PAS

1

Correlation

INTRODUCTION

In our day-to-day life, we find many examples when a mutual relationship exists between two variables, *i.e.*, with fall or rise in the value of one variable, the fall or rise may take place in the value of other variable. For example, price of a commodity rises as the demand for the commodity goes up. Upto a certain time-period, weight of a person increases with the increase in age. Similarly, the temperature rises with the rise in the sun light. These facts indicate that there is certainly some mutual relationship that exists between the demand for a commodity and its price, the age of a person and his weight, and the sunlight and temperature. The correlation refers to the statistical technique used in measuring the closeness of the relationship between the variables.

DEFINITION OF CORRELATION

Some important definitions of correlation are given below:

1. Correlation analysis deals with the association between two or more variables.
—**Simpson and Kafka**
2. If two or more quantities vary in sympathy, so that movement in one tend to be accompanied by corresponding movements in the other, then they are said to be correlated.
—**Conner**
3. Correlation analysis attempts to determine the degree of relationship between variables.
—**Ya-Lun Chou**

Thus, correlation is a statistical technique which helps in analysing the relationship between two or more variables.

UTILITY OF CORRELATION

The study of correlation is of immense significance in statistical analysis and practical life, which is clear from the following points:

(1) Most of variables show same kind of relationship. For example, there is relationship between price and supply, income and expenditure, etc. With the help of correlation analysis, we can measure the degree of relationship in one figure between different variables like supply and price, income and expenditure, etc.

(2) Once we come to know that the two variables are mutually related, then we can estimate the value of one variable on the basis of the value of another. This function is performed by regression technique, which is based on correlation. In other words, the concept of regression is based on correlation.

(3) Correlation is also useful for economists. An economist specifies the relationship between different variables like demand and supply, money supply and price level by way of correlation.

(4) In business, a trader makes the estimation of costs, sales, prices, etc., with the help of correlation and makes appropriate plans.

Thus, in every field of practical life, correlation analysis is extremely useful in making a comparative study of two or more related phenomena and analyzing their mutual relationship.

TYPES OF CORRELATION

Main types of correlation are given below:

(1) Positive and Negative Correlation: On the basis of direction of change of the variables, correlation can be classified into two types:

(*i*) **Positive Correlation:** If two variables X and Y move in the same direction, *i.e.*, if one rises, other rises too and vice versa, then it is a called as positive correlation. Examples of positive correlation are the relationship between price and supply, between money supply and prices, etc.

(*ii*) **Negative Correlation:** If two variables X and Y move in opposite direction, *i.e.*, if one rises, other falls, and if one falls, other rises, then it is called as negative correlation. Examples of negative correlation are the relationship between demand and price, investment and rate of interest, etc.

(2) Linear and Curvi-Linear Correlation: On the basis of change in proportion, correlation is of two types:

(*i*) **Linear Correlation:** If the ratio of change of two variables X and Y ($\Delta Y / \Delta X$) remains constant throughout, then they are said to be linearly correlated, like as when everytime supply of a commodity rises by 20% as often as its price rises by 10%, then such two variables have linear relationship. If values of these two variables are plotted on a graph, then all the points will lie on a straight line.

(*ii*) **Curvi-Linear Correlation:** If the ratio of change between the two variables is not constant but changing, correlation is said to be curvi-linear, like as when everytime price of a commodity rises by 10%, then sometimes its supply rises by 20%, sometimes by 10% and sometimes by 40%, then non-linear or curvi-linear correlation exists between them. In case of curvi-linear correlation, values of the variables plotted on a graph will give a curve.

(3) Simple Partial and Multiple Correlation: On the basis of number of variables studied, correlation may be classified into three types:

(*i*) **Simple Correlation:** When we study the relationship between two variables only, then it is called simple correlation. Relationship between price and demand, height and weight, income and consumption, etc., are all examples of simple correlation.

(*ii*) **Partial Correlation:** When three or more variables are taken but relationship between any two of the variables is studied, assuming other variables as constant, then it is called partial correlation. Suppose, under constant temperature, we study the relationship between amount of rainfall and wheat yield, then this will be called as partial correlation.

(*iii*) **Multiple Correlation:** When we study the relationship among three or more variables, then it is called multiple correlation. For example, if we study the relationship between rainfall, temperature and yield of wheat, then it is called as multiple correlation.

CORRELATION AND CAUSATION

Correlation is a numerical measure of direction and m
between the values of two or more variables. But the presence
the belief that the two correlated variables necessarily have ca
does not always arise from causal relationship but with t
correlation is certain to exist. Presence of high degree of correla
due to the following reasons:

(1) Mutual Dependence: The study of economic theory s
variable may affect other variable. It is possible that the
mutually. In such situation, it is difficult to know which one is
example, price of a commodity is affected by the forces of
of demand, with the rise in price (other things remaining
fall. Here rise in price is the cause and fall in demand is the
demand, price of the commodity falls. Here fall in demand is t
thus there may be high degree of correlation between two va
is difficult to know which one is the cause and which one i

(2) Due to Pure Chance: In a small sample it is possible t
the universe, these variables are unlikely to be correlated,
fluctuations of pure random sampling or due to the bias o
following example makes the point clear:

Income (in Rs.)	5,000	6,000	7,000
Weight (in Kg.)	100	120	140

In the data as stated above, there is perfect positive correl
weight increases with rise in income and the rate of change
such kind of correlation cannot be said to be meanir
spurious or *non-sense*.

(3) Correlation Due to any Third Common Factor: T
a common third factor rather than having direct correlatio
positive correlation between per hectare field of tea and
has risen due to the rich yield of tea. Another reason of
well in time that affects both of these two.

DEGREE OF CORRELATION

Degree of correlation can be known by coefficient of
various types of the degree of correlation:

(1) Perfect Correlation
(2) High Degree of Correlation
(3) Moderate Degree of Correlation
(4) Low Degree of Correlation
(5) Absence of Correlation.

CORRELATION AND CAUSATION

Correlation is a numerical measure of direction and magnitude of the mutual relationship between the values of two or more variables. But the presence of correlation should not be taken as the belief that the two correlated variables necessarily have causal relationship as well. Correlation does not always arise from causal relationship but with the presence of causal relationship, correlation is certain to exist. Presence of high degree of correlation between different variables may be due to the following reasons:

(1) Mutual Dependence: The study of economic theory shows that it is not necessary that only one variable may affect other variable. It is possible that the two variables may affect each other mutually. In such situation, it is difficult to know which one is the cause and which one is the effect. For example, price of a commodity is affected by the forces of demand and supply. According to the law of demand, with the rise in price (other things remaining constant), demand for the commodity will fall. Here rise in price is the cause and fall in demand is the effect. On the other hand, with fall in demand, price of the commodity falls. Here fall in demand is the cause and fall in price is the effect. Thus there may be high degree of correlation between two variables due to mutual dependence, but it is difficult to know which one is the cause and which one is the effect.

(2) Due to Pure Chance: In a small sample it is possible that two variables are highly correlated but in universe, these variables are unlikely to be correlated, such correlation may be due to either the fluctuations of pure random sampling or due to the bias of investigator in selecting the sample. The following example makes the point clear:

Income (in Rs.)	5,000	6,000	7,000	8,000	9,000
Weight (in Kg.)	100	120	140	160	180

In the data as stated above, there is perfect positive correlation between income and weight, *i.e.*, weight increases with rise in income and the rate of change of the two variables is also the same. Still such kind of correlation cannot be said to be meaningful. Such relationship is said to be *spurious or non-sense*.

