UNIT 2 OTHER TYPES OF CORRELATION (PHI-COEFFICIENT)

Structure

2.0 INTRODUCTION

We have learned about the correlation as a concept and also learned about the Pearson's coefficient of correlation. We understand that Pearson's correlation is based on certain assumptions, and if those assumptions are not followed or the data is not appropriate for the Pearson's correlation, then what has to be done ? This unit is answering this practical problem. When either the data type or the assumptions are not followed then the correlation techniques listed in this unit are useful. Out of them some are actually Pearson's correlations with different name and some are non-Pearson correlations. The rank data also poses some issues and hence this unit is also providing the answers to this problem. In this unit we shall learn about Special

Types of Pearson Correlation, Special Correlation of Non-Pearson Type, and correlations for rank-order data. The special types of Pearson correlation are Point-Biserial Correlation and Phi coefficient. The non-Pearson correlations are Biserial and Tetrachoric. The rank order correlations discussed are Spearman's *rho* and Kendall's *tau*.

2.1 OBJECTIVES

After completing this unit, you will be able to:

- describe and explain concept of special correlation;
- explain the concept of special correlation and describe and differentiate between their types;
- describe and explain concept of Point-Biserial and Phi coefficient;
- describe and explain concept of Biserial and Tetrachorich coefficient;
- compute and interpret Special correlations;
- test the significance and apply the correlation to the real data;
- explain concept of Spearman's rho and tau coefficient;
- compute and interpret rho and tau; and
- apply the correlation techniques to the real data.

2.2 SPECIAL TYPES OF CORRELATION

The correlation we have learned in the last unit is Pearson's product moment coefficient of correlation. The Pearson's *r* is one of the computational processes for calculating the correlation between two variables. Nevertheless, this is not the only way to calculate correlations. It is just one of the ways of calculating correlation coefficient.

This correlation can be calculated under various restrictions. The variables X and Y were assumed to be continuous variables. The distribution of these variables is expected to be normal. Some homogeneity among the variables is also expected. Linearity of the relationship is also required for computing the Pearson's *r*. There might be instances when one or more of these conditions are not met. In such cases, one needs to use alternative methods of correlations. Some of them are Pearson's correlation modified for specific kind of data. Others are non-Pearson correlations.

Let us take a quick note on distinction between measures of correlation and measures of association. Howell (2002) made this point quite clear. Measures of correlations are those where some sort of order can be assigned to each of the variable. Increment in scores either represent higher levels (or lower levels) of some quantified attribute. For example, number of friends, BHS hope scores, time taken to complete a task, etc. Measures of association are those statistical procedures that are utilised for variables that do not have a property of order. They are categorical variables, or nominal variables, for example association of gender (male and female) with ownership of residence (own and do not own). Both these variables are nominal variables and do not involve any order.

In this unit we shall learn about Special Types of Pearson Correlation, Special Correlation of Non-Pearson Type, and correlations for rank-order data. The special types of Pearson correlation are Point-Biserial Correlation and Phi coefficient. The non-Pearson correlations are Biserial and Tetrachoric. The rank order correlations discussed are Spearman's *rho* and Kendall's *tau*.

2.3 POINT BISERIAL CORRELATION (r_{PB})

Some variables are dichotomous. The dichotomous variable is the one that can be divided into two sharply distinguished or mutually exclusive categories. Some examples are, male-female, rural-urban, Indian-American, diagnosed with illness and not diagnosed with illness, Experimental group and Control Group, etc. These are the truly dichotomous variables for which no underlying continuous distribution can be assumed. Now if we want to correlate these variables, then applying Pearson's formula have problems because of lack of continuity. Pearson's correlation requires continuous variables.

Suppose we are correlating gender, then male will be given a score of 0, and females will be given a score of 1 (or vice versa; indeed you can give a score of 5 to male and score of 11 to female and it won't make any difference for the correlation calculated).

Point Biserial Correlation (r_{pb}) is Pearson's Product moment correlation between one truly dichotomous variable and other continuous variable. Algebraically, the $r_{pb} = r$. So we can calculate r_{pb} in a similar way.

2.3.1 Calculation of r_{pb}

Let's look at the following data. It is a data of 20 subjects, out of which 9 are male and 11 are females. Their marks in the final examination are also provided. We want to correlate marks in the final examination with sex of the subject. The marks obtained in the final examination are a continuous variable whereas sex is truly dichotomous variable, taking two values male or female. We are using value of 0 for male subject and value of 1 for female subjects. The correlation appropriate for this purpose is Point-Biserial correlation (r_{pb}).

Table 1: Data showing the gender and mark for 20 subjects

Subject	Sex (male) X	Marks (Y)	Subject	Sex (Female (X)	Marks (Y)
1	0	46	11	1	58
2	0	74	12	1	69
3	0	58	13	1	76
4	0	67	14	1	78
5	0	62	15	1	65
6	0	71	16	1	69
7	0	54	17	1	59
8	0	63	18	1	53
9	0	53	19	1	73
10	1	67	20	1	81

$\text{Mean}_{\text{sex}} = 0.55$ $\text{Mean}_{\text{marks}} = 64.8$ $\text{Mean Marks}_{\text{male}} = 60.88$

$S_{\text{sex}} = 0.497$ $S_{\text{marks}} = 9.17$ $\text{Mean Marks}_{\text{female}} = 68$

$Cov_{\text{XY}} = 1.76$

$$r = \frac{Cov_{XY}}{S_X S_Y} = \frac{1.76}{0.497 \times 9.17} = 0.386$$

The Pearson's correlation (point biserial correlation) between sex and marks obtained is 0.386. The sign is positive. The sign is arbitrary and need to be interpreted depending on the coding of the dichotomous group. The interpretation of the sign is the group that is coded as 1 has a higher mean than the group that is coded as 0. The strength of correlation coefficient is calculated in a similar way. The correlation is 0.386, so the percentage of variance shared by both the variables is r^2 for Pearson's correlation. Same would hold true for point biserial correlation. The $r_{\text{pb}}{}^2$ is $0.386^2 = 0.149$. This means that 15% of information in marks is shared by sex.

