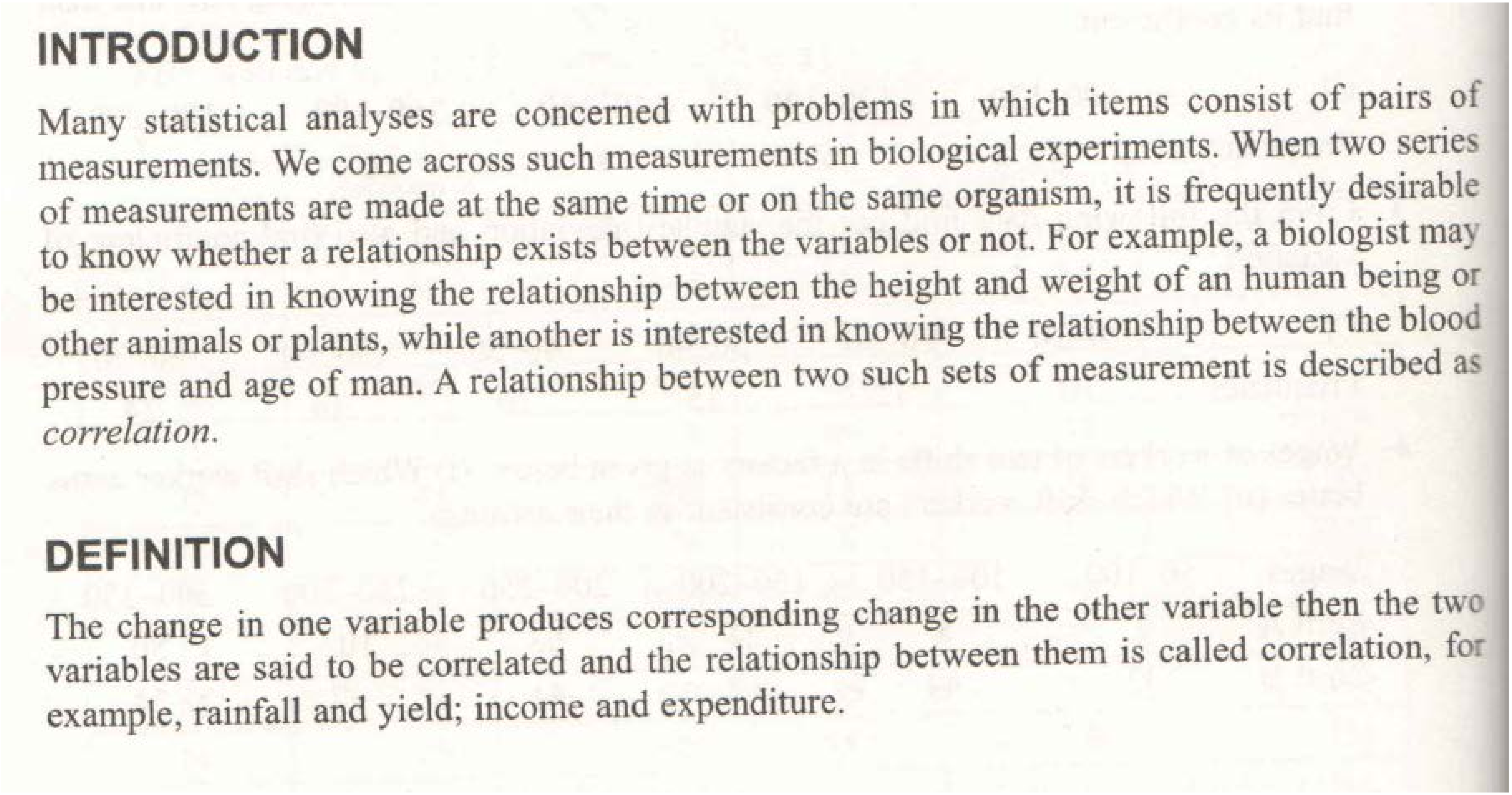
Chapter:7

CORRELATION

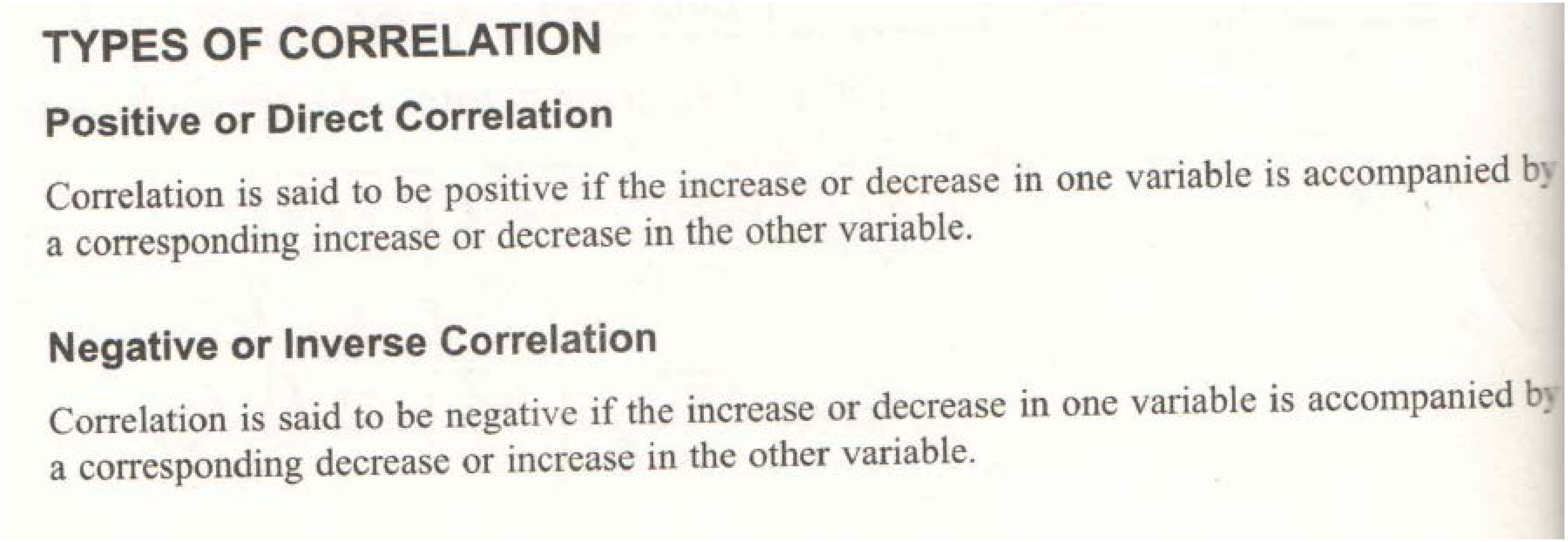


INTRODUCTION

Many statistical analyses are concerned with problems in which items consist of pairs of measurements. We come across such measurements in biological experiments. When two series of measurements are made at the same time or on the same organism, it is frequently desirable to know whether a relationship exists between the variables or not. For example, a biologist may be interested in knowing the relationship between the height and weight of an human being or other animals or plants, while another is interested in knowing the relationship between the blood pressure and age of man. A relationship between two such sets of measurement is described as *correlation*.

DEFINITION

The change in one variable produces corresponding change in the other variable then the two variables are said to be correlated and the relationship between them is called correlation, for example, rainfall and yield; income and expenditure.



TYPES OF CORRELATION

Positive or Direct Correlation

Correlation is said to be positive if the increase or decrease in one variable is accompanied by a corresponding increase or decrease in the other variable.

Negative or Inverse Correlation