(3) Correlation Due to any Third Common Factor: Two variables may be correlated due to some common third factor rather than having direct correlation. For example, if there is high degree of positive correlation between per hectare field of tea and rice, then this does not imply that rice yield has risen due to the rich yield of tea. Another reason of the good yield of these two is the good rainfall well in time that affects both of these two.

DEGREE OF CORRELATION

Degree of correlation can be known by coefficient of correlation (r). The following can be various types of the degree of correlation:

(1) Perfect Correlation
(2) High Degree of Correlation
(3) Moderate Degree of Correlation
(4) Low Degree of Correlation
(5) Absence of Correlation.

(1) Perfect Correlation: When two variables vary at constant ratio in the same direction, it is perfect positive correlation and when the direction of change is opposite, it is perfect negative correlation. In case of perfect positive correlation, correlation coefficient (r) is equal to +1, and in case of perfect negative correlation, correlation coefficient (r) is equal to –1.

(2) High Degree of Correlation: When correlation exists in very large magnitude, then it is called high degree of correlation. In such a case, correlation coefficient ranges between ±0.75 and ±1.

(3) Moderate Degree of Correlation: Correlation coefficient, on being within the limits ±0.25 and ±0.75 is termed as moderate degree of correlation.

(4) Low Degree of Correlation: When correlation exists in very small magnitude, then it is called as low degree of correlation. In such a case, correlation coefficient ranges between 0 and ±0.25.

(5) Absence of Correlation: When there is no relationship between the variables, then correlation is found to be absent. In case of absence of correlation, the value of correlation coefficient is zero.

The degree of correlation on the basis of value of correlation coefficient can be summarised with the following table:

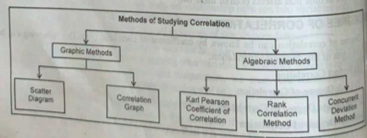
S.No.	Degree of Correlation	Positive	Negative
1.	Perfect Correlation	+1	–1
2.	High Degree of Correlation	Between +0.75 to +1	Between –0.75 to –1
3.	Moderate Degree of Correlation	Between +0.25 to +0.75	Between –0.25 to –0.75
4.	Low Degree of Correlation	Between 0 to + 0.25	Between 0 to –0.25
5.	Absence of Correlation	0	0

METHODS OF STUDYING CORRELATION

Correlation can be determined by the following methods:

(1) Graphic Methods
(*i*) Scatter Diagram
(*ii*) Correlation Graph

(2) Algebraic Methods
(*i*) Karl Pearson's Coefficient of Correlation
(*ii*) Spearman's Rank Correlation Method
(*iii*) Concurrent Deviation Method



(1) GRAPHIC METHOD

(i) Scatter Diagram

Scatter diagram is a graphic method of finding out correlation
method, direction of correlation can be ascertained. For constructin
is represented on X-axis and the Y-variable on Y-axis. Each pai
plotted in two-dimensional space of X—Y. Thus we get a scatter d
values. Different points may be scattered in various ways in the
gives us an idea about the direction and magnitude of correlation i

(*i*) **Perfect Positive Correlation (r = +1):** If all points are
line, passing from the lower corner of left side to the upp
series X and Y have perfect positive correlation, as is cle

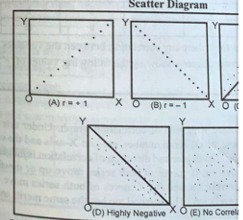
(*ii*) **Perfect Negative Correlation (r = –1):** When all points lie
then X and Y have perfect negative correlation, as is clea

(*iii*) **High Degree of Positive Correlation:** When concentrat
right upward and the points are close to each other, the
positive correlation, as is clear from the diagram (C) bel

(*iv*) **High Degree of Negative Correlation:** When points a
downward, and the points are close to each other, the
negative correlation, as is clear from the diagram (D) be

(*v*) **Zero Correlation (r = 0):** When all the points are sca
there and are lacking in any pattern, then there is absence
diagram (E) below.

Scatter Diagram



Example 1. Given the following pairs of values of the variab

X :	10	20	30	40
Y :	25	50	75	100

(*i*) Make a Scatter Diagram.
(*ii*) Is there any correlation between the variable

(1) GRAPHIC METHOD

(i) Scatter Diagram

Scatter diagram is a graphic method of finding out correlation between two variables. By this method, direction of correlation can be ascertained. For constructing a scatter diagram, X-variable is represented on X-axis and the Y-variable on Y-axis. Each pair of values of X and Y series is plotted in two-dimensional space of X—Y. Thus we get a scatter diagram by plotting all the pair of values. Different points may be scattered in various ways in the scatter diagram whose analysis gives us an idea about the direction and magnitude of correlation in the following ways:

(i) **Perfect Positive Correlation ($r = +1$):** If all points are plotted in the shape of a straight line, passing from the lower corner of left side to the upper corner at right side, then both series X and Y have perfect positive correlation, as is clear from the diagram (A) below.

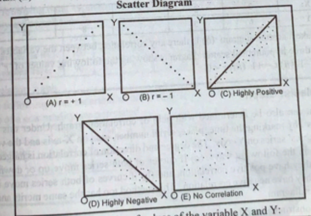
(ii) **Perfect Negative Correlation ($r = -1$):** When all points lie on a straight line from up to down, then X and Y have perfect negative correlation, as is clear from the diagram (B) below.

(iii) **High Degree of Positive Correlation:** When concentration of points moves from left to right upward and the points are close to each other, then X and Y have high degree of positive correlation, as is clear from the diagram (C) below.

(iv) **High Degree of Negative Correlation:** When points are concentrated from left to right downward, and the points are close to each other, then X and Y have high degree of negative correlation, as is clear from the diagram (D) below.

(v) **Zero Correlation ($r = 0$):** When all the points are scattered in four directions here and there and are lacking in any pattern, then there is absence of correlation, as is clear from the diagram (E) below.

Scatter Diagram



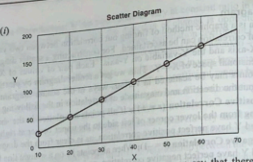
Example 1. Given the following pairs of values of the variable X and Y:

X :	10	20	30	40	50	60
Y :	25	50	75	100	125	150

(i) Make a Scatter Diagram.
(ii) Is there any correlation between the variables X and Y?

Solution: (i)

Scatter Diagram



(ii) By looking at the scatter diagram, we can say that there is perfect positive correlation between X and Y variables.

▶ **Merits and Demerits of Scatter Diagram**

Determining correlation by the method is easy because no mathematical computations are to be done. The major shortcoming of this method is that degree of correlation cannot be determined.

EXERCISE 1.1

1. Given the following pairs of values of the variables X and Y:

X:	2	3	5	6	8	9
Y:	6	5	7	8	12	11

(a) Make a scatter diagram. (b) Is there any correlation between the variables X and Y?

2. Draw three hypothetical scatter diagrams showing the following values of 'r':
(i) $r = -1$ (ii) $r = +1$ (iii) $r = 0$

(ii) Correlation Graph

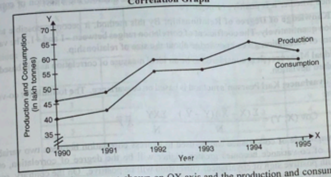
Correlation can also be determined with help of correlation graph. Under this method, two curves are drawn by marking the time, place, serial number, etc., on X-axis and the values of both correlated variables' series on Y-axis. The degree and direction of correlation is judged on the basis of these curves in the following ways: (a) If curves of both series move up or down in the same direction, then they have positive correlation, and (b) If curves of both series move in a opposite direction, then they have negative correlation. This method too has the same merits and demerits as those of a scatter diagram.

