correlation between two variables that have less than perfect reliability is adjusted for unreliability. This is called as disattenuated correlation. If both variables were perfectly reliable then correlation between them is disattenuated correlation.

1.4.1 Outliers

Outliers are extreme score on one of the variables or both the variables. The presence of outliers has deterring impact on the correlation value. The strength and degree of the correlation are affected by the presence of outlier. Suppose you want to compute correlation between height and weight. They are known to correlate positively. Look at the figure below. One of the scores has low score on weight and high score on height (probably, some anorexia patient).

Figure 1.10. Impact of an outlier observation on correlation. Without the outlier, the correlation is 0.95. The presence of an outlier has drastically reduced a correlation coefficient to 0.45.



1.4.2 Curvilinearity

We have already discussed the issue of linearity of the relationship. The Pearson's product moment correlation is appropriate if the relationship between two variables is linear. The relationships are curvilinear then other techniques need to be used. If the degree of cruviliniarity is not very high, high score on both the variable go together, low scores go together, but the pattern is not linear then the useful option is Spearman's *rho*.

1.5 USING RAW SCORE METHOD FOR CALCULATING *r*

The method which we have learned to compute the correlation coefficient is called as deviation scores formula. Now we shall learn another method to calculate Pearson's correlation coefficient. It is called as raw score method. First we will understand how the two formulas are similar. Then we will solve a numerical example for the raw score method. We have learned following formula for calculating *r*.

1.5.1 Formulas for Raw Score

We have already learnt following formula of correlation (eq. 1.8). This is a deviation score formula.

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{nS_X S_Y}$$

The denominator of correlation formula can be written as

$$\sqrt{\sum (X - \bar{X})^2 (Y - \bar{Y})^2} \qquad \text{(eq. 1.10)}$$

Which is

$$\sqrt{(SS_X SS_Y)} \qquad \text{(eq. 1.11)}$$

We have already learnt that

$$SS_X = \sum (X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n} \qquad \text{(eq. 1.12)}$$

and

$$SS_Y = \sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} \qquad \text{(eq. 1.13)}$$

The numerator of the correlation formula can be written as

$$\sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{n} \qquad \text{(eq. 1.14)}$$

So *r* can be calculated by following formula which is a raw score formula:

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(SS_X SS_Y)}} \qquad \text{(eq. 1.15)}$$

1.5.2 Solved Numerical for Raw Score Formula

We shall solve the same numerical example by using the formulas shown above. Table 3 shows how to calculate Pearson's *r* by using raw score formula.

Table 1.3: Table showing the calculation of r by using raw score formula.

Subject	BHS (X)	BDI (Y)	X^2	Y^2	XY
1	11	13	100	676	260
2	13	16	64	529	184
3	16	14	81	529	207
4	9	10	169	676	338
5	6	8	121	576	264
6	17	16	196	900	420
7	7	9	256	729	432
8	12	12	144	729	324
9	5	7	225	841	435
10	14	15	144	625	300
Summation	110	120	1366	1540	1437
	$\bar{X} = 11$	$\bar{Y} = 12$			

$$SS_X = \sum X^2 - (\sum X)^2 / n = 1366 - (110)^2/10 = 156$$

$$SS_Y = \sum Y^2 - (\sum Y)^2 / n = 1540 - (120)^2 / 10 = 100$$

$$\sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{n} \qquad = 117$$

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(SS_X SS_Y)}} \qquad = 0.937$$

Readers might find one of the methods easier. There is nothing special about the methods. One should be able to correctly compute the value of correlation.