Correlation is said to be negative if the increase or decrease in one variable is accompanied by a corresponding decrease or increase in the other variable.

Simple and Multiple Correlations

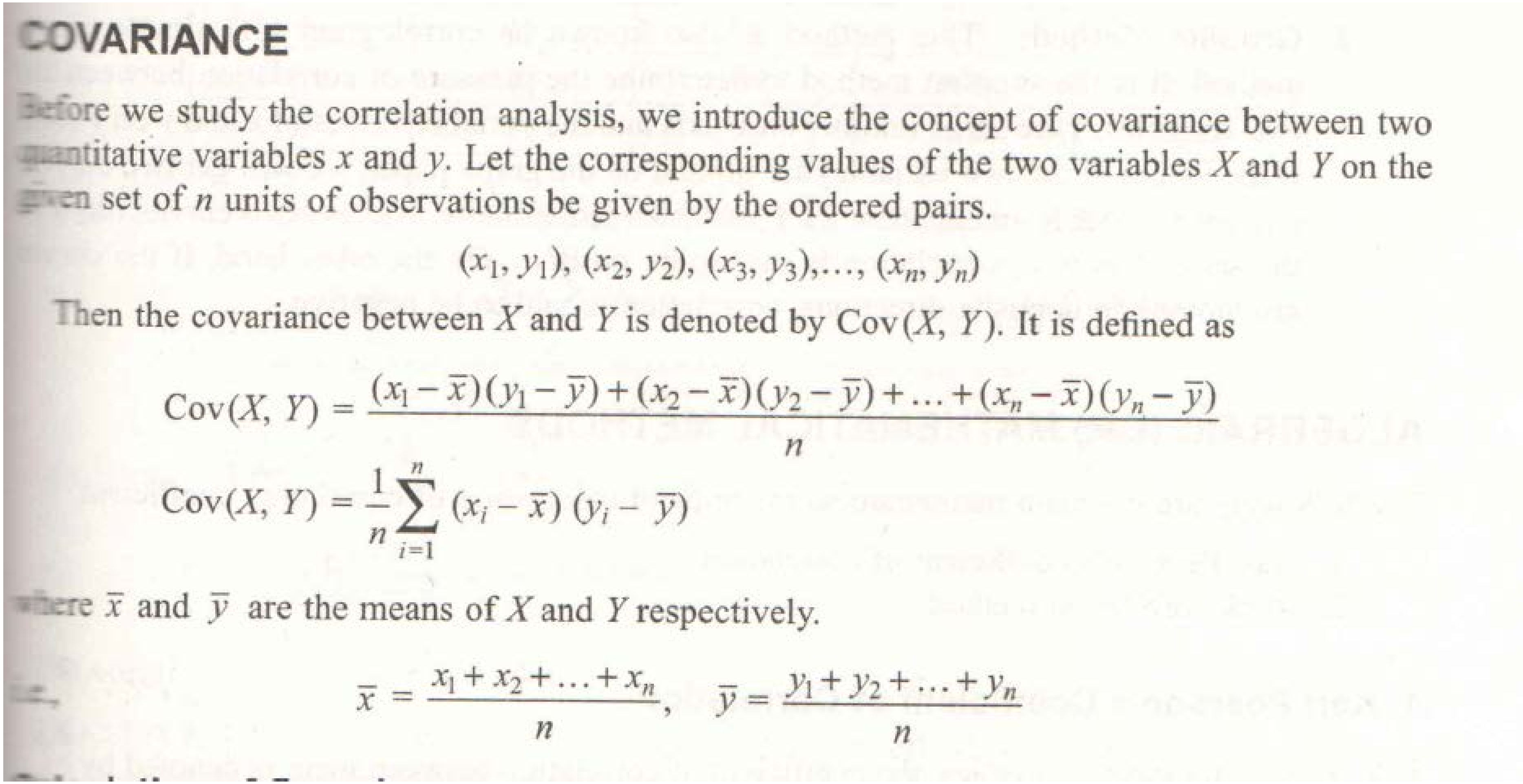
If the relationship is confined to two variables only, then the correlation between them is called *simple correlation.* If the relationship is confined to more than two variables, then the correlation between them is called *multiple correlation.*

Perfect Correlation

If the change in two variables is a constant ratio, then the correlation between them is called *perfect correlation.* If this change is in the same direction with constant ratio it is called *perfect positive correlation.* If this change is in the opposite direction with constant ratio, it is called *perfect negative correlation.*

PROPERTIES

1. The correlation coefficient is a pure number. It does not depend upon the units in which the variables x and y are expressed.
2. The correlation coefficient lies between –1 and 1. i.e., $-1 \leq r \leq 1$.
3. If $r = +1$, the correlation is perfect and positive if $r = -1$, the correlation is perfect and negative.
4. If $r = 0$, there is no correlation between the two variables.
5. The coefficient of correlation is not affected by change and scale of origin.