Example 2: Construct a correlation graph on the basis of the following data and comment on the relationship between production and consumption:

Year:	1990	1991	1992	1993	1994	1995
Production (in lakh tons):	46	48	58	58	64	60
Consumption (in lakh tons):	40	42	54	55	58	57

Solution:

Correlation Graph



In above shown graph, years are shown on OX axis and the production and consumption are shown on OY axis. This graph reveals that the two variables are closely related. Both curves are moving in one direction only. The distance between them also remains almost constant, therefore, there is high degree of positive correlation between them.

EXERCISE 1.2

1. From the following data, ascertain whether the income and expenditure of the 100 workers of a factory are correlated:

Year:	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
Average income (in Rs.):	100	102	105	105	101	112	118	120	125	130
Average expenditure:	90	91	93	95	92	94	100	105	108	110

[Ans. Closely Related]

Use Correlation graph.

2. From the following data, ascertain with the help of correlation graph, whether the demand and price of a commodity are correlated.

Year:	1986	1987	1988	1989	1990	1991	1992	1993
Demand in units:	50	55	62	70	75	78	80	82
Price in Rs.:	40	38	35	30	27	22	20	16

[Ans. Negatively correlated]

(2) ALGEBRAIC METHOD

(i) Karl Pearson's Coefficient of Correlation

It is quantitative method of measuring correlation. This method has been given by Karl Pearson and after his name, it is known as Pearson's coefficient of correlation. This is the best method of working out correlation coefficient. This method has the following main characteristics:

(1) **Knowledge of Direction of Correlation:** By this method, the direction of correlation is determined whether it is positive or negative.

(2) **Knowledge of Degree of Relationship:** By this method, it becomes possible to measure correlation quantitatively. The coefficient of correlation ranges between −1 and +1. The value of the coefficient of correlation gives knowledge about the size of relationship.

(3) **Ideal Measure:** It is considered to be an ideal measure of correlation as it is based on mean and standard deviation.

(4) **Covariance:** Karl Pearson's method is based on co-variance. The formula for co-variance is as follows:

$$\text{Cov}(X, Y) = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{N} = \frac{\Sigma XY}{N} - \overline{X}\,\overline{Y}$$

The magnitude of co-variance can be used to express correlation between two variables. As magnitude of co-variance becomes greater, higher will be the degree of correlation, otherwise lower. With positive sign of covariance, correlation will be positive. On the contrary, correlation will be negative if the sign of covariance is negative.

Calculation of Karl Pearson's Coefficient of Correlation

The calculation of Karl Pearson's coefficient of correlation can be divided into two parts:

(A) Calculation of Coefficient of Correlation in the case of Individual Series or Ungrouped Data

(B) Calculation of Coefficient of Correlation in the case of Grouped Data.

(A) Calculation of Coefficient of Correlation in case of Individual Series or Ungrouped Data

The following are the main methods of calculating the coefficient of correlation in individual series:

(1) Actual Mean Method

This method is useful when arithmetic mean happens to be in whole numbers or integers. This method involves the following steps:

(1) First, we compute the arithmetic mean of X and Y series, *i.e.*, $\overline{X}$ and $\overline{Y}$ are worked out.

(2) Then from the arithmetic means of the two series, deviations of the individual values are taken. The deviations of X-series are denoted by x and of the Y-series by y, *i.e.*, $x = X - \overline{X}$ and $y = Y - \overline{Y}$.

(3) Deviations of the two series are squared and added up to get Σx^2 and Σy^2.

(4) The corresponding deviations of the two series (x and y) are multiplied and summed up to get Σxy.

(5) Finally, correlation coefficient is found out by using the following formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} \text{ or } \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{\Sigma(Y-\overline{Y})^2}}$$

The correlation coefficient has the value always ranging between −1 and +1. The following examples clarify the computation procedure of this method:

Example 3. From the following data, calculate Karl Pearson

X:	2	3	4	5
Y:	4	7	8	9

Solution:

Calculation of Coefficient

X	$(X-\overline{X})$ x	x^2	Y
2	−3	9	4
3	−2	4	7
4	−1	1	8
5	0	0	9
6	+1	1	10
7	+2	4	14
8	+3	9	18
$\Sigma X = 35$ N = 7	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma Y = 70$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{35}{7} = 5, \quad \overline{Y} =$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} =$$

$$= \frac{58}{60.33} = +0.96$$

Thus, there is a high degree of positive correl

Example 4. From the following data, compute the coeffi series.

	X-Seri
Number of items:	15
Arithmetic mean:	25
Squares of deviations from mean:	136

Summation of product of deviations of X arithmetic means = 122.

Solution: We are given: $N = 15$, $\overline{X} = 25$, $\overline{Y} = 18$, $\Sigma x^2 =$

Applying the formula,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} = \frac{1}{\sqrt{136}}$$

$$= \frac{122}{\sqrt{18768}} = \frac{122}{136.996} =$$

Example 3. From the following data, calculate Karl Pearson's coefficient of correlation:

X:	2	3	4	5	6	7	8
Y:	4	7	8	9	10	14	18

Solution:

Calculation of Coefficient of Correlation

X	$(X-\bar{X})$ x	x^2	Y	$(Y-\bar{Y})$ y	y^2	xy
2	–3	9	4	–6	36	+18
3	–2	4	7	–3	9	+6
4	–1	1	8	–2	4	+2
5	0	0	9	–1	1	0
6	+1	1	10	0	0	0
7	+2	4	14	+4	16	+8
8	+3	9	18	+8	64	+24
$\Sigma X = 35$ N = 7	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma Y = 70$	$\Sigma y = 0$	$\Sigma y^2 = 130$	$\Sigma xy = 58$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{35}{7} = 5, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{70}{7} = 10$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{58}{\sqrt{28 \times 130}} = \frac{58}{\sqrt{3640}}$$

$$= \frac{58}{60.33} = +0.96$$

Thus, there is a high degree of positive correlation between the variables X and Y.

Example 4. From the following data, compute the coefficient of correlation between X and Y series.

	X-Series	Y-Series
Number of items:	15	15
Arithmetic mean:	25	18
Squares of deviations from mean:	136	138

Summation of product of deviations of X and Y series from their respective arithmetic means = 122.

Solution: We are given: $N = 15$, $\bar{X} = 25$, $\bar{Y} = 18$, $\Sigma x^2 = 136$, $\Sigma y^2 = 138$, $\Sigma xy = 122$

Applying the formula,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} = \frac{122}{\sqrt{136 \times 138}}$$

$$= \frac{122}{\sqrt{18768}} = \frac{122}{136.996} = +0.89$$

IMPORTANT TYPICAL EXAMPLES

Example 5. From the following table, calculate the coefficient of correlation by Karl Pearson's method:

X:	6	2	10	4	8
Y:	9	11	–	8	7

Arithmetic means of X and Y series are 6 and 8 respectively.

Solution: Let us first find the missing value of Y and let us denote it by a.

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{9+11+a+8+7}{5} = \frac{35+a}{5}$$

$$\Rightarrow \quad 8 = \frac{35+a}{5}$$

$$\therefore \quad 35 + a = 40 \quad \Rightarrow \quad a = 5$$

Thus, the complete series is:

X:	6	2	10	4	8
Y:	9	11	5	8	7

Now we find the coefficient of correlation.