2.3.2 Significance Testing of r_{pb}

The null hypothesis and alternative hypothesis for this purpose are as follows:

H_o: $\tilde{n} = 0$

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

Since the r_{pb} is Pearson's correlation, the significance testing is also similar to it. the *t*- distribution is used for this purpose with $n - 2$ as *df*.

$$t = \frac{r_{pb}\sqrt{n-2}}{\sqrt{1-r_{pb}^2}}$$

(eq. 2.1)

The t value for our data is 1.775. The $df = n - 2 = 20 - 2 = 18$. The value is not significant at 0.05 level. Hence we retain the null hypothesis.

2.4 PHI COEFFICIENT (ϕ)

The Pearson's correlation between one dichotomous variable and another continuous variable is called as point-biserial correlation. When both the variables are dichotomous, then the Pearson's correlation calculated is called as Phi Coefficient (ϕ).

For example, let us say that you have to compute correlation between gender and ownership of the property. The gender takes two levels, male and female. The ownership of property can be measured as either the person owns a property and the person do not own property. Now you have both the variables measured as dichotomous variables. Now if you compute the Pearson's correlation between these two variables is called as Phi Coefficient (ϕ). Both the variables take value of either of 0 or one. Look at the data given in the table below.

Table 2: Data and calculation for correlation between gender and ownership of property

X: Gender	0= Male 1 = Female
Y: Ownership of Property	0=No ownership 1 = Ownership

X	1	0	1	1	0	0	0	0	1	1	1	0
Y	0	1	0	1	1	1	0	1	0	0	1	1

Calculations

$\overline{X} = 0.5$ $\qquad$ $S_x = 0.52$

$\overline{Y} = .58$ $\qquad$ $S_y = 0.51$ $\qquad$ $Cov_{XY} = -0.13$

$$r_{XY} = \phi_{XY} = \frac{Cov_{XY}}{S_X S_Y} = \frac{-0.13}{0.52\times 0.51} = -.465$$

The value of ϕ coefficient is found to be – .465.

The relationship is negative, is function of the way we have assigned the number 0 and 1 to each of the variable. If we assign 0 to females and 1 to males, then we will get the same value of correlation with positive sign. Nevertheless, this does not mean that sign of the relationship cannot be interpreted. Once these numbers have been assigned, then we can interpret the sign. Male is 0 and female is 1; whereas 0 = no ownership and 1 is ownership.

The negative relation can be interpreted as follows: as we move from male to female we move negatively from no ownership to ownership. Meaning that male have more ownership than females. We can also calculate the proportion of variance shared by these two variables.

That is $r^2 = \phi^2 = -.465^2 = 0.216$ percent.

2.4.1 Significance Testing of Phi (ϕ)

The significance can be tested by using the Chi-Square (÷^{2}) distribution.

The ϕ can be converted into the ÷^{2} by obtaining a product of n and ϕ^2.

The Chi-Square of nϕ^2 will have *df* = 1.

The null and alternative hypothesis are as follows:

H$_O$: *ñ* = 0

H$_A$: *ñ* ‘“ 0

÷^{2} = nϕ^2 = 12 × .216 = 2.59 (eq. 2.2)

The value of the chi-square at 1 *df* is 3.84. the obtained value is less than the tabled value. So we accept the null hypothesis which states that the population correlation is zero.

One need to know that this is primarily because of the small sample size. If we take a larger sample, then the values would be significant. Quickly note the relationship between ÷2 and ϕ.

$$\phi = \sqrt{\frac{\chi^2}{n}} \qquad \text{(eq. 2.3)}$$

So one can compute the chi-square and then calculate the phi coefficient.

2.5 BISERIAL CORRELATION

The biserial correlation coefficient (r_b), is a measure of correlation. It is like the point-biserial correlation. But point-biserial correlation is computed while one of the variables is dichotomous and do not have any underlying continuity. If a variable has underlying continuity but measured dichotomously, then the biserial correlation can be calculated.

An example might be mood (happy-sad) and hope, which we have discussed in the first unit. Suppose we measure hope with BHS and measure mood by classifying those who have clinically low *vs*. normal mood. Actually, it is fair to assume that mood is a normally distributed variable.

But this variable is measured discretely and takes only two values, low mood (0) and normal mood (1).

Let's call continuous variable as Y and dichotomized variable as X. the values taken by X are 0 and 1.