COVARIANCE

Before we study the correlation analysis, we introduce the concept of covariance between two quantitative variables x and y. Let the corresponding values of the two variables X and Y on the given set of n units of observations be given by the ordered pairs.

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$$

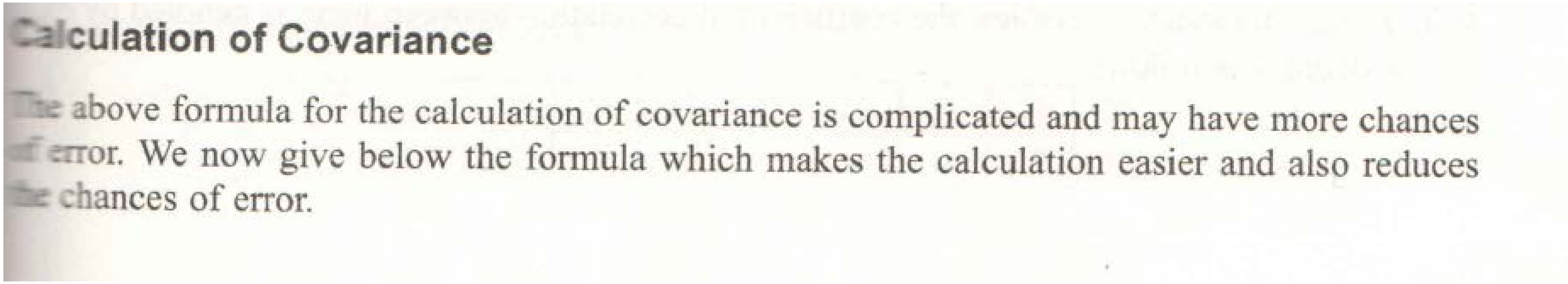
Then the covariance between X and Y is denoted by Cov(X, Y). It is defined as

$$\text{Cov}(X, Y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

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

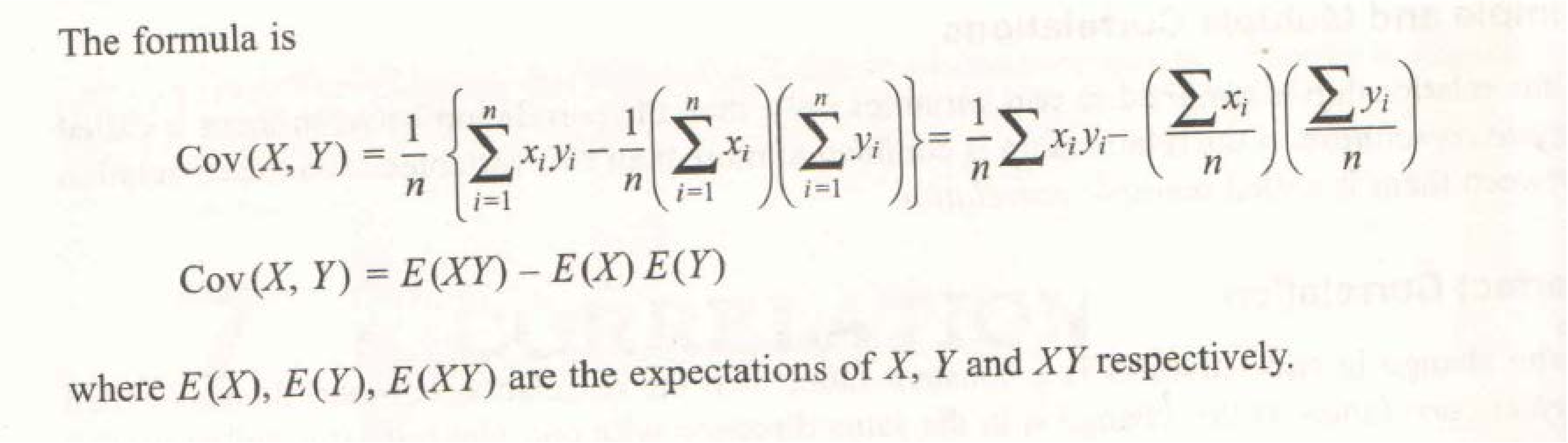
where $\bar{x}$ and $\bar{y}$ are the means of X and Y respectively.

i.e.,
$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \quad \bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$



Calculation of Covariance

The above formula for the calculation of covariance is complicated and may have more chances of error. We now give below the formula which makes the calculation easier and also reduces the chances of error.