Calculation of Coefficient of Correlation

X	$\bar{X}=6$ x	x^2	Y	$\bar{Y}=8$ y	y^2	xy
6	0	0	9	1	1	0
2	–4	16	11	3	9	–12
10	4	16	5	–3	9	–12
4	–2	4	8	0	0	0
8	2	4	7	–1	1	–2
$\Sigma X = 30$	$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma Y = 40$	$\Sigma y = 0$	$\Sigma y^2 = 20$	$\Sigma xy = -26$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$$

Applying the formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{-26}{\sqrt{40 \times 20}}$$

$$= \frac{-26}{\sqrt{800}} = \frac{-26}{28.2843}$$

$$= -0.9192$$

Example 6. From the data given below, find the number of items (N ... $r = 0.5$, $\Sigma xy = 120$, Standard Deviation of Y $(\sigma_y) = 8$, Σ ...
Where, x and y are deviations from arithmetic means.

Solution: Given: $r = 0.5, \Sigma xy = 120, \Sigma x^2 = 90, \sigma_y = 8$

Now, $\sigma_y = \sqrt{\frac{\Sigma y^2}{N}}$ when $y = Y - \bar{Y}$ [Fo ...

$8 = \sqrt{\frac{\Sigma y^2}{N}}$, squaring both sides, we ...

$64 = \frac{\Sigma y^2}{N} \quad \Rightarrow \Sigma y^2 = 64\,N$

Now, $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} \quad \Rightarrow 0.5 = \frac{12 ...}{\sqrt{90 \times ...}}$

Squaring both sides

$0.25 = \frac{(120)^2}{90 \times 64N} \quad \Rightarrow 0.25 = \frac{1440 ...}{5760 ...}$

$$\Rightarrow \quad (0.25)(5760)\,N = 14400$$

$$\Rightarrow \quad (1440)N = 14400$$

$$\therefore \quad N = \frac{14400}{1440} = 10$$

EXERCISE 1.3

1. Calculate Karl Pearson's coefficient of correlation between ... from the following:

Height of fathers (in inches):	65	66	67	...
Height of sons (in inches):	67	68	66	...

2. Calculate Pearson's coefficient of correlation between X a...

X:	14	19	24	21	26
Y:	31	36	48	37	50

3. Calculate the coefficient of correlation using Karl Pearson... value of the series given below:

X:	10	12	15
Y:	14	17	23

4. From the following data, compute Karl Pearson coefficient of correlation:

	X-series	Y-series
Number of items:	7	7
Arithmetic mean:	4	8
Sum of squares of deviations from arithmetic mean:	28	76

Summation of products of deviations of X and Y series from their respective means is 46.

[Ans. $r = 0.997$]

5. If $r = 0.25, \Sigma xy = 45, \sigma_y = 3, \Sigma x^2 = 50$, where x and y denote deviations from their respective means, find the number of observations.

[Ans. $N = 72$]

6. Two variates X and Y when expressed as deviations from their respective means are given as follows:

x:	−4	−3	−1	−2	0	1	2	3	4
y:	3	−3	?	0	4	1	2	−2	−1

Find the coefficient of correlation between them.

[Ans. $r = -0.06$]

[Hint: See Example 51]

7. Calculate Karl Pearson's coefficient of correlation, taking deviations from actual means 52 and 44 of the following data:

X:	44	46	46	48	52	54	?	56	60	60
Y:	36	40	42	40	?	44	46	48	50	52

[Ans. $r = +0.9504$]

8. Determine Pearson's coefficient of correlation from the following data:
$\Sigma X = 250, \Sigma Y = 300, N = 10, \Sigma(X-25)^2 = 480, \Sigma(Y-30)^2 = 600$ and
$\Sigma(X-25)(Y-30) = 150$

[Ans. $r = 0.28$]

(2) Assumed Mean Method

This method is useful when arithmetic mean is not in whole numbers but in fractions. In this method, deviations from assumed means of both the series (X and Y) are calculated. Correlation coefficient by this method can be determined in the following manner:

(1) Any values of X and Y are taken as their assumed mean, Ax and Ay.

(2) Deviations of the individual values of both the series (X and Y) are worked out from their assumed means. Deviations of X series $(X - Ax)$ are denoted by dx and of Y series $(Y - Ay)$ by dy.

(3) Deviations are summed up to get Σdx and Σdy.

(4) Then, squares of the deviations dx^2 and dy^2 are worked out and summed up to get Σdx^2 and Σdy^2 respectively.

(5) Each dx is multiplied by the corresponding dy and the products ($dxdy$) are added upto get $\Sigma dxdy$.

(6) Finally, correlation coefficient is obtained by using any one of following formula:

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx . \Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}} \sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

or

$$r = \frac{N . \Sigma dxdy - \Sigma dx . \Sigma dy}{\sqrt{N . \Sigma dx^2 - (\Sigma dx)^2} . \sqrt{N . \Sigma dy^2 - (\Sigma dy)^2}}$$

or

$$r = \frac{\Sigma dxdy - N(\bar{X} - Ax)(\bar{Y} - Ay)}{N . \sigma_x . \sigma_y}$$

Unless otherwise specifically asked, formula (*ii*) should be use... work very easy.

The following examples clarify the computation process of this ...

Example 7. Find the coefficient of correlation from the following ...

X :	10	12	18	16	15
Y :	30	35	45	44	42

Solution:

Calculation of Coefficient of Cor...

X	A=16 dx	dx^2	Y	A=... dy
10	−6	36	30	−1...
12	−4	16	35	
18	+2	4	45	
16 = A	0	0	44	
15	−1	1	42 = A	
19	+3	9	48	
18	+2	4	47	
17	+1	1	46	
ΣX = 125 N = 8	Σdx = −3	Σdx² = 71	ΣY = 337 N = 8	Σ...

$$\bar{X} = \frac{\Sigma X}{N} = \frac{125}{8} = 15.62 \ , \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{337}{8} ...$$

Since the actual means are not whole numbers, we ... 42 as assumed mean for Y.

Applying the formula,

$$r = \frac{N . \Sigma dxdy - \Sigma ...}{\sqrt{N . \Sigma dx^2 - (\Sigma dx)^2} . \sqrt{N ...}}$$

$$= \frac{8 \times 138 - (-3)(...}{\sqrt{8 \times 71 - (-3)^2} \sqrt{8 \times 2...}}$$

...ES

...ate the coefficient of correlation by Karl Pearson's...

2	10	4	8	
11	–	8	7	

...eries are 6 and 8 respectively.

...lue of Y and let us denote it by a.

$$...+11+a+8+7 \over 5 = \frac{35+a}{5}$$

$\Rightarrow a = 5$

2	10	4	8
11	5	8	7

...f correlation.

...of Coefficient of Correlation

Y	$\bar{Y}=8$ y	y^2	...
9	1	1	0
11	3	9	...
5	−3	9	...
8	0	0	0
7	−1	1	...
ΣY = 40	Σy = 0	Σy² = 20	Σxy...

$= \frac{\Sigma Y}{N} = \frac{40}{5} = 8$

$\frac{\Sigma xy}{...\times \Sigma y^2} = \frac{-26}{\sqrt{40 \times 20}}$

$= \frac{-26}{28.2843}$

...192

Example 6. From the data given below, find the number of items (N):
$r = 0.5$, $\Sigma xy = 120$, Standard Deviation of Y $(\sigma_y) = 8$, $\Sigma x^2 = 90$
Where, x and y are deviations from arithmetic means.

Solution: Given: $r = 0.5, \Sigma xy = 120, \Sigma x^2 = 90, \sigma_y = 8$

Now, $\sigma_y = \sqrt{\frac{\Sigma y^2}{N}}$ when $y = Y - \bar{Y}$ [Formula of S.D.]