So biserial correlation is a correlation coefficient between two continuous variables (X and Y), out of which one is measured dichotomously (X). The formula is very similar to the point-biserial but yet different:

$$r_b = \left[\frac{\overline{Y_1} - \overline{Y_0}}{S_Y}\right]\left[\frac{P_0 P_1}{h}\right] \qquad \text{(eq. 2.4)}$$

where $\overline{Y_0}$ and $\overline{Y_1}$ are the Y score means for data pairs with an X score of 0 and 1, respectively, P_0 and P_1 are the proportions of data pairs with X scores of 0 and 1, respectively, and S_Y is the standard deviation for the Y data, and *h* is ordinate or the height of the standard normal distribution at the point which divides the proportions of P_0 and P_1.

The relationship between the point-biserial and the biserial correlation is as follows.

$$r_b = \frac{r_{pb}\sqrt{p_0 p_1}}{h} \qquad \text{(e.q 2.5)}$$

So once you compute the r_{pb}, its easy to compute the r_b.

2.6 TETRACHORIC CORRELATION (r_{TET})

Tetrachoric correlation is a correlation between two dichotomous variables that have underlying continuous distribution. If the two variables are measured in a more refined way, then the continuous distribution will result. For example, attitude to females and attitude towards liberalisation are two variables to be correlated. Now, we simply measure them as having positive or negative attitude. So we have 0 (negative attitude) and 1 (positive attitude) scores available on both the variables. Then the correlation between these two variables can be computed using Tetrachoric correlation (r_{tet}).

The correlation can be expressed as

$$r = \cos\theta \qquad \text{(eq. 2.6)}$$

Where, è is angle between the vector X and Y. Using this logic, r_{tet} can also be calculated.

$$r_{tet} = \cos\left[\frac{180^0}{1+\sqrt{\frac{ad}{bc}}}\right] \qquad \text{(eq. 2.6)}$$

Look at the following data summarised in table. 3.

Table 3: Data for Tetrachoric correlation.

		X variable: Attitude towards women		
		0 (Negative attitude)	1 (Positive attitude)	Sum of row
Attitude towards Liberalisation	0 (Negative attitude)	68 (a)	32 (b)	100
	1 (Positive attitude)	30 (c)	70 (d)	100
	Sum of columns	98	102	total =200

The table values are self explanatory. Out of 200 individuals, 68 have negative attitude towards both variables, 32 have negative attitude to liberalisation but positive attitude to women, and so on. The tetrachoric correlation can be computed as follows.

$$r_{tet} = \cos\left[\frac{180^0}{1+\sqrt{\frac{ad}{bc}}}\right] = \cos\left[\frac{180^0}{1+\sqrt{\frac{(68)(70)}{(30)(32)}}}\right] = \cos 55.784^0 = .722$$

So the tetrachoric correlation between attitude towards liberalisation and attitude towards women is positive.

2.7 RANK ORDER CORRELATIONS

We have learned about Pearson’s correlation in earlier unit. The Pearson’s correlation is calculated on continuous variables. Pearson’s correlation is not advised under two circumstances: one, when the data are in the form of ranks and two, when the assumptions of Pearson’s correlation are not followed by the data. In this condition, the application of Pearson’s correlations is doubtful. Under such circumstances, rank-order correlations constitute one of the important options. The ordinal scale data is called as rank-order data. Now let us look at these two aspects, rank-order and assumption of Pearson’s correlations, in greater detail.

2.7.1 Rank-Order Data

When the data is in rank-order format, then the correlation that can be computed is called as rank order correlations. The rank-order data present the ranks of the individuals or subjects. The observations are already in the rank order or the rank order is assigned to them. Marks obtained in the unit test will constitute a continuous data. But if only the merit list of the students is displayed then the data is called as rank order data. If the data is in terms of ranks, then Pearson’s correlation need not be done. Spearman’s rho constitutes a good option.

2.7.2 Assumptions Underlying Pearson's Correlation not Satisfied

The statistical significance testing of the Pearson's correlation requires some assumptions about the distributional properties of the variables. We have already delineated these assumptions in the earlier unit. When the assumptions are not followed by the data, then employing the Pearson's correlation is problematic. It should be noted that small violations of the assumptions does not influence the distributional properties and associated probability judgments. Hence it is called as a robust statistics. However, when the assumptions are seriously violated, then application of Pearson's correlation should no longer be considered as a choice. Under such circumstances, Rank order correlations should be preferred over Pearson's correlation.

It needs to be noted that rank-order correlations are applicable under the circumstances when the relationship between two variables is not linear but still it is a *monotonic* relationship. The *monotonic* relationship is one where values in the data are consistently increasing and never decreasing or consistently decreasing and never increasing. Hence, monotonic relationship implies that as X increases Y consistently increase or as X increases Y consistently decrease. In such cases, rank-order is a better option than Pearson's correlation coefficient.

However, some caution should be observed while doing so. A careful scrutiny of Figure 1 below indicates that, in reality, it is a power function. So actually a relationship between X and Y is not linear but curvilinear power function. So, indeed, curve-fitting is a best approach for such data than using the rank order correlation.

The rank-order can be used with this data since the curvilinear relationship shown in figure 1 is also a *monotonic* relationship. It must be kept in mind that all curvilinear relationships would not be monotonic relationships.

In the previous unit, we have discussed the issue of linearity in the section 1.1.2. I have exemplified the non-linear relationship with Yorkes- Dodson law. It states that relationship between stress and performance is non-linear relationship. But this relationship is NOT a monotonic relationship because initially Y increasers with the corresponding increase in X. But beyond the modal value of X, the scores on Y decrease. So this is not a monotonic relationship. Hence, rank-order correlations should not be Calculated for such a data.