The formula is

$$\operatorname{Cov}(X, Y) = \frac{1}{n}\left\{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right\} = \frac{1}{n}\sum x_i y_i - \left(\frac{\sum x_i}{n}\right)\left(\frac{\sum y_i}{n}\right)$$

$$\operatorname{Cov}(X, Y) = E(XY) - E(X)\,E(Y)$$

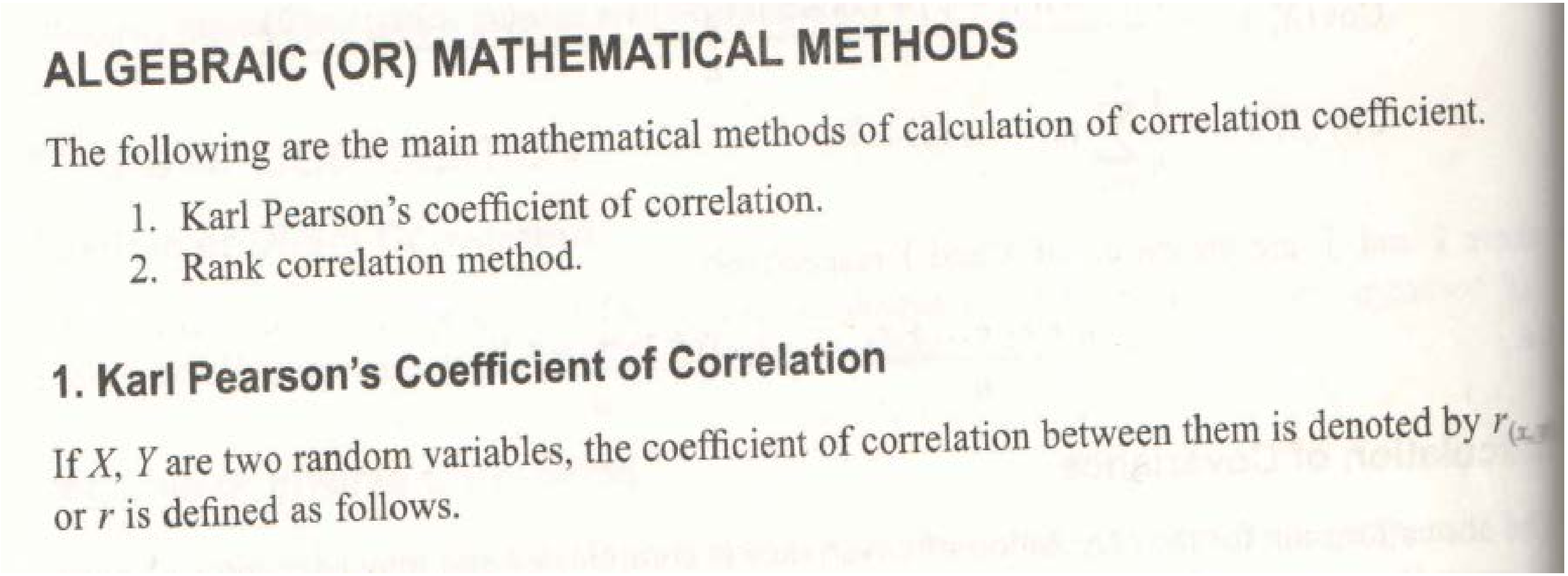
where $E(X)$, $E(Y)$, $E(XY)$ are the expectations of X, Y and XY respectively.

METHOD OF STUDYING CORRELATION

There are two methods which visualize the relationship between two variables, i.e., (1) the scatter diagram, (2) graphic method. These are based on graphs and diagrams.

1.Scatter Diagram Method: It is one of the simplest method s of diagrammatic representation of a bivariate distribution. It provides the simplest tool of determining the correlation between two variables. The term scatter refers to the dispersion (or) spread of the dots on the graph. Taking x values on the x −axis and y on the y −axis . We plotted the points (x_i, y_i) on a graph paper. If all the points are closely scattered, we can accept a correlation between the two variables. If the points are widely scattered , then there is no correlation between the variables. A diagram gives an idea about the correlation between the variables, so it is called a scatter diagram.

2.Graphic Method: **This method is also known as correlogram (or) simple graph method. It is the simplest method to determine the pressure of correlation between the two variables. Take serial number on $x - axis$ and the variable values x and y on the y –axis. If the values of the two variables are plotted on the graph paper, we will get two curves, one for x variable and another for y variables individually. If these two curves moves in the same direction, correlation is said to be positive. On the other hand , if the curves are moving in opposite directions, correlation is said to be negative.**



ALGEBRAIC (OR) MATHEMATICAL METHODS

The following are the main mathematical methods of calculation of correlation coefficient.

1. Karl Pearson's coefficient of correlation.
2. Rank correlation method.

1. Karl Pearson's Coefficient of Correlation

If X, Y are two random variables, the coefficient of correlation between them is denoted by $r_{(x,}$ or r is defined as follows.

$$r = \frac{\operatorname{cov}(x, y)}{\sqrt{v(x)\,v(y)}} = \frac{\operatorname{cov}(x, y)}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}}$$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}},$$ where n is the number of observations.