$8 = \sqrt{\frac{\Sigma y^2}{N}}$, squaring both sides, we get

$64 = \frac{\Sigma y^2}{N} \Rightarrow \Sigma y^2 = 64\,N$

Now, $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} \Rightarrow 0.5 = \frac{120}{\sqrt{90 \times 64\,N}}$

Squaring both sides

$0.25 = \frac{(120)^2}{90 \times 64N} \Rightarrow 0.25 = \frac{14400}{5760\,N}$

$\Rightarrow (0.25)\,(5760)\,N = 14400$

$\Rightarrow (1440)N = 14400$

$\therefore N = \frac{14400}{1440} = 10$

EXERCISE 1.3

1. Calculate Karl Pearson's coefficient of correlation between the heights of fathers and sons from the following:

Height of fathers (in inches):	65	66	67	68	69	70	71
Height of sons (in inches):	67	68	66	69	72	72	69

[Ans. $r = 0.668$]

2. Calculate Pearson's coefficient of correlation between X and Y from the following data:

X:	14	19	24	21	26	22	15	20	19
Y:	31	36	48	37	50	45	33	41	39

[Ans. $r = 0.947$]

3. Calculate the coefficient of correlation using Karl Pearson's formula based on actual mean value of the series given below:

X:	10	12	15	23	20
Y:	14	17	23	25	21

[Ans. $r = 0.864$]

or $$r = \frac{\Sigma dxdy - \dfrac{\Sigma dx \,.\, \Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \dfrac{(\Sigma dx)^2}{N}}\sqrt{\Sigma dy^2 - \dfrac{(\Sigma dy)^2}{N}}} \quad ...(i)$$

or $$r = \frac{N \,.\, \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{\sqrt{N \,.\, \Sigma dx^2 - (\Sigma dx)^2} \,.\, \sqrt{N \,.\, \Sigma dy^2 - (\Sigma dy)^2}} \quad ...(ii)$$

or $$r = \frac{\Sigma dxdy - N(\bar{X} - Ax)(\bar{Y} - Ay)}{N.\sigma_x.\sigma_y} \quad ...(iii)$$

Note: Unless otherwise specifically asked, formula (*ii*) should be used as it makes the computation work very easy.

The following examples clarify the computation process of this method:

Example 7. Find the coefficient of correlation from the following data:

X :	10	12	18	16	15	19	18	17
Y :	30	35	45	44	42	48	47	46

Solution:

Calculation of Coefficient of Correlation

X	A=16 dx	dx^2	Y	A=42 dy	dy^2	$dxdy$
10	−6	36	30	−12	144	72
12	−4	16	35	−7	49	28
18	+2	4	45	+3	9	6
16 = A	0	0	44	+2	4	0
15	−1	1	42 = A	0	0	0
19	+3	9	48	+6	36	18
18	+2	4	47	+5	25	10
17	+1	1	46	+4	16	4
ΣX = 125 N = 8	Σdx = −3	Σdx^2 = 71	ΣY = 337 N = 8	Σdy = 1	Σdy^2 = 283	Σdxdy = 138

$$\bar{X} = \frac{\Sigma X}{N} = \frac{125}{8} = 15.62 \;, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{337}{8} = 42.12$$

Since the actual means are not whole numbers, we take 16 as assumed mean for X and 42 as assumed mean for Y.

Applying the formula,

$$r = \frac{N \,.\, \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{\sqrt{N \,.\, \Sigma dx^2 - (\Sigma dx)^2} \,.\, \sqrt{N \,.\, \Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{8 \times 138 - (-3)(1)}{\sqrt{8 \times 71 - (-3)^2}\sqrt{8 \times [illegible] - (1)^2}}$$

$$= \frac{1104 + 3}{\sqrt{568 - 9}\sqrt{2264 - 1}} = \frac{1107}{\sqrt{559}\sqrt{2263}}$$

$$= \frac{1107}{\sqrt{1265017}} = \frac{1107}{1124.72} = 0.98$$

Aliter: $\bar{X} = 15.62,\ \bar{Y} = 42.12,\ Ax = 16,\ Ay = 42,\ \Sigma dxdy = 138$

$$\sigma_x = \sqrt{\frac{\Sigma dx^2}{N} - \left(\frac{\Sigma dx}{N}\right)^2} = \sqrt{\frac{71}{8} - \left(\frac{-3}{8}\right)^2} = \sqrt{\frac{71}{8} - \frac{9}{64}} = 2.95$$

$$\sigma_y = \sqrt{\frac{\Sigma dy^2}{N} - \left(\frac{\Sigma dy}{N}\right)^2} = \sqrt{\frac{283}{8} - \left(\frac{1}{8}\right)^2} = \sqrt{\frac{283}{8} - \frac{1}{64}} = 5.94$$

Applying the formula

$$r = \frac{\Sigma dxdy - N(\bar{X} - Ax)(\bar{Y} - Ay)}{N.\sigma_x.\sigma_y}$$

$$= \frac{138 - 8(15.62 - 16)(42.12 - 42)}{8 \times 2.95 \times 5.94} = \frac{138 - 8(-0.38)(0.12)}{140.184}$$

$$= \frac{138 - 0.3648}{140.184} = \frac{137.6352}{140.184} = 0.98$$

Example 8. Calculate Karl Pearson's coefficient of correlation from the following data:

X:	24	27	28	28	29	30	32	33	35	35	40
Y:	18	20	22	25	22	28	28	30	27	30	22

(You may use 32 as working mean for X and 25 that for Y.)

Solution:

Calculation of Coefficient of Correlation

X	A=32 dx	dx^2	Y	A=25 dy	dy^2	$dxdy$
24	−8	64	18	−7	49	56
27	−5	25	20	−5	25	25
28	−4	16	22	−3	9	12
28	−4	16	25 = A	0	0	0
29	−3	9	22	−3	9	9
30	−2	4	28	+3	9	−6
32 = A	0	0	28	+3	9	0
33	1	1	30	+5	25	5
35	3	9	27	+2	4	6
35	3	9	30	+5	25	15
40	8	64	22	−3	9	−24
N = 11	Σdx = −11	Σdx^2 = 217		Σdy = −3	Σdy^2 = 173	Σdxdy = [illegible]

$$r=\frac{N\times\Sigma dxdy-\Sigma dx\,\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\,\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$

$$=\frac{10\times 897-(-8)(-24)}{\sqrt{10\times 804-(-8)^2}\,\sqrt{10\times 1262-(-24)^2}}$$

$$=\frac{8970-192}{\sqrt{8040-64}\,\sqrt{12620-576}}=\frac{8778}{\sqrt{7976}\,\sqrt{12044}}$$

$$=\frac{8778}{\sqrt{96062944}}=\frac{8778}{9801.17}=0.895$$

Calculation of Coefficient of Correlation by taking a Common Factor

Common factor may be used to simplify the calculation of coefficient of correlation. It is important to note here that there will be no effects on the formula of coefficient of correlation if this common factor is used. The main reason is that the coefficient of correlation is independent of the change of origin and scale. If the origin is shifted or scale is changed, it will not affect the value of coefficient of correlation.