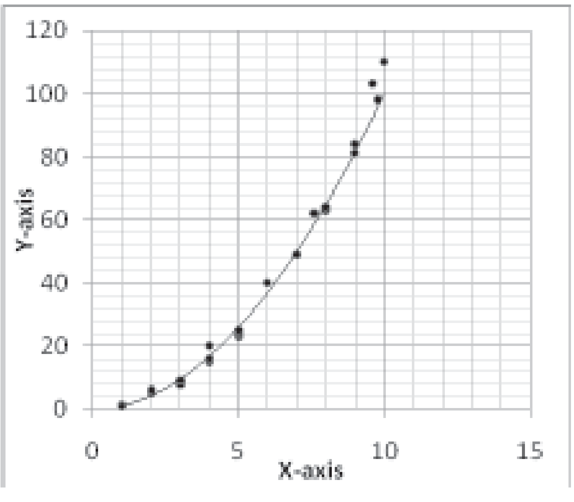


Fig. 1: The figure shows a monotonic relationship between X and Y

2.8 SPEARMAN'S RANK-ORDER CORRELATION OR SPEARMAN'S *RHO* (r_S)

A well-known psychologist and intelligence theorist, Charles Spearman (1904), developed a correlation procedure called in his honor as Spearman's rank-order correlation or Spearman's *rho* (r_s). It was developed to compute correlation when the data is presented on two variables for *n* subjects. It can also be calculated for data of *n* subjects evaluated by two judges for inter-judge agreement. It is suitable for the rank-order data. If the data on X or Y or on both the variables are in rank-order then Spearman's *rho* is applicable. It can also be used with continuous data when the assumptions of Pearson's assumptions are not satisfied. It is used to assess a monotonic relationship.

The range of Spearman's *rho* (r_s) is also from – 1.00 to + 1.00. Like Pearson's correlation, the interpretation of Spearman's *rho* is based on sign of the coefficient and the value of the coefficient.

If the sign of r_s is positive the relationship is positive, if the sign of r_s is negative then the relationship is negative. If the value of r_s is close to zero then relationship is weak, and as the value of r_s approaches to ± 1.00, the strength of relationship increases. When the value of r_s is zero then there is no relationship between X and Y. If r_s is ± 1.00, then the relationship between X and Y is perfect. Whatever the value of r_s may take, it does not directly imply causation. We have already discussed the correlation and causality in the previous unit.

2.8.1 Null and Alternative Hypothesis

The Spearman's *rho* can be computed as a descriptive statistics. We do not carry out statistical hypothesis testing for descriptive use of *rho*. If the r_s is computed as a statistic to estimate population correlation (parameter), then null and alternative hypothesis are required.

The null hypothesis states that

H_O: $\tilde{n}_s = 0$

It means that the value of Spearman's correlation coefficient between X and Y is zero in the population represented by sample.

The alternative hypothesis states that

H_A: $\tilde{n}_s \neq 0$

It means that the value of Spearman's *rho* between X and Y is not zero in the population represented by sample. This alternative hypothesis would require a two-tailed test.

Depending on the theory, the other alternatives could also be written. They are either

H_A: $\tilde{n}_s < 0$

or

H_A: $\tilde{n}_s > 0$.

The first alternative hypothesis, H_A, states that the population value of Spearman's *rho* is smaller than zero. The second H_A denotes that the population value of Spearman's *rho* is greater than zero. Remember, only one of them has to be tested and not both.

You can recall from earlier discussion that one-tailed test is required for this hypothesis.

2.8.2 Numerical Example: for Untied and Tied Ranks

Very obviously, the data on X and Y variables on are required to compute Spearman's *rho*. If the data are on continuous variables then it need to be converted into a rank-orders. The computational formula of Spearman's *rho* (r_s) is as follows:

$$r_s = 1 - \frac{6\sum D^2}{n(n^2 - 1)} \qquad \text{(eq. 2.7)}$$

Where,

r_s = Spearman's rank-order correlation

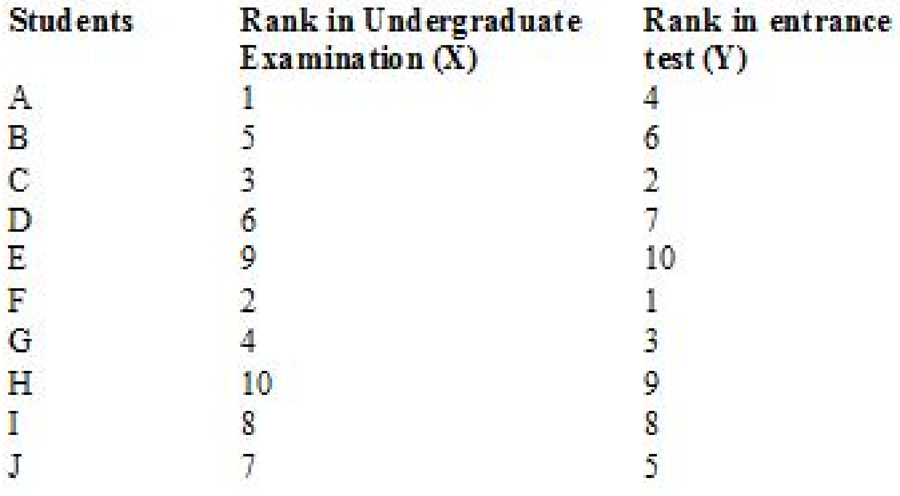
D = difference between the pair of ranks of X and Y

n = the number of pairs of ranks

Steps:

Let's solve an example. We have to appear for entrance examination after the under-graduate studies. We are interested in correlating the undergraduate marks and performance in the entrance test. We have a data of 10 individuals. But we only have ranks of these individuals in undergraduate examination, and merit list of the entrance performance. We want to find the correlation between rank in undergraduate examination and rank in entrance. The data are provided in table 4 and 5. Since this is a rank order data, we can carry out the Spearman's *rho*. (If the data on one or both variable were continuous, we need to transfer this data into ranks for computing the Spearman's *rho*.)