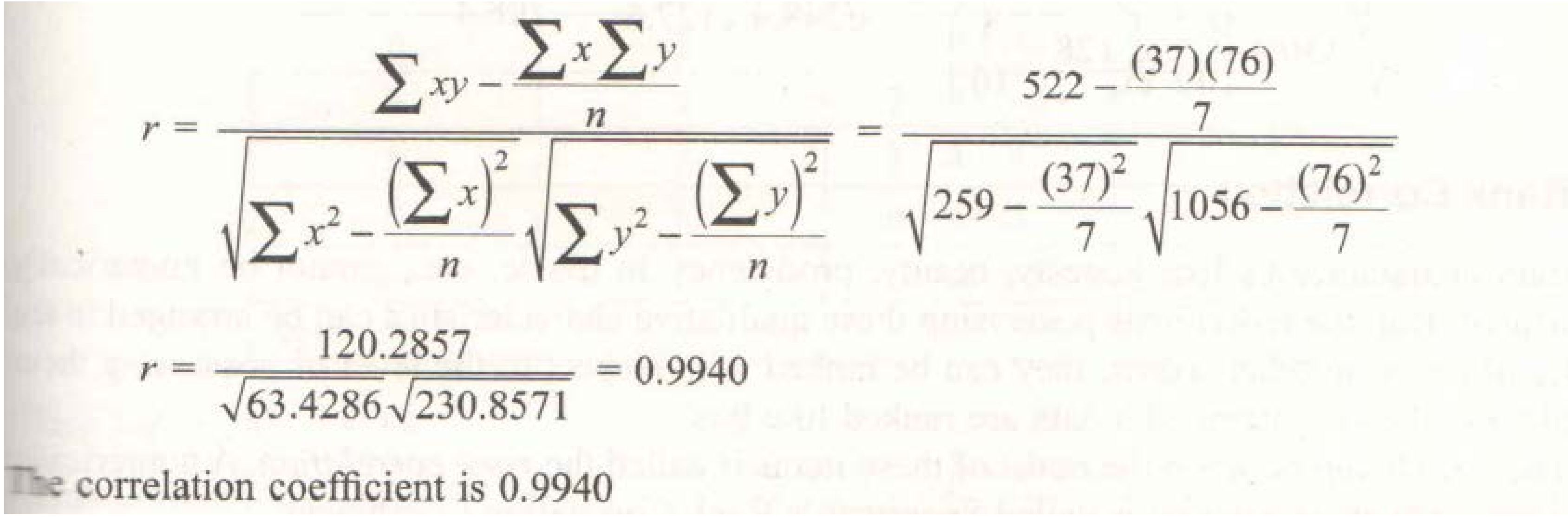
EXAMPLE 7.1: Calculate Karl Pearson's correlation coefficient for the following data.

x	1	2	4	5	7	8	10
y	2	6	8	10	14	16	20

Solution:

x	y	x^2	y^2	xy
1	2	1	4	2
2	6	4	36	12
4	8	16	64	32
5	10	25	100	50
7	14	49	196	98
8	16	64	256	128
10	20	100	400	200
37	76	259	1056	522

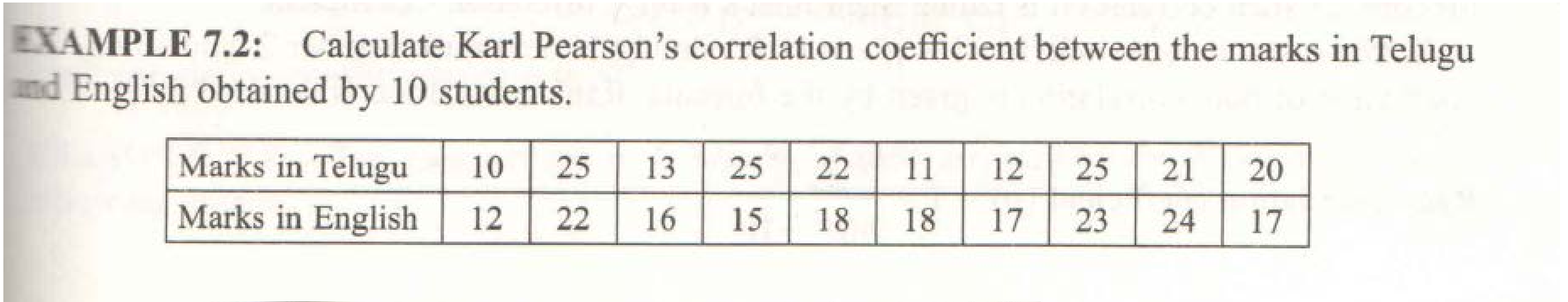
Here $\Sigma x = 37$, $\Sigma y = 76$, $\Sigma x^2 = 259$, $\Sigma y^2 = 1056$, $\Sigma xy = 522$ and $n = 7$.



$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}\sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}} = \frac{522 - \frac{(37)(76)}{7}}{\sqrt{259 - \frac{(37)^2}{7}}\sqrt{1056 - \frac{(76)^2}{7}}}$$

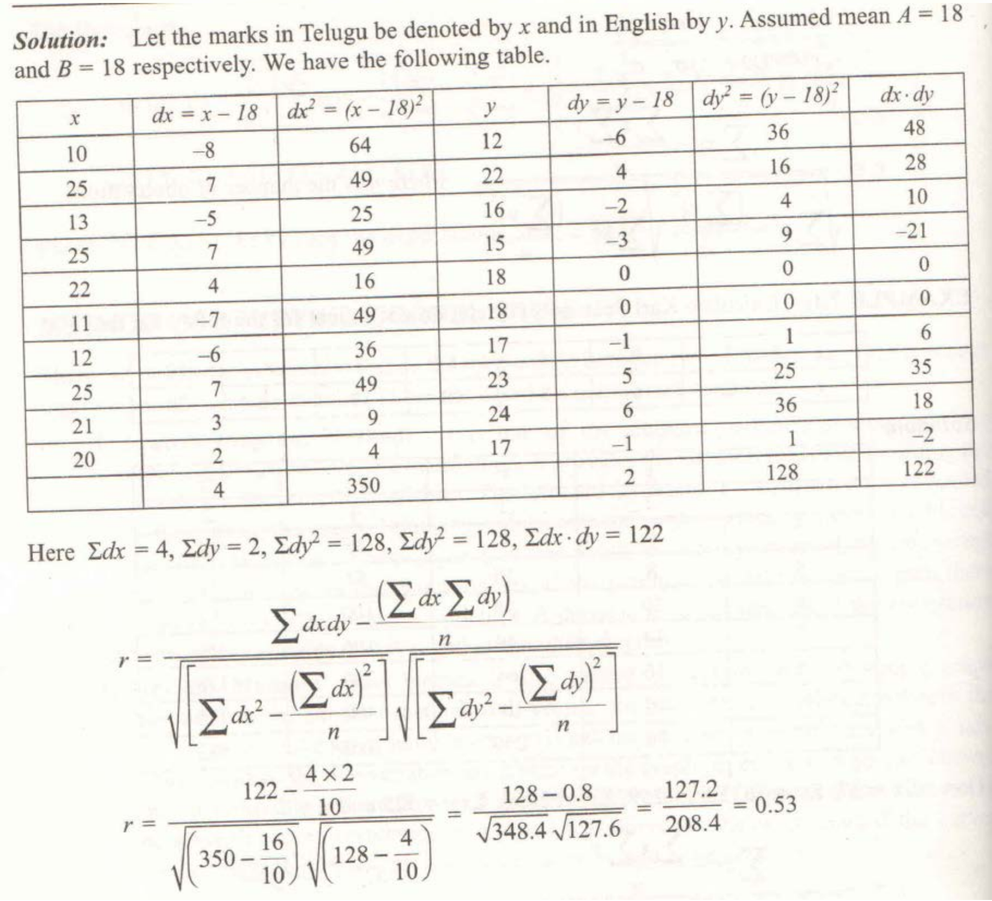
$$r = \frac{120.2857}{\sqrt{63.4286}\sqrt{230.8571}} = 0.9940$$

