Correlation Analysis

SESSION 8

Structure

8.1 INTRODUCTION

In Lab Sessions 6 and 7, you have learnt how to compute measures of central tendency, measures of dispersion, moments, skewness and kurtosis, which analyse data involving a single variable. But in many situations, we are interested in studying the relationship between two or more variables. In Unit 6 of MST-002 (Descriptive Statistics), we have explained that the coefficient of correlation measures the strength and direction of the linear relationship between two variables. We use the Karl Pearson's correlation coefficient to measure the extent or degree of relationship between two variables. When we have more than two variables, we use the multiple correlation coefficient to determine the joint effect of two or more variables on a single variable. We use partial correlation to obtain the relationship between two variables, eliminating the effects of other variables, as explained in Units 11 and 12 of MST-002.

Prerequisite

- Lab Sessions 6 and 7 of MSTL-001 (Basic Statistics Lab).
- Units 6, 7, 11 and 12 of MST-002 (Descriptive Statistics).

You have also learnt in Unit 7 of MST-002 that the rank correlation is applied to measure the association between two ordinal (or rank) variables.

In this lab session, you will learn how to determine the simple and rank correlation coefficients for two variables using MS Excel 2007. You will also learn how to compute multiple and partial correlation coefficients for more than two variables using Excel 2007.

Objectives

After performing the activities of this session, you should be able to:

- prepare the spreadsheet in MS Excel 2007;
- compute the simple correlation coefficient;
- determine the multiple and partial correlation coefficients; and
- obtain the rank correlation coefficient.

8.2 PROBLEM DESCRIPTION

1. Suppose we are interested in investigating the relationship between the harvested yield of a particular crop (kg/ha), the amount of rainfall (mm) and the amount of fertiliser (kg/ha). The data for 35 years for the three variables are recorded in Table 1.

Table 1: Crop yield data

Yield of a Crop (kg/h)	Rainfall (mm)	Fertiliser (kg/h)
133	304	168
72	247	45
91	270	80
59	70	10
90	264	80
84	112	66
115	167	123
112	257	124
96	117	91
61	131	17
168	297	237
92	281	83
97	130	102
74	143	38
105	176	119
136	158	171
95	201	86
143	266	186
113	193	118
151	283	202
144	228	191
210	378	320
220	259	337
157	284	215
219	233	336
200	249	298
218	275	339
287	287	477
203	232	307
341	503	158
237	393	214
290	305	480
323	418	542
397	501	698
334	497	574

For investigating the relationship between any two or all three of these variables, we need to

i) compute the Pearson's correlation coefficients between (a) the yield of the crop and the amount of rainfall, (b) the yield of the crop and the amount of fertiliser and (c) the amount of rainfall and fertiliser for the given data;

ii) determine the multiple correlation coefficients $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$; and

iii) obtain the partial correlation coefficients $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$.

2. Twenty five students participated in a science talent contest and were assessed in two parts: A and B. Part A was based on written examination while Part B was based on oral quiz. Scores between 0 and 100 were awarded separately for both parts. The scores of each contestant for both parts are recorded in Table 2.

Table 2: Scores of students in the science talent contest

Part A	51	78	68	58	62	65	96	60	76	66	85	95	91	55	81
	99	61	53	98	73	83	84	94	63	80					
Part B	55	67	69	57	60	62	86	73	68	75	81	100	72	59	79
	65	64	53	97	96	85	76	95	50	83					

For investigating the relationship between the students' performance in both parts for the given data, we need to compute the Spearman's rank correlation coefficient between the scores obtained.

3. A survey was conducted to investigate the impact of the average number of hours spent in watching television per day by students on the marks obtained by them in examinations. A sample of fifteen students was taken. The number of hours spent in watching TV and the marks obtained are recorded in Table 3.

Table 3: Time spent in watching TV and marks obtained

S. No.	Marks	TV Viewing Hours
1	60	6
2	72	1
3	64	2
4	71	2
5	77	3
6	65	2
7	62	1
8	78	2
9	65	4
10	72	2
11	73	1
12	79	1
13	67	3
14	65	3
15	61	5

For investigating the impact of viewing TV on marks obtained, we need to compute the Spearman's rank correlation coefficient between the number of hours spent in watching TV and marks obtained in the examination for the given data.

8.3 SIMPLE CORRELATION COEFFICIENT

You have learnt how to compute Karl Pearson's correlation coefficient in Unit 6 of MST-002. This is also known as simple or product moment correlation coefficient. It is the most widely used statistical method for determining the degree of linear relationship between two variables. Here we briefly mention the main steps as follows:

Step 1: If X and Y are two random variables considering n observations $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, the covariance between X and Y is given by

$$\text{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad \ldots (1)$$

Step 2: We compute the variances of the variables X and Y as

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \text{ and} \quad \ldots (2)$$

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \ldots (3)$$

Step 3: The correlation coefficient (r) between X and Y is given by

$$r = \frac{\text{Cov}(x, y)}{\sqrt{V(x)\,V(y)}} \quad \ldots (4)$$

Note that the correlation coefficient lies between –1 and +1.

We consider Problem 1 and explain the computation of the correlation coefficient between two variables in Excel 2007. In order to compute Karl Pearson's correlation coefficient, we follow the steps given below:

Step 1: We enter the data of Table 1 in an Excel spreadsheet as shown in Fig. 8.1.

	A	B	C
1	Yield of a crop (X_1) (kg/h)	Rainfall (X_2) (mm)	Fertiliser (X_3) (kg/h)
2	133	304	168
3	72	247	45
4	91	270	80
5	59	70	10
6	90	264	80
7	84	112	66
8	115	167	123
9	112	257	124
10	96	117	91

Correlation

Fig. 8.1: Partial screenshot of the given data in Excel spreadsheet.

Step 2: We determine the correlation coefficient between the yield of the crop and the amount of rainfall, and denote it by r_{12}. To determine Karl Pearson's correlation coefficient, we use the built-in function ***Correl*** of Excel. As shown in Fig. 8.2, we

1. select Cell B38,
2. click on the ***Formulas*** tab, and

3. click on *More Functions* → *Statistical* → *Correl*.

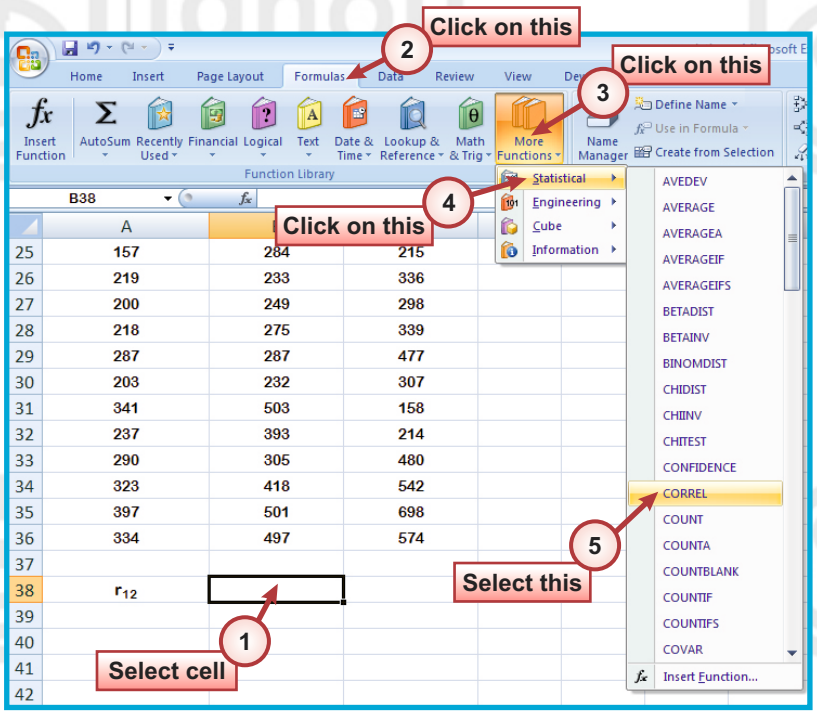
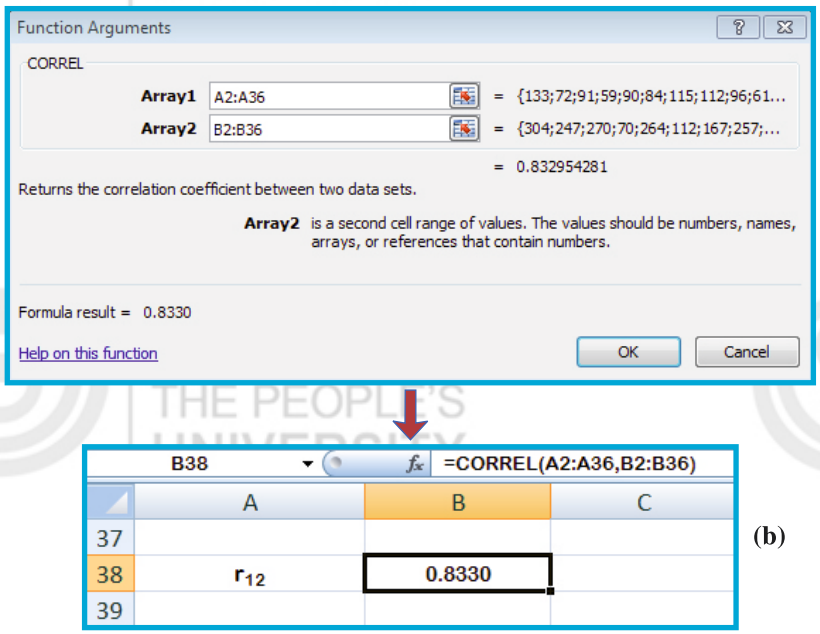