Table 4: Data for Spearman's rho.



Students	Rank in Undergraduate Examination (X)	Rank in entrance test (Y)
A	1	4
B	5	6
C	3	2
D	6	7
E	9	10
F	2	1
G	4	3
H	10	9
I	8	8
J	7	5

The steps for computation of r_s are given below:

Step 1: List the names/serial number of subjects (students, in this case) in column 1.

Step 2: Write the scores of each subject on X variable (undergraduate examination) in the column labeled as X (column 2), and write the scores of each subject on Y variable (Entrance test) in the column labeled as Y (column 3). We will skip this step because we do not have original scores in undergraduate examination and entrance test.

Step 3: Rank the scores of X variable in ascending order. Give rank 1 to the lowest score, 2 to the next lowest score, and so on. In case of our data, the scores are already ranked.

Step 4: Rank the scores of Y variable in ascending order. Give rank 1 to the lowest score, 2 to the next lowest score, and so on. This column is labeled as R_Y (Column 5). Do cross-check your ranking by calculating the sum of ranks. In case of our data, the scores are already ranked.

Step 5: Now find out D, where D = $R_X - R_Y$ (Column 6).

Step 6: Square each value of D and enter it in the next column labeled as D^2 (Column7). Obtain the sum of the D^2. This is written at the end of the column D^2. This $\sum D^2$ is 18 for this example.

Step 7: Use the equation 4.1 (given below) to compute the correlation between rank in undergraduate examination and rank in entrance test.

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} \qquad \text{(eq. 2.8)}$$

Table 5: Table showing the data on rank obtained in undergraduate examination and ranks in entrance examination. It also shows the computation of Spearman's *rho*.

Students	Rank in Undergraduate Examination (X)	Rank in entrance test (Y)	R_X	R_Y	$D = R_X - R_Y$	D^2
A	1	4	1	4	-3	9
B	5	6	5	6	-1	1
C	3	2	3	2	1	1
D	6	7	6	7	-1	1
E	9	10	9	10	-1	1
F	2	1	2	1	1	1
G	4	3	4	3	1	1
H	10	9	10	9	1	1
I	8	8	8	8	0	0
J	7	5	7	5	2	4
n = 10						$\sum D^2=20$

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6\times 20}{10(10^2-1)} = 1 - \frac{180}{990} = 1 - 0.1818 = 0.818$$

Now the Spearman's *rho* has been computed for this example. The value of *rho* is 0.818. This value is positive value. It shows that the correlation between the ranks in undergraduate examination and the ranks in entrance test is positive. It indicates that the relationship between them is positively monotonic. The value of the correlation coefficient is very close to 1.00 which indicates that the strength association between the two set of ranks is very high. The tied ranks were not employed in this example since it was the first example. Now I shall introduce you to the problem of tied ranks.

Interesting point need to be noted about the relationship between Pearson's correlation and Spearman's *rho*. The Pearson's correlation on ranks of X and Y (i.e., R_X and R_Y) is equal the Spearman's *rho* on X and Y. That's the relationship between Pearson's *r* and Spearman's *rho*. The Spearman's *rho* can be considered as a special case of Pearson's *r*.

2.8.3 Spearman's *rho* with Tied Ranks

The ranks are called as *tied ranks* when two or more subjects have the same score on a variable. We usually get larger than the actual value of Spearman's *rho* if we employ the formula in the equation 2.1 for the data with the tied ranks. So the formula in equation 2.1 is not appropriate for tied ranks. A correction is required in this formula in order to calculate correct value of Spearman's *rho*. The recommended procedure of correction for tied ranks is computationally tedious. So we shall use a computationally more efficient procedure. The easier procedure of correction actually uses Pearson's formula on the ranks. The formula and the steps are as follows:

$$r = r_s = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sqrt{\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right]\left[\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right]}} \qquad \text{(eq. 2.10)}$$

Where,

r_s = Spearman's *rho*

X = ranks of variable X

Y = rank on variable Y

n = number of pairs

Look at the example we have solved for Pearson's correlation. It is an example of relationship between BHS and BDI. The data is different than the one we have used in the earlier unit. We shall solve this example with Spearman's *rho*.

2.8.4 Steps for r_s with Tied Ranks

If the data are not in ranks, then convert it into rank-order. In this example, we have assigned ranks to X and Y (column 2 and 3) in column 4, and 5.

Appropriately rank the ties (Cross-check the ranking by using sum of ranks check). This is the basic information for the Spearman's *rho*.

Compute the square of rank of X and rank of Y for all the observations. It is in columns 6 and 7.

Multiply the rank of X by rank of Y for each observation. It is provided in column 8.