The correlation coefficient is 0.9940



EXAMPLE 7.2: Calculate Karl Pearson's correlation coefficient between the marks in Telugu and English obtained by 10 students.

Marks in Telugu	10	25	13	25	22	11	12	25	21	20
Marks in English	12	22	16	15	18	18	17	23	24	17



Solution: Let the marks in Telugu be denoted by x and in English by y. Assumed mean $A = 18$ and $B = 18$ respectively. We have the following table.

x	$dx = x - 18$	$dx^2 = (x-18)^2$	y	$dy = y - 18$	$dy^2 = (y-18)^2$	$dx \cdot dy$
10	–8	64	12	–6	36	48
25	7	49	22	4	16	28
13	–5	25	16	–2	4	10
25	7	49	15	–3	9	–21
22	4	16	18	0	0	0
11	–7	49	18	0	0	0
12	–6	36	17	–1	1	6
25	7	49	23	5	25	35
21	3	9	24	6	36	18
20	2	4	17	–1	1	–2
	4	350		2	128	122

Here $\Sigma dx = 4$, $\Sigma dy = 2$, $\Sigma dy^2 = 128$, $\Sigma dy^2 = 128$, $\Sigma dx \cdot dy = 122$

$$r = \frac{\sum dxdy - \frac{\left(\sum dx \sum dy\right)}{n}}{\sqrt{\left[\sum dx^2 - \frac{\left(\sum dx\right)^2}{n}\right]}\sqrt{\left[\sum dy^2 - \frac{\left(\sum dy\right)^2}{n}\right]}}$$

$$r = \frac{122 - \frac{4 \times 2}{10}}{\sqrt{\left(350 - \frac{16}{10}\right)}\sqrt{\left(128 - \frac{4}{10}\right)}} = \frac{128 - 0.8}{\sqrt{348.4}\sqrt{127.6}} = \frac{127.2}{208.4} = 0.53$$

2. Rank Correlation

Certain characteristics like honesty, beauty, proficiency in music, etc., cannot be numerically measured. But, the individuals possessing these qualitative characteristics can be arranged in the order of merit; in other words, they can be ranked with respect to the level of possessing these attributes when the items of a data are ranked like this.

The correlation between the ranks of these items is called the *rank correlation*. A numerically measure of such correlation is called Spearman's Rank Correlation Coefficient.

In this case, the best individual is given rank number 1, next rank number 2, and so on. The coefficient of rank correlation is given by the formula. Rank correlation is denoted by ρ.

$$\text{Rank correlation coefficient } (\rho) = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

where d_i = difference between the two corresponding ranks

$d_i = x_i - y_i$

N = Number of samples or observations or values.

Merits

1. It is easier to understand and calculate as compared to Karl Pearson's method.
2. It is useful for qualitative data such as beauty, honesty and efficiency.
3. It is a useful method when the actual data is not given but only ranks are given.

Limitations

1. It cannot be used for grouped frequency distribution.
2. It is not as accurate as Karl Pearson's coefficient of correlation.
3. It cannot be used in a continuous series.

When the ranks are given

EXAMPLE 7.3: Calculate the coefficient of correlation of ranks obtained by 10 students of a class in Mathematics and Physics.

Mathematics	1	2	3	4	5	6	7	8	9	10
Physics	3	8	1	7	10	2	9	4	6	5

Solution: Let Mathematics be denoted by x and Physics by y.

x	y	$d_i = x_i - y_i$	d_i^2
1	3	$1 - 3 = -2$	4
2	8	$2 - 8 = -6$	36
3	1	$3 - 1 = 2$	4
4	7	$4 - 7 = -3$	9
5	10	$5 - 10 = -5$	25
6	2	$6 - 2 = 4$	16
7	9	$7 - 9 = -2$	4
8	4	$8 - 4 = 4$	16
9	6	$9 - 6 = 3$	9
10	5	$10 - 5 = 5$	25
Total			148

Here $\Sigma d_i^2 = 148$, $n = 10$

Rank correlation coefficient $(\rho) = 1 - \dfrac{6\sum d_i^2}{n(n^2-1)} = 1 - \dfrac{6 \times 148}{10(10^2-1)} = 1 - \dfrac{888}{990} = 1 - 0.8970 = 0.103$

Rank correlation coefficient $(\rho) = 0.103$

EXAMPLE 7.4: Ten competitors in a beauty contest are ranked by three judges in the following orders:

1st judge	1	6	5	10	3	2	4	9	7	8
2nd judge	3	5	8	4	7	10	2	1	6	9
3rd judge	6	4	9	8	1	2	3	10	5	7

Use the correlation coefficient to determine which pair of judges has the nearest approach to common taste in beauty.