Example 10. Calculate coefficient of correlation from the following data:

X:	100	200	300	400	500	600
Y:	110	120	135	140	160	165

Solution: To simplify the calculation, let

$$dx=\frac{X-400}{100},\quad dy=\frac{Y-140}{5}$$

Calculation of Coefficient of Correlation

X	dx	dx^2	Y	dy	dy^2	dxdy
100	−3	9	110	−6	36	18
200	−2	4	120	−4	16	8
300	−1	1	135	−1	1	1
400	0	0	140	0	0	0
500	+1	1	160	4	16	4
600	+2	4	165	5	25	10
N = 6	$\Sigma dx=-3$	$\Sigma dx^2=19$		$\Sigma dy=-2$	$\Sigma dy^2=94$	$\Sigma dxdy=41$

$$r=\frac{N\times\Sigma dxdy-\Sigma dx\,\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\,\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$

$$=\frac{6\times 41-(-3)(-2)}{\sqrt{6\times 19-(-3)^2}\times\sqrt{6\times 94-(-2)^2}}$$

$$=\frac{246-6}{\sqrt{105}\,\sqrt{560}}=\frac{240}{\sqrt{58800}}=\frac{240}{242.487}=0.9897$$

IMPORTANT TYPICAL EXAMPLES

Example 11. From the following data, calculate the Karl Pearso[n]… between age of students and their playing habits:

Age:	15	16	17
No. of students:	250	200	150
Regular players:	200	150	90

Solution: Since it is asked to find the correlation between age a… to find the percentage of regular players which is ob…

No. of students	Regular players
250	200
200	150
150	90
120	48
100	30
80	12

Now we calculate the correlation coefficient betw… players. Denoting the age by X and percentage o…

X	dx	dx^2	Y	dy
15	−2	4	80	+20
16	−1	1	75	+15
17 = A	0	0	60 = A	0
18	+1	1	40	−20
19	+2	4	30	−30
20	+3	9	15	−45
N = 6	$\Sigma dx=3$	$\Sigma dx^2=19$		$\Sigma dy=$

$$r=\frac{N\times\Sigma dxdy-\ldots}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\ldots}$$

$$=\frac{6\times(-270)-(3\ldots}{\sqrt{6\times 19-(3)^2}\times\sqrt{6\times\ldots}}$$

$$r=\frac{\Sigma dxdy-\frac{\Sigma dx\,.\,\Sigma dy}{N}}{\sqrt{\Sigma dx^2-\frac{(\Sigma dx)^2}{N}}\,\sqrt{\Sigma dy^2-\frac{(\Sigma dy)^2}{N}}}$$

$$=\frac{98-\frac{(-11)(-3)}{11}}{\sqrt{217-\frac{(-11)^2}{11}}\times\sqrt{173-\frac{(-3)^2}{11}}}=\frac{98-3}{\sqrt{217-11}\,\sqrt{173-0.82}}$$

$$=\frac{95}{\sqrt{206\times 172.18}}=\frac{95}{188.33}=0.504$$

Aliter: r can be calculated by using the formula:

$$r=\frac{\Sigma dxdy\times N-\Sigma dx\,\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\,\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$

$$=\frac{98\times 11-(-11)(-3)}{\sqrt{11\times 217-(-11)^2}\times\sqrt{11\times 173-(-3)^2}}$$

$$=\frac{1078-33}{\sqrt{2387-121}\,\sqrt{1903-9}}=\frac{1045}{\sqrt{2266}\,\sqrt{1894}}$$

$$=\frac{1045}{\sqrt{4291804}}=\frac{1045}{2071.66}=0.504$$

Example 9. Deviations of the items of two series X and Y from assumed mean are as under:

Deviations of X:	+5	−4	−2	+20	−10	0	+3	0	−15	−5
Deviations of Y:	+5	−12	−7	+25	−10	−3	0	+2	−9	−15

Calculate Karl Pearson's coefficient of correlation.

Solution:

dx	dx^2	dy	dy^2	dxdy
+5	25	+5	25	25
−4	16	−12	144	48
−2	4	−7	49	14
+20	400	+25	625	500
−10	100	−10	100	100
0	0	−3	9	0
+3	9	0	0	0
0	0	+2	4	0
−15	225	−9	81	135
−5	25	−15	225	75
$\Sigma dx=-8$	$\Sigma dx^2=804$	$\Sigma dy=-24$	$\Sigma dy^2=1262$	$\Sigma dxdy=897$

$$= \frac{-1620+180}{\sqrt{114-9}\sqrt{23700-3600}} = \frac{-1440}{\sqrt{2110500}} = \frac{-1440}{1452.75} - 0.99$$

There is a high degree of negative correlation between age and playing habits. It shows that as age increases, the tendency to play decreases.

Example 12. From the following data, calculate Karl Pearson's coefficient of correlation between age and blindness:

Age	No. of persons (in thousands)	Blinds
0—10	100	55
10—20	60	40
20—30	40	40
30—40	36	40
40—50	24	36
50—60	11	22
60—70	6	18
70—80	3	15

Solution: First, we shall find the number of blinds per lakh of population in each group as:

No. of persons ('000)	Blinds	No. of blinds (per lakh)
100	55	$\frac{55}{100000}\times 100000 = 55$
60	40	$\frac{40}{60000}\times 100000 = 67$
40	40	$\frac{40}{40000}\times 100000 = 100$
36	40	$\frac{40}{36000}\times 100000 = 111$
24	36	$\frac{36}{24000}\times 100000 = 150$
11	22	$\frac{22}{11000}\times 100000 = 200$
6	18	$\frac{18}{6000}\times 100000 = 300$
3	15	$\frac{15}{3000}\times 100000 = 500$

Denoting the Mid Value of Age by X and No. of Blinds per lakh by Y, we find coefficient of correlation.

Age	MV (X)	A = 35, $dx = \frac{X-35}{10}$	dx^2	Y	A=185, $dy = Y-185$
0—10	5	−3	9	55	−130
10—20	15	−2	4	67	−118
20—30	25	−1	1	100	−85
30—40	35	0	0	111	−74
40—50	45	+1	1	150	−35
50—60	55	+2	4	200	+15
60—70	65	+3	9	300	+115
70—80	75	+4	16	500	+315
$N = 8$		$\Sigma dx = 4$	$\Sigma dx^2 = 44$		$\Sigma dy = 3$

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx . \Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}} = \frac{}{\sqrt{44 -}}$$

$$= \frac{2311 - \frac{(4)(3)}{8}}{\sqrt{44 - \frac{(4)^2}{8}}\sqrt{157425 - \frac{(3)^2}{8}}} = \frac{2311-1.5}{\sqrt{42}\sqrt{157423.87}}$$

$$= \frac{2309.5}{\sqrt{42}\sqrt{157423.87}} = \frac{2309.5}{6.48\times 396.76} = \frac{2309.5}{2571.004} = +0$$

Example 13. From the following data, calculate the coefficient of cor
Y-series.