Obtain sum of all the columns. Now all the basic data for the computation is available.

Enter this data into the formula shown in the equation 2.2 and calculate r_s.

Table 6: Spearman's *rho* for tied ranks

Subject	BHS (X)	BDI (Y)	Rank X	Rank Y	(Rank X)^2	(Rank Y)^2	(Rank X) (Rank Y)
1	7	8	3.5	2.5	12.25	6.25	8.75
2	11	16	6.5	9.5	42.25	90.25	61.75
3	16	14	9	7	81	49	63
4	9	12	5	5.5	25	30.25	27.5
5	6	8	2	2.5	4	6.25	5
6	17	16	10	9.5	100	90.25	95
7	7	9	3.5	4	12.25	16	14
8	11	12	6.5	5.5	42.25	30.25	35.75
9	5	7	1	1	1	1	1
10	14	15	8	8	64	64	64
Sum			55	55	384	383.5	375.75

$$r_s = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sqrt{\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right]\left[\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right]}} = \frac{375.75 - \frac{(55)(55)}{10}}{\sqrt{\left[384 - \frac{55^2}{10}\right]\left[383.5 - \frac{55^2}{10}\right]}} = \frac{73.25}{81.2496} = 0.902$$

The Spearman's *rho* for this example is 0.902. Since this is a positive value, the relationship between them is also positive. This value is rather near to 1.00. So the strength of association between the ranks of BDI and BHS are very high. This is a simpler way to calculate the Spearman's *rho* with tied ranks. Now, we shall look at the issue of significance testing of the Spearman's *rho*.

2.8.5 Significance Testing of Spearman's *rho*

Once the statistics of Spearman's *rho* is calculated, then the significance of Spearman's *rho* need to be found out. The null hypothesis tested is

H_O: $\tilde{n}_s = 0$

It states that the value Spearman's *rho* between X and Y is zero in the population represented by sample.

The alternative hypothesis is

H_A: $\tilde{n}_s \neq 0$

It states that the value Spearman's *rho* between X and Y is not zero in the population represented by sample. This alternative hypothesis requires a two-tailed test. We have already discussed about writing a directional alternative which requires one-tailed test.

We need to refer to Appendix D for significance testing. The appendix in statistics book, provides critical values for one-tailed as well as two-tailed tests. Let us use the table for the purpose of hypothesis testing for the first example of correlation between ranks in undergraduate examination and entrance test (table 2).

The obtained Spearman's *rho* is 0.818 on the sample of 10 individuals. For *n* = 10, and two-tailed level of significance of 0.05, the critical value of r_s = 0.648. The critical value of r_s = 0.794 at the two-tailed significance level of 0.01.

The obtained value of 0.818 is larger than the critical value at 0.01. So the obtained correlation is significant at 0.01 level (two-tailed). We reject the null hypothesis and accept the alternative hypothesis. It indicates that the value of the Spearman's *rho* is not zero in the population represented by the sample.

For the second example (table 3), the obtained r_s value is 0.902 on the sample of 10 individuals. For *n* = 10, the critical value is 0.794 at the two-tailed significance level of 0.01. The obtained value of 0.902 is larger than the critical value at 0.01. So the obtained correlation is significant at 0.01 level (two-tailed). Hence, we reject the null hypothesis and accept the alternative hypothesis.

When the sample size is greater than ten, then the *t*-distribution can be used for computing the significance with *df* = *n* – 2. Following equation is used for this purpose.

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \qquad \text{(eq. 2.10)}$$

For the example shown in table 2, the *t*-value is computed using equation 2.11.

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} = \frac{0.818\sqrt{10-2}}{\sqrt{1-0.818^2}} = 4.022 \qquad \text{(eq.2.11)}$$

At the *df* = 10 – 2 = 8, the critical *t*-value at 0.01 (two-tailed) is 3.355. The obtained *t*-value is larger than the critical *t*-value. Hence, we reject the null hypothesis and accept the alternative hypothesis.

2.9 KENDALL'S *TAU* (τ)

Kendall's *tau* is another useful measure of correlation. It is as an alternative to Spearman's *rho* (r_s).

This correlation procedure was developed by Kendall (1938). Kendall's *tau* is based on an analysis of two sets of ranks, X and Y. Kendall's *tau* is symbolised as ô, which is a lowercase Greek letter *tau*. The parameter (population value) is symbolised as ô and the statistics computed on the sample is symbolised as. The range of *tau* is from – 1.00 to + 1.00. The interpretation of *tau* is based on the sign and the value of coefficient. The *tau* value closer to ±1.00 indicates stronger relationship. Positive value of *tau* indicates positive relationship and vice versa. It should be noted that Kendall's Concordance Coefficient is a different statistics and should not be confused with Kendall's *tau*.

2.9.1 Null and Alternative Hypothesis

When the Kendall's *tau* is computed as a descriptive statistics, statistical hypothesis testing is not required. If the sample statistic $\tilde{\tau}$ is computed to estimate population correlation ô, then null and alternative hypothesis are required.

The null hypothesis states that

H_O: ô = 0

It stated that the value Kendall’s *tau* between X and Y is zero in the population represented by sample.

The alternative hypothesis states that

H_A: ô $\neq$ 0

It states that the value Kendall’s *tau* between X and Y is not zero in the population represented by sample. This alternative hypothesis requires a two-tailed test.