Fig. 8.2

Step 3: A new dialog box opens, which requires the values of the two variables to compute the correlation between them. As shown in Fig. 8.3, we

1. select Cells A2:A36 as *Array 1*,
2. select Cells B2:B36 as *Array 2*,
3. click on *OK* (Fig. 8.3a), and
4. obtain the value of the correlation coefficient in Cell B38 (Fig. 8.3b).



(a)

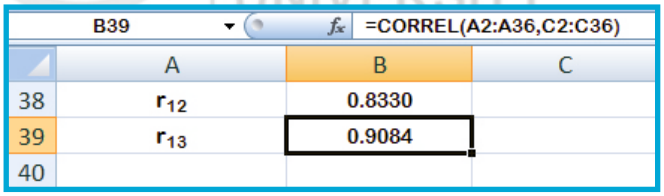
(b)

Fig. 8.3

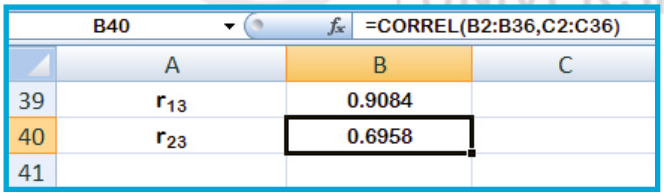
You can directly type "=Correl(A2 : A36, B2: B36)" in Cell B38 to obtain the correlation coefficient between the yield of the crop and the amount of rainfall.

Notice from Fig. 8.3b that the correlation coefficient between the yield of the crop and the amount of rainfall is 0.8330. We can conclude that there is high positive correlation between the yield of the crop and the amount of rainfall for the given data.

Step 4: We now compute the correlation coefficients between (i) the yield of the crop and the amount of fertiliser (r_{13}) and (ii) the amounts of fertiliser and rainfall (r_{23}) in Cells B39 and B40 in the same way as explained in Steps 2 and 3. The values of r_{13} and r_{23} are shown in Figs. 8.4a and 8.4b, respectively.



(a)



(b)

Fig. 8.4

Notice from Fig. 8.4a that the correlation coefficient between the yield of the crop and the amount of fertiliser is 0.9084. It means that the crop yield and the amount of fertiliser are highly positively correlated. However, the correlation coefficient between the amount of fertiliser and rainfall is 0 6958 (Fig 8 4b) It means that the amounts of fertiliser and rainfall are moderately positively correlated for the given data.

8.4 MULTIPLE CORRELATION COEFFICIENTS

You have learnt in Unit 11 of MST-002 that multiple correlation determines the combined influence of two or more variables on a single variable. Here we briefly mention the main steps as under:

Step 1: Suppose, X_1, X_2 and X_3 are three variables having n observations or units. We compute the correlation coefficient using equation (4) between (i) X_1 and X_2 denoted by r_{12}, (ii) X_1 and X_3 denoted by r_{13}, and (iii) X_2 and X_3 denoted by r_{23}.

Step 2: The multiple correlation coefficient of X_1 on X_2 and X_3 is the simple correlation coefficient between X_1 and the joint effect of X_2 and X_3. It is denoted by $R_{1.23}$ and given by

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \quad \ldots (5)$$

where r_{12} – correlation coefficient between variables X_1 and X_2,

r_{23} – correlation coefficient between variables X_2 and X_3, and

r_{13} – correlation coefficient between variables X_1 and X_3.

Step 3: The multiple correlation coefficients $R_{2.13}$ and $R_{3.12}$ are given by

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} \quad \ldots (6)$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}} \quad \ldots (7)$$

Here we describe how to compute multiple correlation coefficients using Excel 2007. We continue with the same Excel sheet prepared in Sec. 8.3 to compute the multiple correlation coefficients and follow the steps given below:

Step 1: The multiple correlation coefficient of the crop yield (X_1) on rainfall (X_2) and fertiliser (X_3) is the simple correlation coefficient between the crop yield and the joint effect of rainfall and fertiliser. It is denoted by $R_{1.23}$. We use equation (5) to compute the multiple correlation coefficient and type "=Sqrt((B38*B38+B39*B39-2*B38*B39*B40)/(1-B40*B40))" in Cell B41 as shown in Fig. 8.5a. When we press ***Enter***, we get the output shown in Fig. 8.5b.

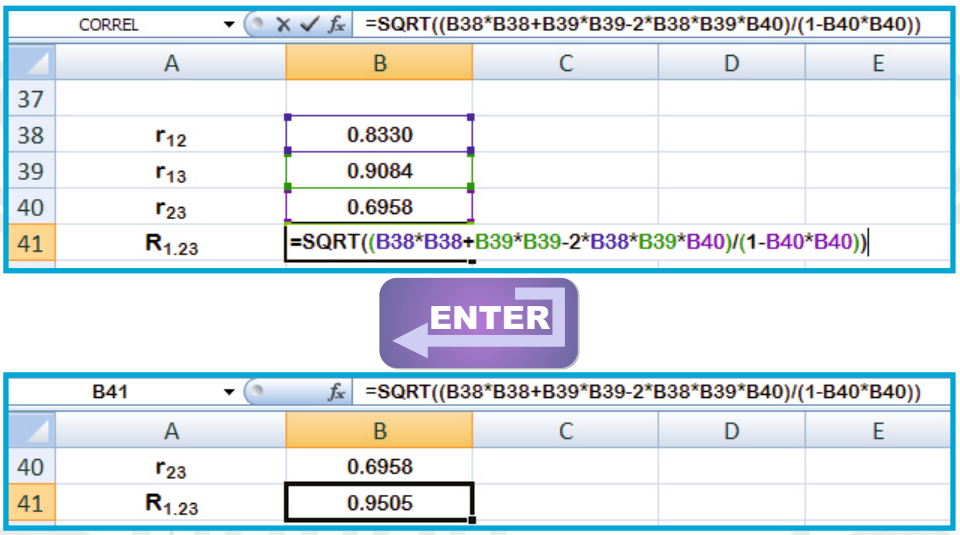


Fig. 8.5

Notice from Fig. 8.5b that the multiple correlation coefficient of the crop yield on the amounts of rainfall and fertiliser is 0.9505. It means that the amounts of rainfall and fertiliser are highly positively correlated with the crop yield for the given data.