Solution: Let R_1, R_2, R_3 respectively be the ranks given by first, second and third judge. Let ρ_{ij} be the rank correlation coefficient between the rank given by ith and jth judges. $i \neq j$. $i =$ 1, 2, 3 and j = 1, 2, 3. Let $d_i = R_i - R_j$ be the difference of ranks of an individual given by ith and jth judge.

R_1	R_2	R_3	$d_{12} = R_1 - R_2$	$d_{13} = R_1 - R_3$	$d_{23} = R_2 - R_3$	d_{12}^2	d_{13}^2	d_{23}^2
1	3	6	–2	–5	–3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	–3	–4	–1	9	16	1
10	4	8	6	2	–4	36	4	16
3	7	1	–4	2	6	16	4	36
2	10	2	–8	0	8	64	0	64
4	2	3	2	1	–1	4	1	1
9	1	10	8	–1	9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	–1	1	2	1	1	4
			0	0	0	200	60	214

Here $\Sigma d_{12} = 0, \Sigma d_{13} = 0, \Sigma d_{23} = 0, \Sigma d_{12}^2 = 200, \Sigma d_{13}^2 = 60, \Sigma d_{23}^2 = 214, n = 10.$

$$\rho_{12} = 1 - \frac{6\sum d_{12}^2}{n(n^2-1)} = 1 - \frac{6\times 200}{10(10^2-1)} = -\,2121$$

$$\rho_{13} = 1 - \frac{6\sum d_{13}^2}{n(n^2-1)} = 1 - \frac{6\times 60}{10(10^2-1)} = 0.6364$$

$$\rho_{23} = 1 - \frac{6\sum d_{23}^2}{n(n^2-1)} = 1 - \frac{6\times 214}{10(10^2-1)} = -\,0.2970$$

Since ρ_{13} is maximum, so the pair of the first and third judge has the nearest approach to the common taste of beauty.

When the ranks are not given

In this case, we are given only the data. We assign the ranks to both the series *X* and *Y* by giving the ranks one to highest values in both the series, and so on.

EXAMPLE 7.5: Ten students got the following percentage of marks in English and Hindi.

English	8	36	98	25	75	82	92	62	65	35
Hindu	84	51	91	60	68	62	86	58	35	49

Find the coefficient of rank correlation.

Solution: Let English be denoted by x and Hindi by y.

x	R_1 *rank in x*	y	R_2 = *rank in y*	$d_i = R_1 - R_2$	d_i^2
8	10	84	3	7	49
36	7	51	8	–1	1
98	1	91	1	0	0
25	9	60	6	3	9
75	4	68	4	0	0
82	3	62	5	–2	4
92	2	86	2	0	0
62	6	58	7	–1	1
65	5	35	10	–5	25
35	8	49	9	–1	1
				0	90

Here $\Sigma d_i = 0$, $\Sigma d_i^2 = 90$, $n = 10$

Rank correlation coefficient

$$(\rho) = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 90}{10(10^2-1)} = 1 - \frac{540}{990} = 1 - 0.5455 = 0.4545$$

Rank correlation coefficient $(\rho) = 0.4545$

EXAMPLE 7.6: The marks obtained by the students in Chemistry and Statistics are as follows.

Marks in Chemistry	36	24	47	18	11	44	10	7	29
Marks in Statistics	31	34	46	24	9	50	13	5	32

Compute their ranks in the two subjects and the coefficient of correlation of ranks.

Solution: Let the marks in Chemistry denoted by x and in Statistics by y.

x	R_1 = *rank in x*	y	R_2 = *rank in y*	$d_i = R_1 - R_2$	d_i^2
36	3	31	5	–2	4
24	5	34	3	2	4
48	1	46	2	–1	1
17	6	24	6	0	0
11	7	9	8	–1	1
44	2	50	1	1	1
10	8	13	7	1	1
7	9	5	9	0	0
29	4	32	4	0	0
					12

Contd

Here $\Sigma d_i = 0$, $\Sigma d_i^2 = 12$, $n = 9$

$$\text{Rank correlation coefficient } (\rho) = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6\times 12}{9(9^2-1)} = 1 - \frac{72}{720} = 1 - 0.1 = 0.9$$

Rank correlation coefficient $(\rho) = 0.9$

When equal ranks are given to more than two attributes or tie cases

If two or more individuals are placed together in any classification with respect to an attribute, i.e., if in case of variable data, there are more than one item with the same value in either or both the series (Tie rank), then the Spearman's Rank Correlation Coefficient formula does not give the correlation coefficient for Tie rank. The problem is solved by assigning average rank to each of these individuals who are put in tie. In order to find the Rank Correlation Coefficient of Repeated Ranks or Tie Ranks, an adjustment or correction factor is added to the Spearman's Rank Correlation formula.