Depending on the theory, the other alternatives could be written. They are either

1) H_A: $\hat{o} < 0$ or

2) H_A: $\hat{o} > 0$.

The first H_A denotes that the population value of Kendall’s *tau* is smaller than zero.

The second H_A denotes that the population value of Kendall’s *tau* is greater than zero. Remember, only one of them has to be tested and not both. One-tailed test is required for these hypotheses.

2.9.2 Logic of τ and Computation

The *tau* is based on concordance and discordance among two sets of ranks. For example, table 4.4 shows ranks of four subjects on variables X and Y as R_X and R_Y. In order to obtain concordant and discordant pairs, we need to order one of the variables according to the ranks, from lowest to highest (we have ordered X in this fashion).

Take a pair of ranks for two subjects A (1,1) and B (2,3) on X and Y.

Now, if sign or the direction of $R_X - R_X$ for subject A and B is similar to the sign or direction of $R_Y - R_Y$ for subject A and B, then the pair of ranks is said to be concordant (i.e., in agreement).

In case of subject A and B, the $R_X - R_X$ is $(1 - 2 = -1)$ and $R_Y - R_Y$ is also $(1 - 3 = -2)$. The sign or direction of A and B pair is in agreement. So pair A and B is called as concordant pair.

Look at second example of B and C pair. The $R_X - R_X$ is $(2 - 3 = -1)$ and $R_Y - R_Y$ is also $(3 - 2 = +1)$. The sign or the direction of B and C pair is not in agreement. This pair is called as discordant pair.

Table 7: Small data example for *tau* on four subjects

Subject	R_X	R_Y
A	1	1
B	2	3
C	3	2
D	4	4

How many such pair we need to evaluate? They will be $n(n-1)/2 = (4 \times 3)/2 = 6$, so six pairs. Here is an illustration: AB, AC, AD, BC, BD, and CD. Once we know the concordant and discordant pairs, then we can calculate by using following equation.

$$\tilde{\tau} = \frac{n_C - n_D}{\left[\frac{n(n-1)}{2}\right]} \qquad \text{(eq. 2.13)}$$

Where,

$\tilde{\tau}$ = value of ô obtained on sample

n_C = number of concordant pairs

n_D = number of discordant pairs

n = number of subjects

Now, I illustrate a method to obtain the number of concordant (n_C) and discordant (n_D) pairs for this small data in the table above. We shall also learn a computationally easy method later.

Step 1. First, Ranks of X are placed in second row in the ascending order.

Step 2. Accordingly ranks of Y are arranged in the third row.

Step 3. Then the ranks of Y are entered diagonally.

Step 4. Start with the first element in the diagonal which is 1 (row 4).

Step 5. Now move across the row.

Step 6. Compare it (1) with each column element of Y. If it is smaller then enter C in the intersection. If it is larger, then enter D in the intersection. For example, 1 is smaller than 3 (column 3) so C is entered.

Step 7. In the next row (row 5), 3 is in the diagonal which is greater than 2 (column 4) of Y, so D is entered in the intersection.

Step 8. Then “C and “D are computed for each row.

Step 9. The n_C is obtained from $\sum\sum$C (i.e., 5) and

Step 10. n_D is obtained from $\sum\sum$D (i.e., 1).

Step 11. These values are entered in the equation 4.4 to obtain.

Table 8. Computation of concordant and discordant pairs.

Subjects	A	B	C	D	∑C	∑D
Rank of X	1	2	3	4		
Rank of Y	1	3	2	4		
	1	C	C	C	3	0
		3	D	C	1	1
			2	C	1	0
				4	0	0
					∑∑C = 5	∑∑D = 1

$$\tilde{\tau} = \frac{n_C - n_D}{\left[\frac{n(n-1)}{2}\right]} = \frac{5-1}{\left[\frac{4(4-1)}{2}\right]} = \frac{4}{6} = 0.667$$

2.9.3 Computational Alternative for τ

This procedure of computing the *tau* is tedious. I suggest an easier alternative. Suppose, we want to correlate rank in practice sessions and rank in sports competitions. We also know the ranks of the sportspersons on both variables. The data are given below for 10 sportspersons.

Table 9: Data of 10 subjects on X (rank in practice session) and Y (ranks in sports competition)

	Subjects being ranked									
	A	B	C	D	E	F	G	H	I	J
Practice session (Ranks on X)	1	2	3	4	5	6	7	8	9	10
Sports competition (Ranks on Y)	2	1	5	3	4	6	10	8	7	9

First we arrange the ranks of the students in ascending order (in increasing order; begin from 1 for lowest score) according to one variable, X in this case. Then we arrange the ranks of Y as per the ranks of X. I have drawn the lines to connect the comparable ranking of X with Y. Please note that lines are not drawn if the subject gets the same rank on both the variables. Now we calculate number of inversions. Number of inversions is number of intersection of the lines. We have five intersections of the lines.

So the following equation can be used to compute $\tilde{\tau}$

$$\tilde{\tau} = 1 - \frac{2(n_s)}{\frac{n(n-1)}{2}} \qquad \text{(eq. 2.14)}$$

Where

$\tilde{\tau}$ = sample value of ô

n_s = number of inversions

n = number of subjects

$$\tilde{\tau} = 1 - \frac{2(n_s)}{\frac{n(n-1)}{2}} = 1 - \frac{2(5)}{\frac{10(10-1)}{2}} = 1 - \frac{10}{45} = 1 - 0.222 = 0.778$$

The value of Kendall's *tau* for this data is 0.778. The value is positive. So the relationship between X and Y is positive. This means as the rank on time taken increases the rank on subject increases. Interpretation of *tau* is straightforward. For example, if the $\tilde{\tau}$ is 0.778, then it can be interpreted as follows: if the pair of subjects is sampled at random, then the probability that their order on two variables (X and Y) is similar is 0.778 higher than the probability that it would be in reverse order. The calculation of *tau* need to be modified for tied ranks. Those modifications are not discussed here.