Step 2: We follow Step 1 to compute the multiple correlation coefficient of amount of rainfall (X_2) on the crop yield (X_1) and the amount of fertiliser (X_3), i.e., $R_{2.13}$. We use equation (6) and type "=Sqrt((B38*B38+B40* B40-2*B38*B39*B40)/(1-B39*B39))" in Cell B42 as shown in Fig. 8.6.

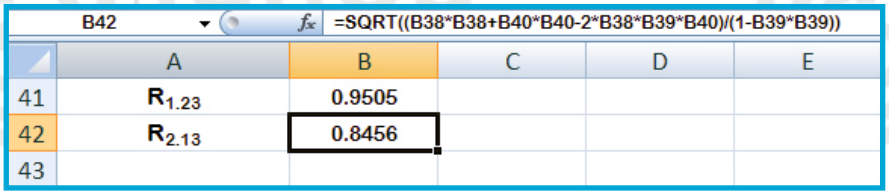


Fig. 8.6

Fig. 8.6 reveals that the multiple correlation coefficient of the amount of rainfall on the crop yield and the amount of fertiliser is 0.8456. It means that the yield of the crop and the amount of fertiliser are highly positively correlated with the amount of rainfall for the given data.

Step 3: We compute the multiple correlation coefficient of the amount of

fertiliser (X_3) on the crop yield (X_1) and the amount of rainfall (X_2), i.e., $R_{3.12}$. We use equation (7) and type "=Sqrt((B39*B39+B40*B40-2*B38*B39*B40)/(1-B38*B38))" in Cell B43 as shown in Fig. 8.7.

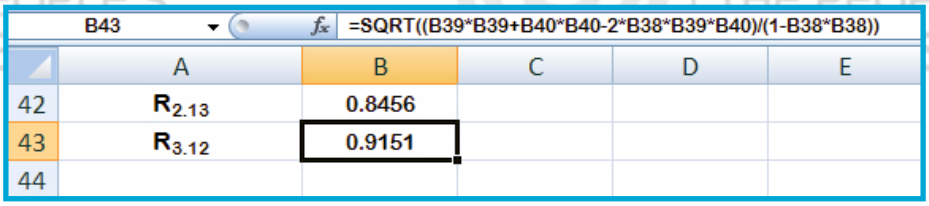


Fig. 8.7

Notice from Fig. 8.7 that the multiple correlation coefficient of the amount of fertiliser on the yield of the crop and the amount of rainfall is 0.9151. It means that the yield of the crop and the amount of rainfall are highly positively correlated with the amount of fertiliser for the given data.

8.5 PARTIAL CORRELATION COEFFICIENTS

You have learnt in Unit 12 of MST-002 that many-a-times, the correlation between two variables is only partly due to the third variable. In such situations, we obtain the relationship between two variables ignoring the effect of the third variable. This is known as partial correlation. Here we briefly mention the main steps as follows:

Step 1: Suppose, X_1, X_2 and X_3 are three variables having n observations or units. We compute the correlation coefficient using equation (4) between (i) X_1 and X_2 denoted by r_{12}, (ii) X_1 and X_3 denoted by r_{13}, and (iii) X_2 and X_3 denoted by r_{23}.

Step 2: The partial correlation is the correlation between two variables, after eliminating the linear effects of other variables on them. The correlation coefficient between X_1 and X_2 after eliminating the linear effect of X_3 on X_1 and X_2 is called the partial correlation coefficient. It is denoted by $r_{12.3}$ and given by

$$r_{12.3} = \frac{(r_{12} - r_{13}r_{23})}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \quad \ldots (8)$$

Step 3: The partial correlation coefficients $r_{13.2}$ and $r_{23.1}$ are given by

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} \quad \ldots (9)$$

and

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}} \quad \ldots (10)$$

We now describe how to compute the partial correlation coefficients using Excel 2007 and continue with the Excel sheet prepared in Sec. 8.4. In order to compute the partial correlation coefficients, we follow the steps given below:

Step 1: The partial correlation coefficient of the yield of the crop (X_1) and the amount of rainfall (X_2) after eliminating the linear effect of the amount of fertiliser (X_3) is denoted by $r_{12.3}$. We use equation (8) to

compute the partial correlation coefficient and type "=(B38-B39*B40)/(Sqrt((1-B39*B39)*(1-B40*B40)))" in Cell B44 as shown in Fig. 8.8a. When we press ***Enter***, we get the output shown in Fig. 8.8b.

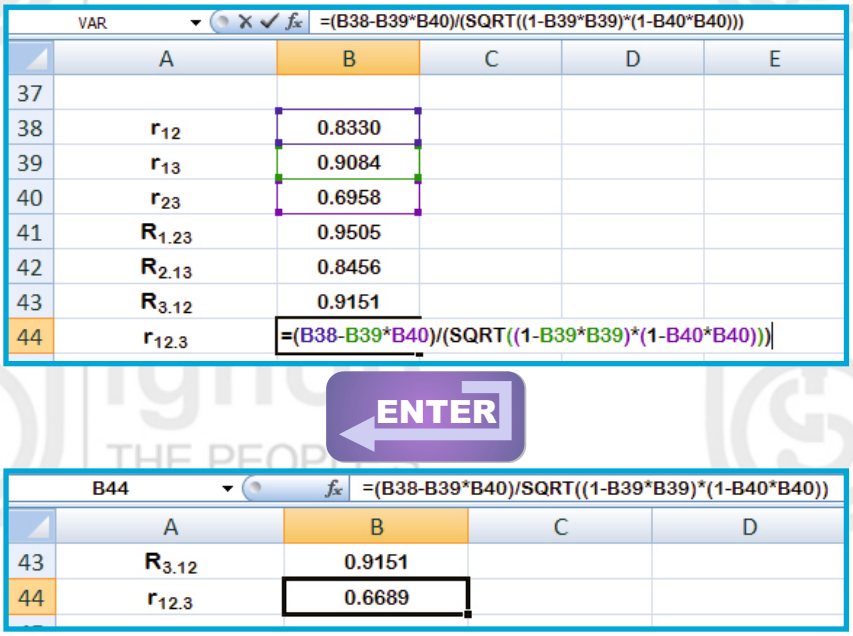


Fig. 8.8

Notice from Fig. 8.8b that the partial correlation coefficient between the yield of the crop and the amount of rainfall after eliminating the effect of the amount of fertiliser is 0.6689. It means that the yield of the crop and the amount of rainfall are moderately positively correlated after eliminating the effect of the amount of fertiliser for the given data.

Step 2: We follow Step 1 to compute the partial correlation coefficient of the yield of the crop (X_1) and the amount of fertiliser (X_3) after eliminating the linear effect of the amount of rainfall (X_2) denoted by $r_{13.2}$. We use equation (9) and type "=(B39-B38*B40)/(Sqrt((1-B38*B38)*(1-B40*B40)))" in Cell B45 (Fig. 8.9).

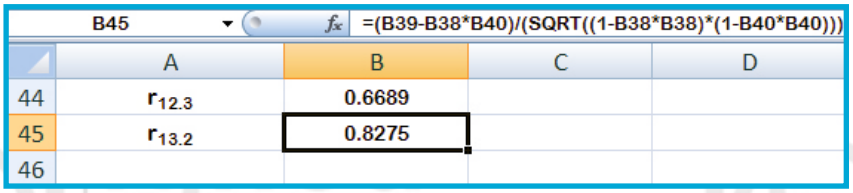


Fig. 8.9

Fig. 8.9 reveals that the partial correlation coefficient between the yield of the crop and the amount of fertiliser after eliminating the effect of the amount of rainfall is 0.8275. It means that the yield of the crop and the amount of the fertiliser are highly positively correlated after eliminating the effect of the amount of rainfall for the given data.

Step 3: We compute the partial correlation coefficient $r_{23.1}$ of the amounts of rainfall (X_2) and fertiliser (X_3) after eliminating the linear effect of the yield of the crop (X_1). We use equation (10) and type "=(B40-B38*B39)/(Sqrt((1-B38*B38)*(1-B39*B39)))" in Cell B46 as shown in Fig. 8.10.