	X-series
Mean	74.5
Assumed mean	69
Standard deviation (σ)	13.07

Sum of products of corresponding deviations of X ar
mean $(\Sigma\, dxdy) = 2176$ and no. of pairs of observatio

Solution: Given:

$N = 8,\ \bar{X} = 74.5,\ A_x = 69,\ \sigma_x = 13.07,$

$\bar{Y} = 125.5,\ A_y = 112,\ \sigma_y = 15.85,\ \Sigma dxdy = 2176$

Applying the formula:

$$r = \frac{\Sigma dxdy - N(\bar{X} - A_x)(\bar{Y} - A}{N . \sigma_x . \sigma_y}$$

$$\frac{N \times \Sigma dxdy - \Sigma dx\ \Sigma dy}{x^2 - (\Sigma dx)^2\ \sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$\frac{10\times 897 - (-8)(-24)}{804 - (-8)^2\ \sqrt{10\times 1262 - (-24)^2}}$$

$$\frac{8970-192}{-64\ \sqrt{12620-576}} = \frac{8778}{\sqrt{7976}\sqrt{12044}}$$

$$\frac{78}{62944} = \frac{8778}{9801.17} = 0.895$$

relation by taking a Common Factor

plify the calculation of coefficient of correlation
effects on the formula of coefficient of correlation
is that the coefficient of correlation is independent
shifted or scale is changed, it will not affect the v

rrelation from the following data:

200	300	400	500	
120	135	140	160	

on, let

$dy = \frac{Y-140}{5}$

lation of Coefficient of Correlation

dx^2	Y	dy	dy^2
9	110	−6	36
4	120	−4	16
1	135	−1	1
0	140	0	0
1	160	4	16
4	165	5	25
$\Sigma dx^2 = 19$		$\Sigma dy = -2$	$\Sigma dy^2 = 94$

$$= \frac{N \times \Sigma dxdy - \Sigma dx\ \Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{6\times 41 - (-3)(-2)}{\sqrt{6\times 19 - (-3)^2}\times\sqrt{6\times 94 - (-2)^2}}$$

$$= \frac{246-6}{\sqrt{105}\sqrt{560}} = \frac{240}{\sqrt{58800}} = \frac{240}{242.487} = 0.9897$$

IMPORTANT TYPICAL EXAMPLES

Example 11. From the following data, calculate the Karl Pearson's coefficient of correlation between age of students and their playing habits:

Age:	15	16	17	18	19	20
No. of students:	250	200	150	120	100	80
Regular players:	200	150	90	48	30	12

Solution: Since it is asked to find the correlation between age and playing habits, it is required to find the percentage of regular players which is obtained as follows:

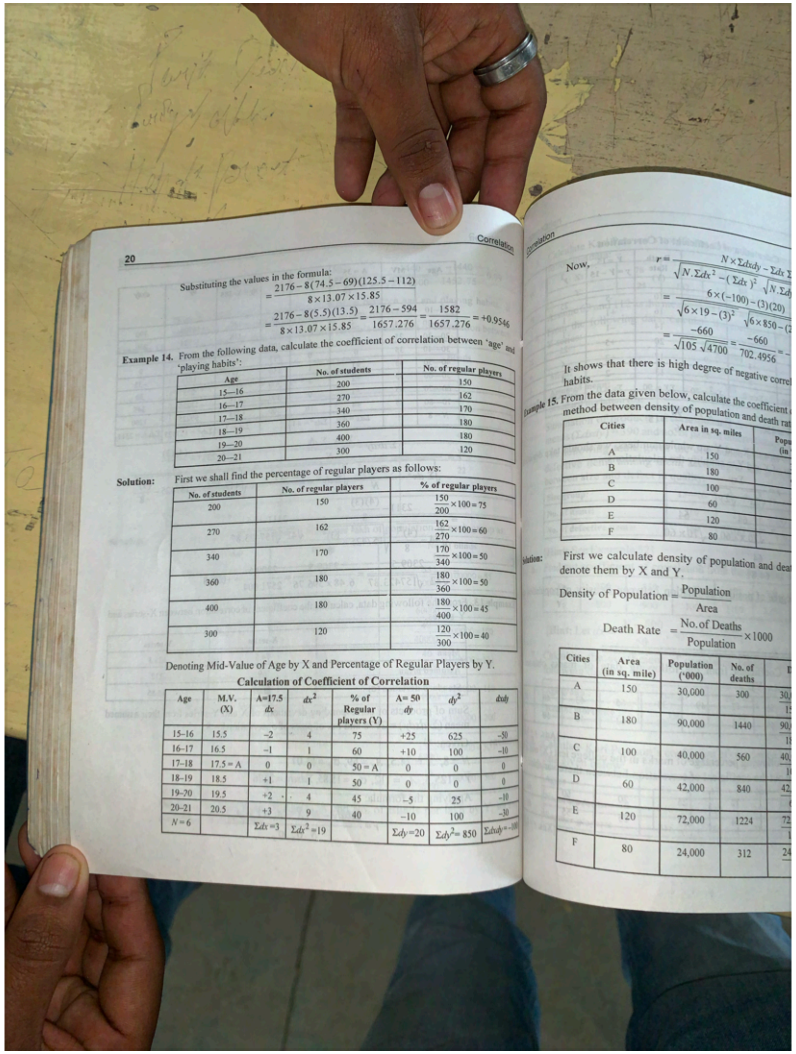
No. of students	Regular players	% of Regular players
250	200	$\frac{200}{250}\times 100 = 80$
200	150	$\frac{150}{200}\times 100 = 75$
150	90	$\frac{90}{150}\times 100 = 60$
120	48	$\frac{48}{120}\times 100 = 40$
100	30	$\frac{30}{100}\times 100 = 30$
80	12	$\frac{12}{80}\times 100 = 15$

Now we calculate the correlation coefficient between age and percentage of regular players. Denoting the age by X and percentage of regular players by Y.

X	dx	dx^2	Y	dy	dy^2	$dxdy$
15	−2	4	80	+20	400	−40
16	−1	1	75	+15	225	−15
17 = A	0	0	60 = A	0	0	0
18	+1	1	40	−20	400	−20
19	+2	4	30	−30	900	−60
20	+3	9	15	−45	2025	−135
N = 6	$\Sigma dx = 3$	$\Sigma dx^2 = 19$		$\Sigma dy = -60$	$\Sigma dy^2 = 3950$	$\Sigma dxdy = -270$

$$r = \frac{N \times \Sigma dxdy - \Sigma dx\ \Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{6\times(-270) - (3)(-60)}{\sqrt{6\times 19 - (3)^2}\times\sqrt{6\times 3950 - (-60)^2}}$$


20 Correlation

Substituting the values in the formula:

$$= \frac{2176 - 8(74.5-69)(125.5-112)}{8 \times 13.07 \times 15.85}$$

$$= \frac{2176 - 8(5.5)(13.5)}{8 \times 13.07 \times 15.85} = \frac{2176-594}{1657.276} = \frac{1582}{1657.276} = +0.9546$$

Example 14. From the following data, calculate the coefficient of correlation between 'age' and 'playing habits':

Age	No. of students	No. of regular players
15—16	200	150
16—17	270	162
17—18	340	170
18—19	360	180
19—20	400	180
20—21	300	120

Solution: First we shall find the percentage of regular players as follows:

No. of students	No. of regular players	% of regular players
200	150	$\frac{150}{200}\times 100 = 75$
270	162	$\frac{162}{270}\times 100 = 60$
340	170	$\frac{170}{340}\times 100 = 50$
360	180	$\frac{180}{360}\times 100 = 50$
400	180	$\frac{180}{400}\times 100 = 45$
300	120	$\frac{120}{300}\times 100 = 40$

Denoting Mid-Value of Age by X and Percentage of Regular Players by Y.

Calculation of Coefficient of Correlation

Age	M.V. (X)	A=17.5 dx	dx^2	% of Regular players (Y)	A= 50 dy	dy^2	$dxdy$
15–16	15.5	–2	4	75	+25	625	–50
16–17	16.5	–1	1	60	+10	100	–10
17–18	17.5 = A	0	0	50 = A	0	0	0
18–19	18.5	+1	1	50	0	0	0
19–20	19.5	+2	4	45	–5	25	–10
20–21	20.5	+3	9	40	–10	100	–30
$N=6$		$\Sigma dx = 3$	$\Sigma dx^2 = 19$		$\Sigma dy = 20$	$\Sigma dy^2 = 850$	$\Sigma dxdy = -10$[illegible]

Correlation

Now,

$$r = \frac{N \times \Sigma dxdy - \Sigma dx\ \Sigma}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N.\Sigma d}}$$

$$= \frac{6 \times (-100) - (3)(20)}{\sqrt{6 \times 19 - (3)^2}\ \sqrt{6 \times 850 - (2}}$$

$$= \frac{-660}{\sqrt{105}\ \sqrt{4700}} = \frac{-660}{702.4956} = -$$

It shows that there is high degree of negative correl habits.