2.9.4 Significance Testing of ô

The statistical significance testing of Kendall's *tau* is carried out by using either Appendix E and referring to the critical value provided in the Appendix E. The other way is to use the z transformation. The z can be calculated by using following equation

$$z = \frac{\tilde{\tau}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} \qquad \text{(eq. 2.15)}$$

You will realise that the denominator is the standard error of *tau*. Once the Z is calculated, you can refer to Appendix A for finding out the probability.

For our example in table 4, the value of $\tilde{\tau}$ = 0.664 for the *n* = 4. The Appendix E provides the critical value of 1.00 at two-tailed significance level of 0.05. The obtained value is smaller than the critical value. So it is not statistically significant. Hence, we retain the null hypothesis which states H_O: ô = 0. So we accept this hypothesis. It implies that the underlying population represented by the sample has no relationship between X and Y.

For example in table 6, the obtained value of *tau* is 0.778 with the *n* = 10. From the Appendix E, for the *n* = 10, the critical value of *tau* is 0.644 at two-tailed 0.01 level of significance. The value obtained is 0.778 which is higher than the critical value of 0.664. So the obtained value of *tau* is significant at 0.01 level. Hence, we reject the null hypothesis H_O: ô = 0 and accept the alternative hypothesis H_A: ô ≠ 0. It implies that the value of *tau* in the population represented by sample is other than zero. So there exists a positive relationship between practice ranks and sports competition ranks.

Other way of testing significance is to convert the obtained value of the tau into z. Then use the z distribution for testing the significance of the tau. For this purpose, following formula can be used.

$$z = \frac{\tilde{\tau}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} = \frac{0.778}{\sqrt{\frac{2(2\times10+5)}{9\times10(10-1)}}} = 3.313$$

The z table (normal distribution table) in the Appendix A has a value of z = 1.96 at 0.05 level and 2.58 at 0.01 level. The obtained value of z = 3.313 is far greater than these values. So we reject the null hypothesis at 0.01 level of significance.

Kendall's *tau* is said to better alternative to Spearman's *rho* under the conditions of tie ranks. The *tau* is also supposed to do better than Pearson's *r* under the conditions of extreme non-normality. This holds true only under the conditions of very extreme cases. Otherwise, Pearson's *r* is still a coefficient of choice.

2.10 LET US SUM UP

In this unit, we have learned the specific types of correlations that can be used under circumstances that are special. These correlations are either Pearsons correlations with different names or non-Pearson correlations. We have also learned to compute the values as well as test the significances of these correlations. We have also learned the coorelations that can be calculated for the ordinal data. They are Spearman's rho and tau. Indeed we also got to know that Spearman's rho can be considered as a special case of Pearson's correlation. The tau is useful under ties ranks. This will help you to handle the correlation data of various types.

2.11 UNIT END QUESTIONS

1) What are the special types of correlations and why are they to be used?

2) Discuss the point biserial correlation and indicate its advantages.

3) Calculate point biserial for the following data:

Subject	**Sex (male) X**	**Marks (Y)**	**Subject**	**Sex (Female (X)**	**Marks (Y)**
1	0	30	11	1	38
2	0	56	12	1	69
3	0	68	13	1	78
4	0	48	14	1	58
5	0	52	15	1	55
6	0	80	16	1	89
7	0	78	17	1	82
8	0	72	18	1	85
9	0	55	19	1	73
10	0	48	20	1	62

4) How will you do the significance of testing for point biserial correlation

5) When do we use Phi Coefficient?

6) Calculate phi coefficient for the following data

X: Gender	0= Male 1 = Female
Y: Ownership of Property	0=No ownership 1 = Ownership

X	1	0	1	1	0	1	1	0	0	1	1	0
Y	1	1	0	0	1	0	0	1	1	0	1	1

7) What is biserial correlation? When do we use biserial correlation?

8) Discuss the use of Tetrachoric correlation.

9) What are the important assumptions of rank order correlation?

10) Discuss in detail Spearman's Rank Correlation and compare it with Kendall's tau.

11) Calculate Rho for the following data and test the significance of Rho

Students	**Marks in history**	**Marks in English**
A	50	60
B	45	48
C	63	72
D	65	76
E	48	58
F	59	60
G	62	68

12) Discuss Kendall's Tau.

13) Discuss the significance testing of Tau.

14) Calculate Tau for the following data

	A	B	C	D	E	F	G	H	I	J
Practice session (Ranks on X)	1	2	3	4	5	6	7	8	9	10
Sports competition (Ranks on Y)	5	1	2	4	4	10	6	7	9	8

2.12 SUGGESTED READINGS

Garrett, H.E. (19). *Statistics In Psychology And Education*. Goyal Publishing House, New Delhi.

Guilford, J.P.(1956). *Fundamental Statistics in Psychology and Education.* McGraw Hill Book company Inc. New York.