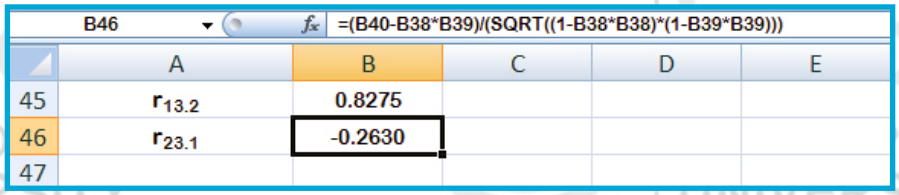


Fig. 8.10

Notice from Fig. 8.10 that the partial correlation coefficient between the amounts of fertiliser and rainfall after eliminating the effect of the yield of the crop is – 0.2630. It means that the amounts of fertiliser and rainfall are poorly negatively correlated after eliminating the effect of the yield of the crop for the given data.

8.6 CORRELATION COEFFICIENTS USING DATA ANALYSIS TOOLPAK

We can also use the ***Data Analysis ToolPak*** for determining Karl Pearson's correlation coefficient in Excel. It provides the correlation matrix among the variables. To obtain the correlation matrix for Problem 1, we use data on the Excel spreadsheet entered in Step 1 of Sec. 8.3. We follow the steps given below:

Step 1: As shown in Fig. 8.11, we

1. click on the ***Data*** tab,
2. select ***Data Analysis*** (Fig. 8.11a),
3. select the ***Correlation*** option,
4. click on ***OK*** (Fig. 8.11b), and
5. get a new dialog box (Fig. 8.11c).

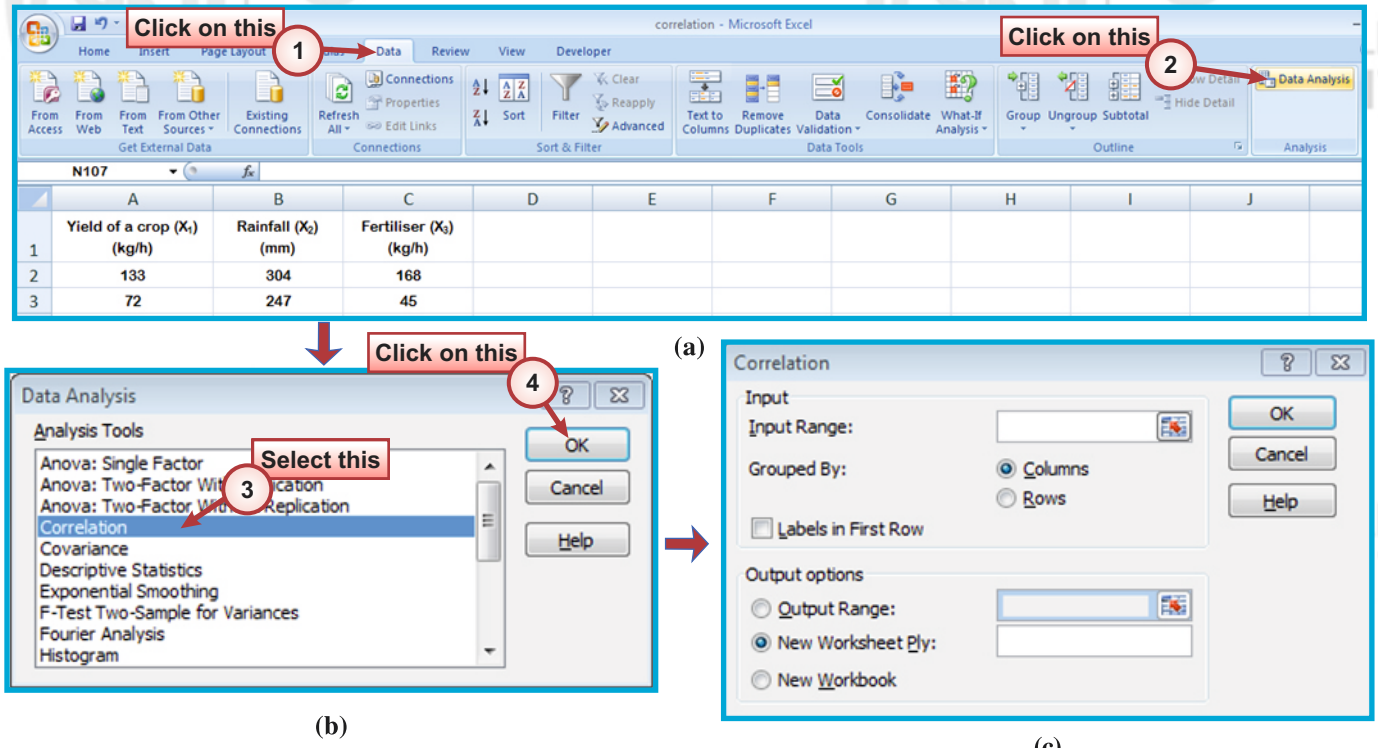


Fig. 8.11

Step 2: As shown in Fig. 8.12, we

1. specify the entire data with labels in ***Input Range***, i.e., Cells A1:C36,
2. tick on the box ***Labels in first row*** since we have included data labels given in Cells A1:C1 in the ***Input Range***,

3. provide the ***Output Range*** under ***Output Options***. Here we select Cell A48 for the output, and
4. tick on the ***Columns*** box under the ***Grouped by*** option as we have different variables in Columns A, B and C and then click on ***OK***.

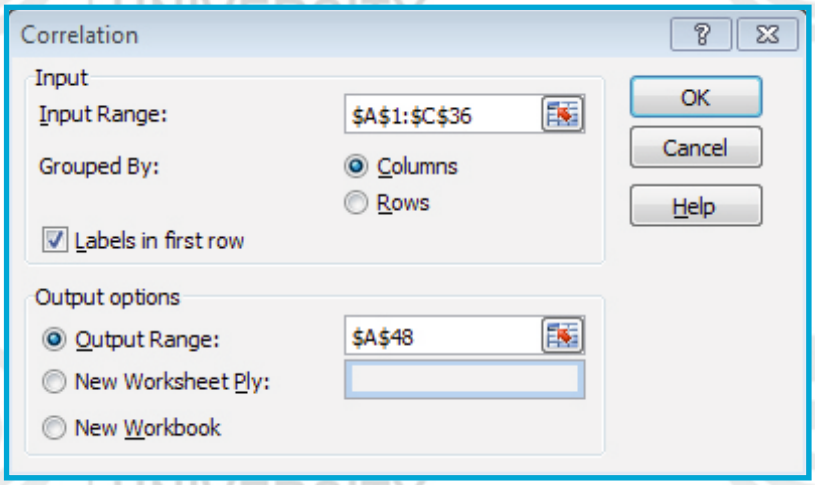


Fig. 8.12

Step 3: After completing Step 2, we obtain the output shown in Fig. 8.13.

	A	B	C	D	E
38	r_{12}	0.8330			
39	r_{13}	0.9084			
40	r_{23}	0.6958			
41	$R_{1.23}$	0.9505			
42	$R_{2.13}$	0.8456			
43	$R_{3.12}$	0.9151			
44	$r_{12.3}$	0.6689			
45	$r_{13.2}$	0.8275			
46	$r_{23.1}$	-0.2630			
47					
48		Yield of a crop (X1)	Rainfall (X2)	Fertiliser (X3)	
49	Yield of a crop (X1)	1			
50	Rainfall (X2)	0.8330	1		
51	Fertiliser (X3)	0.9084	0.6958	1	
52					

Fig. 8.13

Notice from Fig. 8.13 that the ***Correlation*** option under the ***Data Analysis ToolPak*** provides the following information:

- ✓ the correlation coefficient between the yield of the crop and the amount of rainfall is 0.8330 (Cell B50).
- ✓ the correlation coefficient between the yield of the crop and the amount of fertiliser is 0.9084 (Cell B51).
- ✓ the correlation coefficient between the amount of fertiliser and rainfall is 0.6958 (Cell C51).