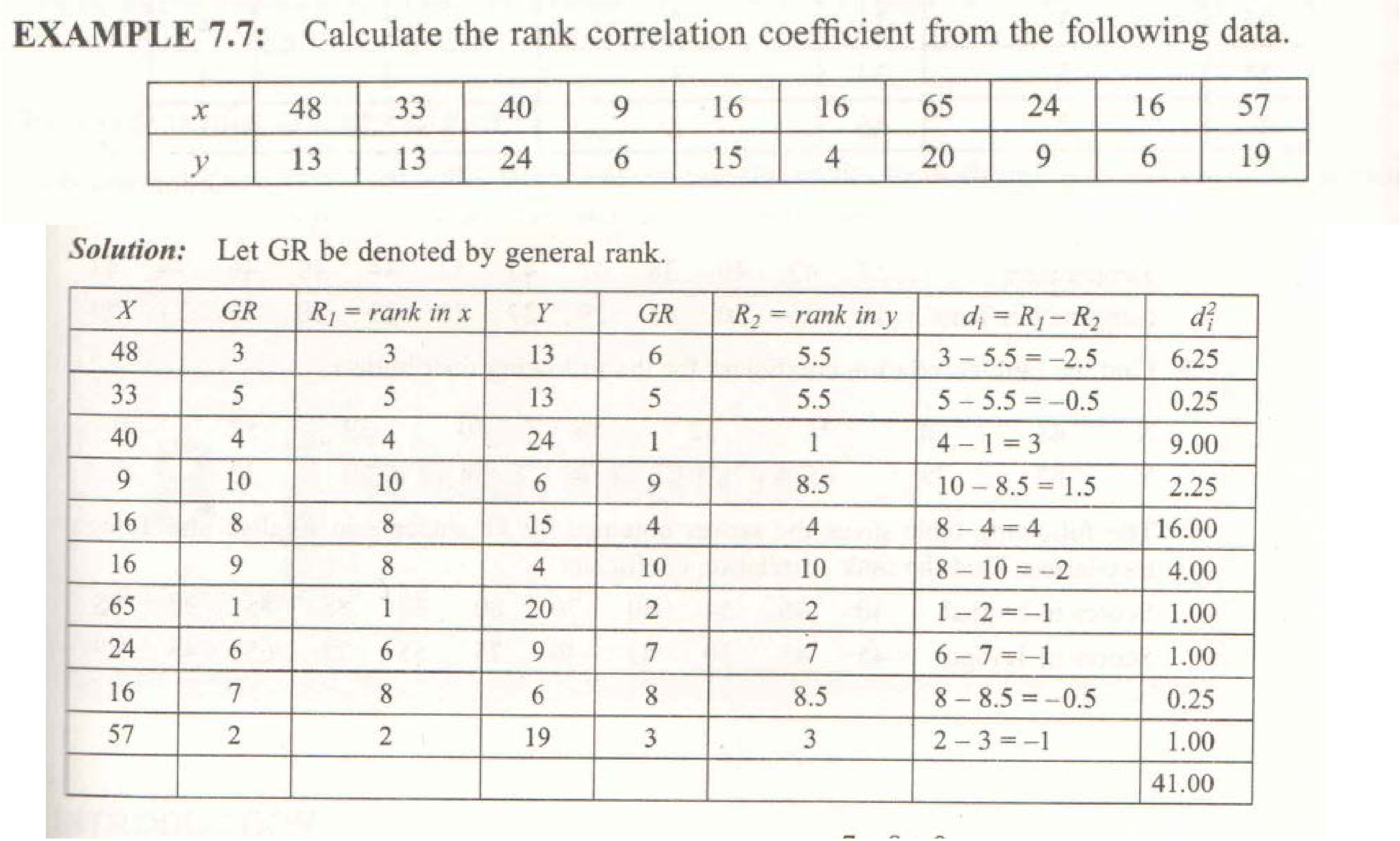
$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

Correction factor: In the formula $\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$, add the factor $\frac{m(m^2 - 1)}{12}$ to Σd^2, where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both the series.

The modified formula for tie rank correlation coefficient is given by

$$\rho = 1 - \frac{6\left[\sum d_i^2 + \sum_{j=1}^{k} \frac{m(m_j^2 - 1)}{12}\right]}{n(n^2 - 1)}$$

where $m_1, m_2, \ldots, m_k$ are the number of times a value is repeated.



EXAMPLE 7.7: Calculate the rank correlation coefficient from the following data.

x	48	33	40	9	16	16	65	24	16	57
y	13	13	24	6	15	4	20	9	6	19

Solution: Let GR be denoted by general rank.

X	GR	R_1 = *rank in x*	Y	GR	R_2 = *rank in y*	$d_i = R_1 - R_2$	d_i^2
48	3	3	13	6	5.5	$3 - 5.5 = -2.5$	6.25
33	5	5	13	5	5.5	$5 - 5.5 = -0.5$	0.25
40	4	4	24	1	1	$4 - 1 = 3$	9.00
9	10	10	6	9	8.5	$10 - 8.5 = 1.5$	2.25
16	8	8	15	4	4	$8 - 4 = 4$	16.00
16	9	8	4	10	10	$8 - 10 = -2$	4.00
65	1	1	20	2	2	$1 - 2 = -1$	1.00
24	6	6	9	7	7	$6 - 7 = -1$	1.00
16	7	8	6	8	8.5	$8 - 8.5 = -0.5$	0.25
57	2	2	19	3	3	$2 - 3 = -1$	1.00
							41.00

In the x series 16 number is repeated 3 times then the rank is $\frac{7+8+9}{3} = 8$.

In the y series 13 number is repeated 2 times then the rank is $\frac{5+6}{2} = 5.5$

In the y series 6 number is repeated 2 times then the rank is $\frac{8+9}{2} = 8.5$

16 value is repeated 3 times, i.e., $m_1 = 3 \cdot \frac{m_1(m_1^2-1)}{12} = \frac{3(3^2-1)}{12} = \frac{3\times 8}{12} = 2$

13 value is repeated 2 times, i.e., $m_2 = 2 \cdot \frac{m_2(m_2^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{2\times 3}{12} = 0.5$

6 value is repeated 2 times, i.e., $m_3 = 2 \cdot \frac{m_2(m_2^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{2\times 3}{12} = 0.5$

Correction factor $= \frac{\sum m_j(m_j^2-1)}{12} = 2 + 0.5 + 0.5 = 3$ and $\Sigma d_i^2 = 41$

$$\rho = 1 - \frac{6\left[\sum d_i^2 + \sum_{j=1}^{k} \frac{m(m_j^2-1)}{12}\right]}{n(n^2-1)} = 1 - \frac{6[41+3]}{10(10^2-1)} = 1 - \frac{6\times 44}{990} = 1 - 0.2667 = 0.7333$$

The rank correlation, $\rho = 0.7333$.

END