1.6 LET US SUM UP

In this unit we started with definition and meaning of correlation and followed it up with how the correlation could be depicted in graphical form and scatter diagram. We then learnt about linear and non-linear and curvilinear relationship amongst variables. We also learnt that the direction of relationship could be either in the positive or in the negative direction. It can also be no correlation in that it can be a zero correlation. Then we learnt the methods of measurement of correlation and

learnt how to use the formula for calculating the Pearson's r, that Pearson's Product Moment Coefficient of Correlation. We then discussed about the building blocks of correlation which included variance and co variance. We also learnt how to test the level of significance of a particular coefficient of correlation calculated by us. Then we learnt about the interpretation of the correlation coefficient and also its ramifications. Then we looked into the unreliability of a correlation and the causes for the same such as the inclusion of outliers etc.

1.7 UNIT END QUESTIONS

1) Problem:

Plot scatter diagram for the following data. Compute Pearson's correlation between x and y. Write the null hypothesis stating that population correlation is zero. Test the significance of the correlation coefficient.

X	Y
12	20
13	22
15	28
17	31
11	22
9	24
8	18
10	21
11	23
7	16

2) Plot scatter for following example. The data was collected on Perceived stress and anxiety on 10 subjects. Compute the Pearson's correlation between them State the null hypothesis. Test the null hypothesis using this hypothesis. Do the similar exercise after deleting a pair that clearly looks an outlier observation.

Perceived stress	Anxiety
9	12
8	11
7	9
4	5
8	9
4	6
6	8
14	2
7	11
11	9
9	11

3) Data showing scores on time taken to complete 200 meters race and duration of practice for 5 runners. Plot the scatter. Compute mean, variance, SD, and covariance. Compute correlation coefficient. Write the null hypothesis.

Time taken (in Seconds)	Duration of Practice (in months)
31	11
32	14
36	9
26	15
38	7

4) Data showing scores on dissatisfaction with work and scores on irritability measured by standardised test for thirteen individuals. Plot the scatter. Compute mean, variance, SD, and covariance. Compute correlation coefficient. Write the null hypothesis stating no relationship. Test the significance at 0.05 level of significance.

Dissatisfaction with work	Irritability scores
12	5
16	7
19	9
27	13
30	16
25	11
22	6
26	14
11	7
17	9
19	14
21	18
23	19

5) Check whether the following statements are true or false.

1)	Positive correlation means as X increases Y decreases.	True/False
2)	Negative correlation means as X decreases Y decreases.	True/False
3)	Generally, in a scatter, lower scores on X are paired with lower scores on Y for negative correlation.	True/False
4)	$-1.00 \leq$ Pearson's correlation $\leq +1.00$	True/False
5)	Generally, in a scatter, lower scores on X are paired with higher scores on Y in positive correlation.	True/False
6)	The scatter diagram cannot indicate the direction of the relationship.	True/False
7)	Percentage of shared variance by X and Y can be obtained by squaring the value of correlation.	True/False

Answers: 1) = False, 2) = False, 3) = False, 4) = True, 5) = False, 6) = False, 7) = True

Answer in brief.

6) What is correlation coefficient?

7) What is the range of correlation coefficient?

8) Is correlation coefficient a percentage?

9) How to calculate common variance from correlation coefficient?

10) What is the percentage of variance shared by X and Y if the $r_{xy} = 0.77$?

11) What is the percentage of variance shared by X and Y if the $r_{xy} = -0.56$?

Answers:

A number expressing the relationship between two variables.

The range of correlation coefficient is from –1.00 to + 1.00.

No. Correlation is not a percentage. But it can be converted into percentage of variance shared.

Common variance is calculated from correlation coefficient by using a formula: $r_{xy}^2 \times 100$.

59.29%

31.36%

1.8 SUGGESTED READINGS

Aron, A., Aron, E. N., Coups, E.J. (2007). *Statistics for Psychology*. Delhi: Pearson Education.

Minium, E. W., King, B. M., & Bear, G. (2001). *Statistical Reasoning in Psychology and Education*. Singapore: John-Wiley.

Guilford, J. P., & Fructore, B. (1978). *Fundamental Statistics for Psychology and Education*. N.Y.: McGraw-Hill.

Wilcoxon, R. R. (1996). *Statistics for Social Sciences*. San Diego: Academic Press.