Note that the correlation coefficients computed by ***Correlation*** under the ***Data Analysis ToolPak*** in Cells B50, B51 and C51 are the same as computed by the ***Correl*** function in Cells B38, B39 and B40 (Fig. 8.13).

8.7 RANK CORRELATION COEFFICIENT

You have studied in Unit 7 of MST-002 that the rank correlation coefficient was given by Spearman. Hence, it is known as Spearman's rank correlation coefficient. We denote the rank correlation coefficient by r_s. In this section, we assume that no two (or more) values of the variables have the same rank for either one or both variables. Here we mention the main steps as follows:

Step 1: Suppose we have n observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ corresponding to the variables X and Y, respectively. We compute the ranks of x_i and y_i denoted by R_{xi} and R_{yi}, respectively.

Step 2: We calculate the difference between the ranks of the observation of the variables as

$$d_i = R_{xi} - R_{yi} \qquad \ldots (11)$$

Step 3: Spearman's rank correlation coefficient when two or more values of the variables do not have the same rank in either one or both the variables is given by

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \ldots (12)$$

Here we describe how to compute the rank correlation coefficient using Excel 2007. We first consider Problem 2 and follow the steps given below:

Step 1: We enter the data of Table 2 in an Excel spreadsheet as shown in Fig. 8.14.

	A	B	C
1	S. No.	Part A	Part B
2	1	51	55
3	2	78	67
4	3	68	69
5	4	58	57
6	5	62	60
7	6	65	62
8	7	96	86
9	8	60	73
10	9	76	68
11	10	66	75
12	11	85	81
13	12	95	100

Rank Correlation

Fig. 8.14 : Partial screenshot of the data entered in Excel spreadsheet.

Step 2: We compute the rank for the marks given in Cell B2 for Part A of the first student. As shown in Fig. 8.15, we

1. select Cell D2,
2. click on the ***Formulas*** tab, and
3. click on ***More Functions*** → ***Statistical*** → ***Rank***.

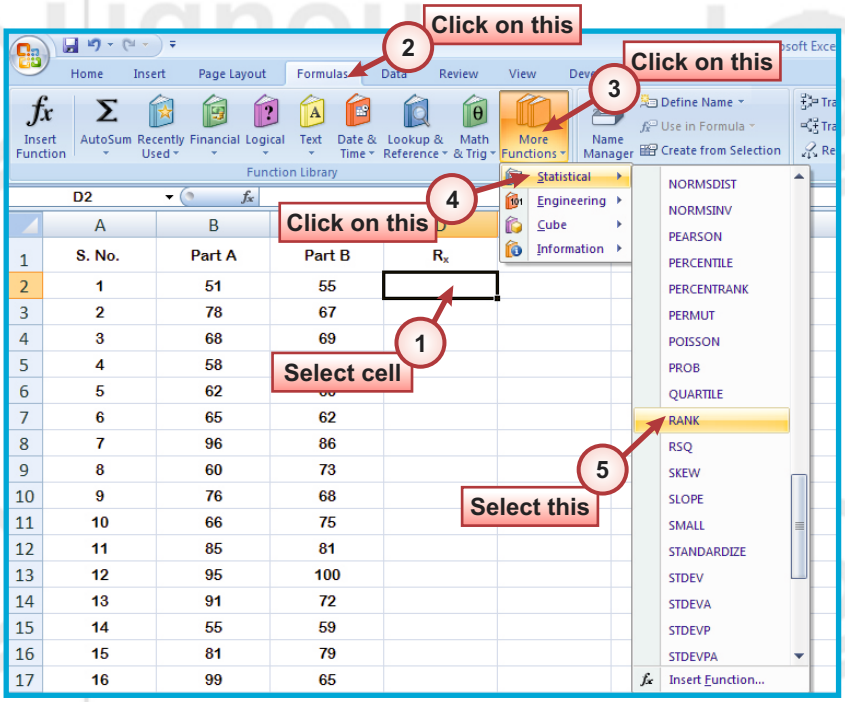


Fig. 8.15

Step 3: A new dialog box opens. As shown in Fig. 8.16a, we

- ✓ select Cell B2 as ***Number*** to compute the rank for the marks of the first student in Part A entered in Cell B2,
- ✓ select Cells B2:B26 as ***Ref***. It means the reference of the data, i.e., the marks of the Part A. It assigns ranks on that basis,
- ✓ type "1" in ***Order*** since we want ranks in ascending order (if we want the ranks in descending order, we type "0" or leave it blank), and
- ✓ click on ***OK*** (Fig. 8.16a). The value of rank for the marks given for part A to the first student is shown in Fig. 8.16b.

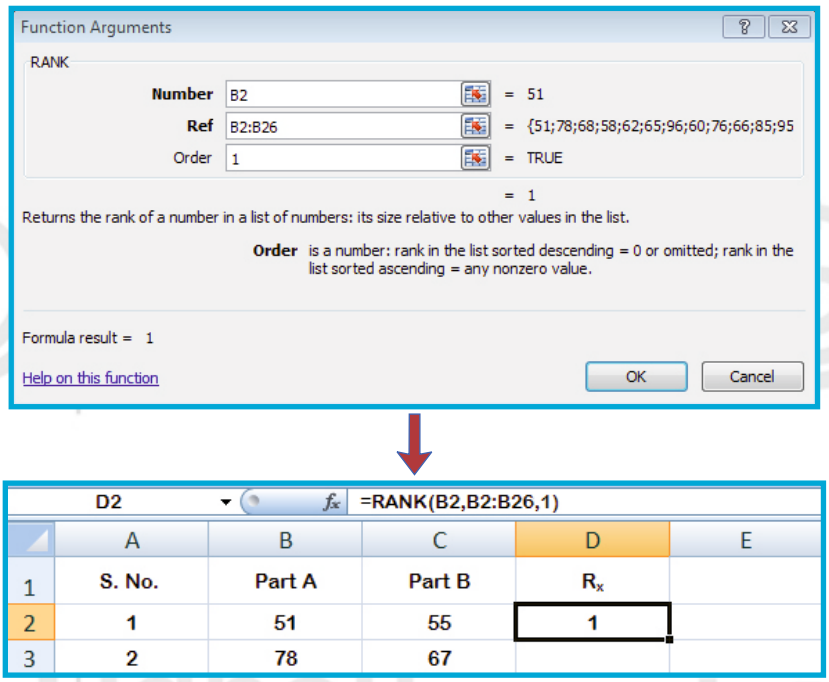


Fig. 8.16

Step 4: We now fix the range given in ***Ref***, i.e., B2:B26 by putting the "$" sign, i.e., $B$2:$B$26 as shown in Fig. 8.17. Note that it helps us in copying the formula in Cells D3:D26 as the reference range is fixed for Part A.

D2 =RANK(B2,$B$2:$B$26,1)

	A	B	C	D	E
1	S. No.	Part A	Part B	R_x	
2	1	51	55	1	
3	2	78	67		
4	3	68	69		
5	4	58	57		

Fig. 8.17

Step 5: We follow Steps 2 and 3 to compute the rank for the marks given for Part B (Cell C2) to the first student in Cell E2. We also fix the range as explained in Step 4. The output is shown in Fig. 8.18.

E2 =RANK(C2,$C$2:$C$26,1)

	B	C	D	E	F
1	Part A	Part B	R_x	R_y	
2	51	55	1	3	
3	78	67			
4	68	69			
5	58	57			

Fig. 8.18

Step 6: We compute the value of the difference between the ranks obtained for Parts A and B in Cells D2 and E2, respectively, i.e., we compute $d = R_x - R_y$ by typing "=D2-E2" in Cell F2 as shown in Fig. 8.19a. We get the output when we press ***Enter*** (Fig. 8.19b).

(a)

MINVERSE =D2-E2

	C	D	E	F	G
1	Part B	R_x	R_y	$d = R_x - R_y$	
2	55	1	3	=D2-E2	
3	67				
4	69				
5	57				

ENTER

(b)

F2 =D2-E2

	C	D	E	F	G
1	Part B	R_x	R_y	$d = R_x - R_y$	
2	55	1	3	-2	
3	67				
4	69				
5	57				

Fig. 8.19

Step 7: We compute the square of the difference between the two ranks, i.e., d^2, by typing "=F2*F2" in Cell G2 and pressing ***Enter*** as shown in Fig. 8.20.

RANK | =F2*F2

	E	F	G	H
1	R_y	$d = R_x - R_y$	d^2	
2	3	-2	=F2*F2	
3				

(a)

ENTER

G2 | =F2*F2

	F	G	H
1	$d = R_x - R_y$	d^2	
2	-2	4	
3			

(b)

Fig. 8.20

Step 8: We now select Cells D2:G2 and drag them down up to Row 26 to determine the required values for the remaining students as shown in Fig. 8.21.

	C	D	E	F	G	H
1	Part B	R_x	R_y	$d = R_x - R_y$	d^2	
2	55	1	3	-2	4	
3	67					
4	69					
5	57					

DRAG THEM DOWN

(a)

	C	D	E	F	G	H
1	Part B	R_x	R_y	$d = R_x - R_y$	d^2	
2	55	1	3	-2	4	
3	67	14	10	4	16	
4	69	11	12	-1	1	
5	57	4	4	0	0	
6	60	7	6	1	1	
7	62	9	7	2	4	
8	86	23	21	2	4	
9	73	5	14	-9	81	
10	68	13	11	2	4	
11	75	10	15	-5	25	
12	81	19	18	1	1	
13	100	22	25	-3	9	

(b)

Fig. 8.21

Step 9: We now compute the value of $\sum d^2$ using "=Sum(G2:G26)" function in Cell G27 as shown in Fig. 8.22.

G27 | =SUM(G2:G26)

	E	F	G	H
26	19	-4	16	
27		Σd^2	664	
28				

Fig. 8.22

Step 10: We type the value of n (= 25), i.e., the number of students who participated in the contest in Cell G28 as shown in Fig. 8.23.

	F	G	H
27	Σd^2	664	
28	n	25	
29			

Fig. 8.23

Step 11: The formula for computing the rank correlation coefficient is given by equation (12). We type "=1-((6*G27)/(G28*(G28*G28-1)))" in Cell G29 as shown in Fig. 8.24a and press ***Enter***. The output is shown in Fig. 8.24b.

(a)

RANK =1-((6*G27)/(G28*(G28*G28-1)))

	E	F	G	H	I
26	19	-4	16		
27		Σd^2	664		
28		n	25		
29		r_s	=1-((6*G27)/(G28*(G28*G28-1)))		
30					

ENTER

(b)

G29 =1-((6*G27)/(G28*(G28*G28-1)))

	E	F	G	H	I
27		Σd^2	664		
28		n	25		
29		r_s	0.7446		
30					

Fig. 8.24

Notice from Fig. 8.24 that the rank correlation coefficient between the marks of Parts A and B is 0.7446. We can conclude that there is high positive correlation between the marks obtained in both parts for the given data.

Step 12: Note that we can also compute Spearman's rank correlation coefficient by using ***Correl*** function since Spearman's rank correlation coefficient [equation (12)] is the same as Karl Pearson's correlation coefficient [equation (4)] computed for the ranks of both the variables instead of their values. We use ***Correl*** function (as explained in Sec. 8.3) in Cell G30. We select Cells D2:D26 as ***Array 1***, Cells E2:E26 as ***Array 2***, and click on ***OK*** to obtain the value of the correlation coefficient in Cell G30 (Fig. 8.25).

G30 =CORREL(D2:D26,E2:E26)

	E	F	G	H
29		r_s	0.7446	
30	Karl Pearson	r_s	0.7446	
31				

Fig. 8.25

Notice from Fig. 8.25 that Karl Pearson's correlation coefficient given in Cell G30 computed from equation (4) is the same as the Spearman's rank correlation coefficient given in Cell G29 computed from equation (12). Hence we can apply ***Correl*** function to compute the Spearman's rank correlation coefficient directly.

In this section, we have explained the computation of the rank correlation coefficient when two or more values of the variables do not have the same rank. But there might be a situation when two or more values have the same rank in either one or both variables. Then the rank is said to be tied. We consider this situation in the next section.

8.8 RANK CORRELATION COEFFICIENT FOR TIED OR REPEATED RANKS

You have studied in Unit 7 of MST-002 that to compute Spearman's rank correlation coefficient, we use the correction factor when two or more values of the variables are the same. Here we mention the main steps as follows:

Step 1: If two or more values of the variables are the same, we assign common ranks to the repeated values. This common rank is the average of ranks that these repeated values would receive if there were no repetition.

Step 2: We use the correction factor $\frac{m(m^2-1)}{12}$ when there is a repetition of ranks in Spearman's rank correlation coefficient given in equation (12).

Step 3: If $(R_{xi})_c$ and $(R_{yi})_c$ are the corrected ranks of x_i and y_i, for $i = 1, 2, \ldots, n$, respectively, the corrected Spearman's rank correlation coefficient is given by

$$(r_s)_c = \frac{\frac{(n^3-n)}{6} - \left(\sum_{i=1}^{n} d_{ci}^2 + T_x + T_y\right)}{\sqrt{\left[\left(\frac{n^3-n}{6} - 2T_x\right)\left(\frac{n^3-n}{6} - 2T_y\right)\right]}} \quad \ldots (13)$$

where

$$T_x = \sum_{i=1}^{p} \frac{(m_{xi}^3 - m_{xi})}{12}, \quad \ldots (14)$$

$$T_y = \sum_{i=1}^{q} \frac{(m_{yi}^3 - m_{yi})}{12}, \quad \ldots (15)$$

m_{xi} – number of times the rank of x_i is repeated,

m_{yi} – number of times the rank of y_i is repeated,

p – number of ranks of variable X, which have repeated values, and

q – number of ranks of variable Y, which have repeated values.

Note that this correction factor is added for every repetition of rank in both variables.

Step 4: In Unit 7 of MST-002, we have used the correction factors only in the numerator. So equation (13) becomes

$$(r_s)_c = 1 - \frac{6\left(\sum_{i=1}^{n} d_{ci}^2 + T_x + T_y\right)}{n(n^2-1)} \quad \ldots (16)$$

Note that here we are using equation (13) because most of the statistical software including Excel uses this formula.

Here, we describe how to compute the rank correlation coefficient for tied or repeated ranks using Excel 2007. We first consider Problem 3. In order to compute the rank correlation coefficient, we follow the steps given below:

Step 1: We enter the data of Table 3 in an Excel spreadsheet and follow Steps 1 to 5 of Sec. 8.7 and drag down Cells D2:E2 up to Row 16 to compute the ranks for the given data. The output is shown in Fig. 8.26.

The value 65 is repeated 3 times and its position will be 5th, 6th and 7th in the data given in Cells B2:B16. The ***Rank*** function of Excel assigns the lowest rank, i.e., 5 to each 65. But we assign the ranks as an average of 5, 6 and 7, i.e., 6 to each 65. So we use a correction factor in Excel to modify the ranks.

	A	B	C	D	E
1	S. No.	Marks	TV Viewing Hour	R_x	R_y
2	1	60	6	1	15
3	2	72	1	10	1
4	3	64	2	4	5
5	4	71	2	9	5
6	5	77	3	13	10
7	6	65	2	5	5
8	7	62	1	3	1
9	8	78	2	14	5
10	9	65	4	5	13
11	10	72	2	10	5
12	11	73	1	12	1
13	12	79	1	15	1
14	13	67	3	8	10
15	14	65	3	5	10
16	15	61	5	2	14

Fig. 8.26

Notice from Fig. 8.26 that the ranks of the first variable (marks), i.e., 5 (highlighted with orange colour) and 10 (highlighted with purple colour) are repeating 3 and 2 times, respectively. Also, the ranks of the second variable (number of hours spent in watching TV), i.e., 1 (highlighted with green colour), 5 (highlighted with yellow colour) and 10 (highlighted with blue colour) are repeating 4, 5 and 3 times, respectively.

Step 2: Excel assigns the lowest rank for all repeated values of the variable. But we assign the average of the ranks to all repeated values. So we add a correction factor to these ranks, i.e., (Count(reference) + 1 – Rank(number, reference, 0) – Rank(number, reference, 1))/2.

Here, number is the value of the variable for which we need to compute rank and reference is all given values of that variable. To compute the corrected rank for the first variable, we type "=D2+(Count($B$2:$B$16)+1-Rank(B2, $B$2:$B$16,0)-Rank(B2,$B$2:$B$16,1))/2" in Cell F2 as shown in Fig. 8.27.

F2 | =D2+(COUNT($B$2:$B$16) + 1 - RANK(B2, $B$2:$B$16, 0)- RANK(B2, $B$2:$B$16, 1))/2

	A	B	C	D	E	F	G	H
1	S. No.	Marks	TV Viewing Hour	R_x	R_y	$(R_x)_c$		
2	1	60	6	1	15	1		
3	2	72	1	10	1			

Fig. 8.27

Step 3: To compute the corrected rank for the second variable, we type "=E2+(Count($C$2:$C$16)+1-Rank(C2,$C$2:$C$16,0)-Rank(C2,$C$2:$C$16,1))/2" in Cell G2 as shown in Fig. 8.28.

G2 =E2+(COUNT($C$2:$C$16) + 1 - RANK(C2, $C$2:$C$16, 0)- RANK(C2, $C$2:$C$16, 1))/2

	A	B	C	D	E	F	G	H
1	S. No.	Marks	TV Viewing Hour	R_x	R_y	$(R_x)_c$	$(R_y)_c$	
2	1	60	6	1	15	1	15	
3	2	72	1	10	1			
4	3	64	2	4	5			

Fig. 8.28

Step 4: We compute the value of the difference between the ranks, i.e., d_c and the square of the difference between the two ranks, i.e., d_c^2 by typing "=F2-G2" and "=H2*H2" in Cells H2 and I2, respectively, as shown in Fig. 8.29.

	F	G	H	I
1	$(R_x)_c$	$(R_y)_c$	$d_c = (R_x)_c - (R_y)_c$	d_c^2
2	1	15	-14	196

Type "=F2-G2" here

Type "=H2*H2" here

Fig. 8.29

Step 5: We now select Cells F2:I2 and drag them down up to Row 16 to determine the required values for the remaining students as shown in Fig. 8.30.

(a)

	F	G	H	I	J
1	$(R_x)_c$	$(R_y)_c$	$d_c = (R_x)_c - (R_y)_c$	d_c^2	
2	1	15	-14	196	
3					
4					

DRAG THEM DOWN

(b)

	F	G	H	I	J
1	$(R_x)_c$	$(R_y)_c$	$d_c = (R_x)_c - (R_y)_c$	d_c^2	
2	1	15	-14	196	
3	10.5	2.5	8	64	
4	4	7	-3	9	
5	9	7	2	4	
6	13	11	2	4	
7	6	7	-1	1	
8	3	2.5	0.5	0.25	
9	14	7	7	49	
10	6	13	-7	49	
11	10.5	7	3.5	12.25	
12	12	2.5	9.5	90.25	
13	15	2.5	12.5	156.25	
14	8	11	-3	9	
15	6	11	-5	25	
16	2	14	-12	144	
17					

Fig. 8.30

Step 6: We compute the value of $\sum d_c^2$ using "=Sum(I2:I16)" function in Cell I17 as shown in Fig. 8.31.

I17 =SUM(I2:I16)

	H	I	J
16	-12	144	
17	Σd_c^2	813	
18			

Fig. 8.31

Step 7: We type the value of n (= 15), i.e., the number of students in Cell I18 (Fig. 8.32). We also type "=(I18^3-I18)/6" in Cell I19 to compute the value of $\frac{(n^3 - n)}{6}$ as shown in Fig. 8.32.

I19 =(I18^3-I18)/6

	H	I	J
17	Σd_c^2	813	
18	n	15	
19	$(n^3-n)/6$	560	
20			

Fig. 8.32

Step 8: We type the values of the corrected tie ranks in Cells A19 and A20 for the first variable and in Cells D19 to D21 for the second variable. We also type the number of times the particular tied rank is repeated in Cells B19 to B20 for the first variable (m_x) and in Cells E19 to E21 for the second variable (m_y) as shown in Fig. 8.33.

	A	B	C	D	E
17					
18	$Tie\ (R_x)_c$	m_x		$Tie\ (R_y)_c$	m_y
19	6	3		2.5	4
20	10.5	2		7	5
21				11	3

Fig. 8.33

Step 9: We type "=(B19^3-B19)/12" in Cell C19 to compute the value of $\frac{(m_x^3 - m_x)}{12}$ and drag down Cell C19 up to Cell C20 (Fig. 8.34).

(a)

C19 =(B19^3-B19)/12

	A	B	C	D
18	$Tie\ (R_x)_c$	m_x	$(m_x^3-m_x)/12$	
19	6	3	2	
20	10.5	2		

DRAG IT DOWN

(b)

	A	B	C	D
18	$Tie\ (R_x)_c$	m_x	$(m_x^3-m_x)/12$	
19	6	3	2	
20	10.5	2	0.5	
21				

Fig. 8.34

Step 10: We type "=(E19^3-E19)/12" in Cell F19 to compute the value of $\frac{(m_y^3 - m_y)}{12}$ and drag Cell F19 down up to Cell F21 (Fig. 8.35).

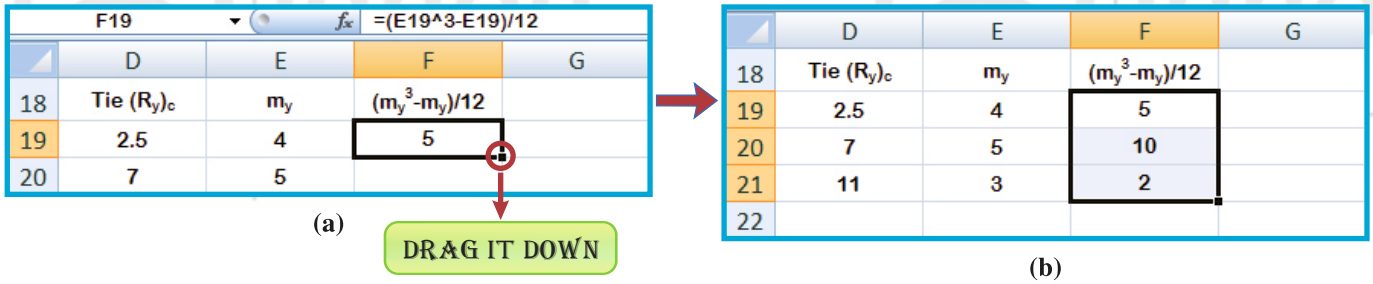


Fig. 8.35

Step 11: We use equation (14) and compute the value of T_x using "=Sum(C19:C20)" function in Cell C22 as shown in Fig. 8.36.

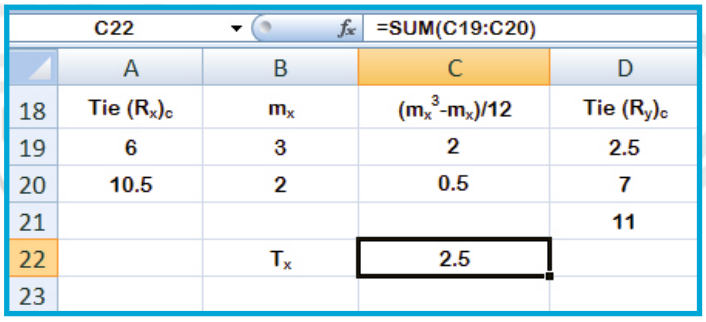


Fig. 8.36

Step 12: We use equation (15) and compute the value of T_y using "=Sum(F19:F21)" function in Cell F22 as shown in Fig. 8.37.

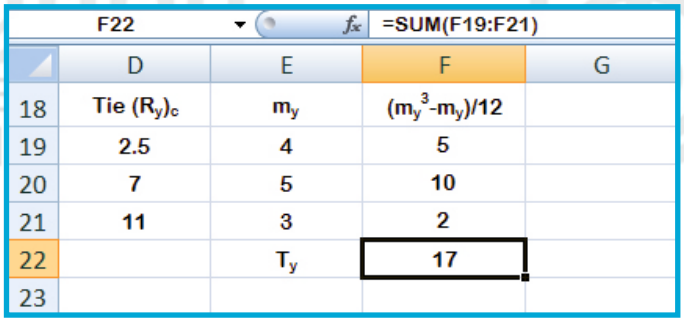


Fig. 8.37

Step 13: The formula for computing the corrected rank correlation coefficient is given by equation (13). We type "=(I19-(I17+C22+F22))/((Sqrt(I19-2*C22))*(Sqrt(I19-2*F22)))" in Cell I20 as shown in Fig. 8.38.

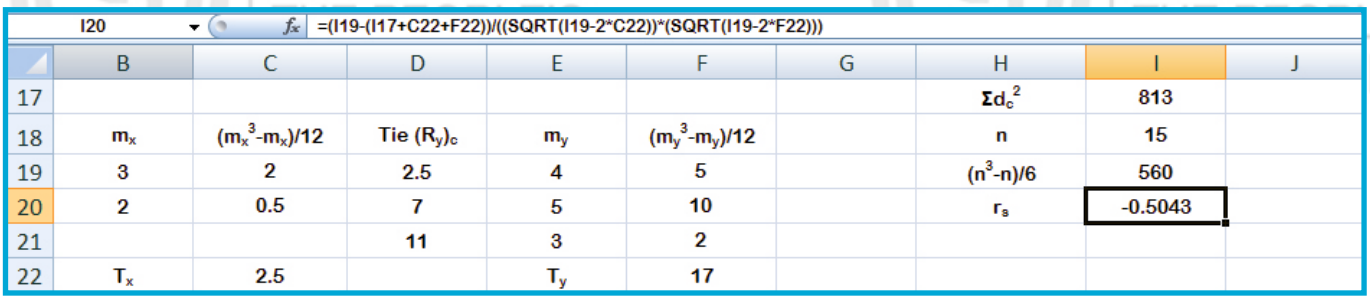


Fig. 8.38

Notice from Fig. 8.38 that the rank correlation coefficient between the marks of Parts A and B is – 0.5043. We can conclude that there is

negative correlation between the number of hours spent in watching TV and the marks obtained for the given data, i.e., an increase in the number of hours spent in watching TV is correlated with a decrease in the marks obtained in the examination.

Step 14: We can also compute Spearman's rank correlation coefficient using the ***Correl*** function as explained in Sec. 8.7. We use the ***Correl*** function, as explained in Sec. 8.3, in Cell I21. We select Cells F2:F16 as ***Array 1***, Cells G2:G16 as ***Array 2***, and click on ***OK***. We obtain the value of the correlation coefficient in Cell I21 (Fig. 8.39).

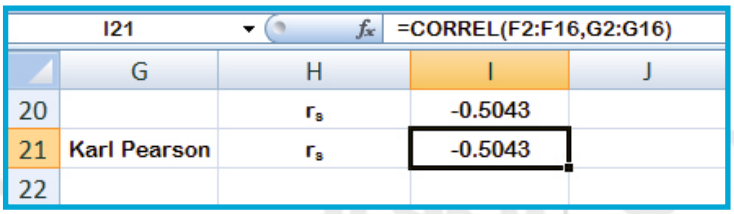


Fig. 8.39

Notice from Fig. 8.39 that the Karl Pearson's correlation coefficient computed from equation (4) is the same as the rank correlation coefficient computed from equation (13).

You can now try the following exercises for practice.

Activity

Work out the following exercises with the help of MS Excel 2007 and interpret the results:

A1) Examples 1 to 3 given in Unit 6 of MST-002.

A2) Exercises E4 to E6 given in Unit 6 of MST-002.

A3) Examples 1 to 4 given in Unit 7 of MST-002.

A4) Exercises E1 to E3 given in Unit 7 of MST-002.

A5) Examples 1 to 3 given in Unit 11 of MST-002.

A6) Exercises E1 to E5 given in Unit 11 of MST-002.

A7) Examples 1 to 3 given in Unit 12 of MST-002.

A8) Exercises E1 to E4 given in Unit 12 of MST-002.

Match the results with the manual computation of the correlation coefficients done in Units 6, 7, 11 and 12 of MST-002.

Continuous Assessment 8

1. Suppose we are interested in determining the correlation coefficients between the electricity consumption, size of the house and the number of hours an AC is used in a household during summers. For this purpose, a

sample of 40 houses having one AC was selected. We have recorded the electricity consumption (in kWh), size of the house (in square feet) and number of hours of AC use for one month during summers in Table 4.

Table 4: Electricity consumption data

S. No.	Unit (in kWh)	Area (in sq ft)	AC (in hours)
1	1060	1316	5
2	1150	1420	7
3	1365	1556	12
4	1275	1488	9
5	1425	1612	13
6	1310	1516	10
7	1365	1556	12
8	1075	1352	6
9	925	1168	4
10	1340	1540	11
11	1425	1612	13
12	1150	1420	8
13	1060	1316	5
14	1545	1680	15
15	1140	1388	7
16	1075	1352	6
17	1620	1736	16
18	1050	1296	5
19	1310	1516	10
20	1645	1760	16
21	1565	1696	15
22	1215	1464	9
23	1275	1488	10
24	1465	1632	13
25	1080	1356	7
26	975	1196	4
27	1040	1256	5
28	1340	1540	11
29	865	1144	4
30	1175	1440	8
31	1080	1356	7
32	1500	1652	15
33	1175	1440	9
34	1050	1296	5
35	1365	1580	12
36	1465	1632	15
37	1215	1464	9
38	1365	1580	12
39	1140	1388	7
40	1005	1224	4

i) Compute Pearson's correlation coefficients between (a) electricity consumption and size of the house, (b) electricity consumption and number of hours the AC is used and (c) size of the house and number of hours the AC is used for the given data.

ii) Determine the multiple correlation coefficients $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$.

iii) Obtain the partial correlation coefficients $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$.

2. 25 singers participated in a singing contest and were evaluated by two judges A and B. The judges awarded scores between 0 and 100 to each contestant, which are recorded in Table 5.

Table 5: Scores of the singing contest

Judge A	48	75	65	55	59	62	93	57	73	63	82	99	88
	52	78	61	58	50	95	70	80	81	97	47	77	
Judge B	51	76	65	53	56	58	94	69	64	71	77	98	87
	55	75	61	60	49	93	92	81	72	96	46	79	

Compute Spearman's rank correlation coefficient between the scores awarded by both judges for the given data.

3. Suppose we are interested in determining the relationship between the ranking of 15 universities of the country and their success rates. The rank from 0 to 10 is assigned to a university on the basis of different factors and the success rate is based on the percentage of students passing out. The ranks and success rates of 15 universities are recorded in Table 6.

Table 6: Ranks and success rates of 15 universities

S. No.	Rank of the University	Success Rate (in %)
1	2	57
2	9	98
3	5	70
4	9	95
5	6	75
6	4	70
7	8	96
8	8	92
9	7	87
10	8	75
11	7	85
12	3	76
13	5	64
14	7	82
15	5	58

Compute Spearman's rank correlation coefficient between the ranks and success rates of the universities for the given data.

Home Work: Do It Yourself

1) Follow the steps explained in Secs. 8.3 to 8.8 to compute the various correlation coefficients for the data of Tables 1, 2 and 3. Take the screenshots and keep them in your record book.
2) Develop the spreadsheets for the exercises given in "Continuous Assessment 8" as explained in this lab session. Take screenshots of the final spreadsheets.
3) **Do not forget** to keep the screenshots in your record book as these will contribute to your continuous assessment in the Laboratory.