Example 15. From the data given below, calculate the coefficient method between density of population and death rat

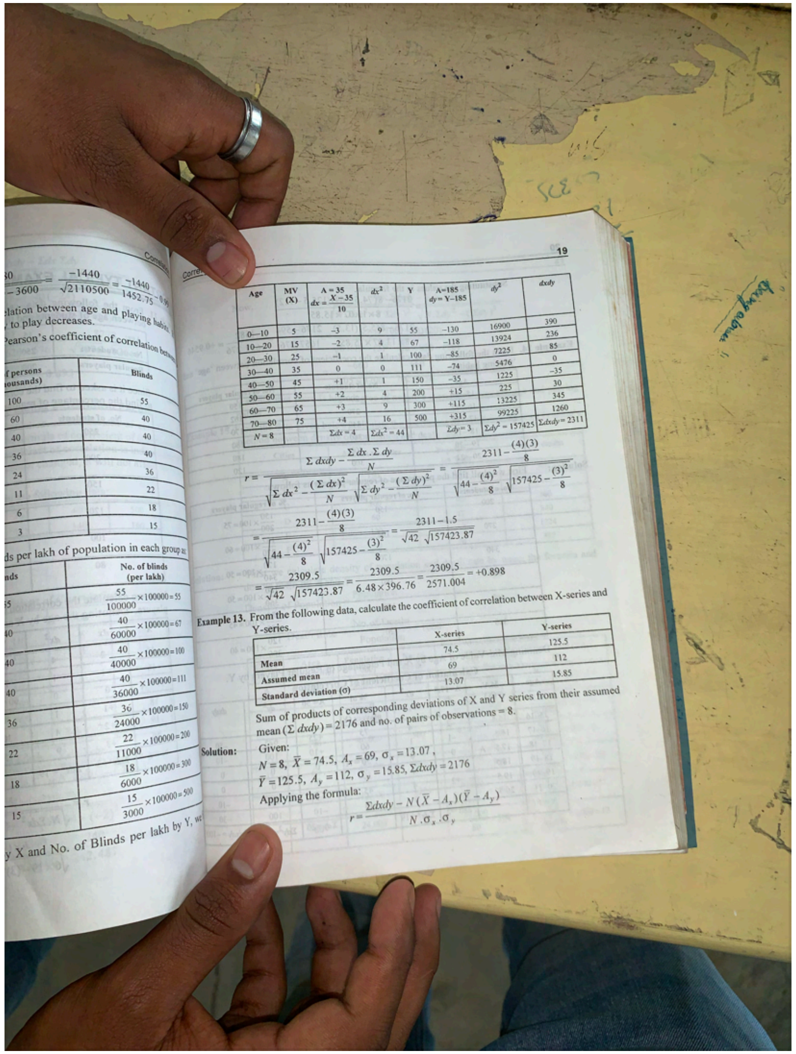
Cities	Area in sq. miles	Popu (in
A	150	
B	180	
C	100	
D	60	
E	120	
F	80	

Solution: First we calculate density of population and deat denote them by X and Y.

$$\text{Density of Population} = \frac{\text{Population}}{\text{Area}}$$

$$\text{Death Rate} = \frac{\text{No. of Deaths}}{\text{Population}} \times 1000$$

Cities	Area (in sq. mile)	Population ('000)	No. of deaths	D
A	150	30,000	300	30, 15
B	180	90,000	1440	90, 18
C	100	40,000	560	40, 10
D	60	42,000	840	42, 6
E	120	72,000	1224	72, 1
F	80	24,000	312	24


$$\frac{\ldots 0}{\ldots -3600} = \frac{-1440}{\sqrt{2110500}} = \frac{-1440}{1452.75} = -0.99$$

...lation between age and playing habits ... to play decreases.

...Pearson's coefficient of correlation betw...

...f persons ...ousands)	Blinds
100	55
60	40
40	40
36	40
24	36
11	22
6	18
3	15

...ds per lakh of population in each group as

...nds	No. of blinds (per lakh)
55	$\frac{55}{100000}\times 100000 = 55$
40	$\frac{40}{60000}\times 100000 = 67$
40	$\frac{40}{40000}\times 100000 = 100$
40	$\frac{40}{36000}\times 100000 = 111$
36	$\frac{36}{24000}\times 100000 = 150$
22	$\frac{22}{11000}\times 100000 = 200$
18	$\frac{18}{6000}\times 100000 = 300$
15	$\frac{15}{3000}\times 100000 = 500$

...y X and No. of Blinds per lakh by Y, we

19

Age	MV (X)	A = 35 $dx = \frac{X-35}{10}$	dx^2	Y	A=185 $dy = Y-185$	dy^2	$dxdy$
0—10	5	–3	9	55	–130	16900	390
10—20	15	–2	4	67	–118	13924	236
20—30	25	–1	1	100	–85	7225	85
30—40	35	0	0	111	–74	5476	0
40—50	45	+1	1	150	–35	1225	–35
50—60	55	+2	4	200	+15	225	30
60—70	65	+3	9	300	+115	13225	345
70—80	75	+4	16	500	+315	99225	1260
$N = 8$		$\Sigma dx = 4$	$\Sigma dx^2 = 44$		$\Sigma dy = 3$	$\Sigma dy^2 = 157425$	$\Sigma dxdy = 2311$

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx . \Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\ \sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}} = \frac{2311 - \frac{(4)(3)}{8}}{\sqrt{44 - \frac{(4)^2}{8}}\ \sqrt{157425 - \frac{(3)^2}{8}}}$$

$$= \frac{2311 - \frac{(4)(3)}{8}}{\sqrt{44 - \frac{(4)^2}{8}}\ \sqrt{157425 - \frac{(3)^2}{8}}} = \frac{2311 - 1.5}{\sqrt{42}\ \sqrt{157423.87}}$$

$$= \frac{2309.5}{\sqrt{42}\ \sqrt{157423.87}} = \frac{2309.5}{6.48 \times 396.76} = \frac{2309.5}{2571.004} = +0.898$$

Example 13. From the following data, calculate the coefficient of correlation between X-series and Y-series.

	X-series	Y-series
Mean	74.5	125.5
Assumed mean	69	112
Standard deviation (σ)	13.07	15.85

Sum of products of corresponding deviations of X and Y series from their assumed mean $(\Sigma\, dxdy) = 2176$ and no. of pairs of observations = 8.

Solution: Given:

$N = 8,\ \bar{X} = 74.5,\ A_x = 69,\ \sigma_x = 13.07,$

$\bar{Y} = 125.5,\ A_y = 112,\ \sigma_y = 15.85,\ \Sigma dxdy = 2176$

Applying the formula:

$$r = \frac{\Sigma dxdy - N(\bar{X} - A_x)(\bar{Y} - A_y)}{N . \sigma_x . \sigma_y}$$

Now,
$$r=\frac{N\times\Sigma dxdy-\Sigma dx\,\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$
$$=\frac{6\times(-100)-(3)(20)}{\sqrt{6\times19-(3)^2}\sqrt{6\times850-(20)^2}}$$
$$=\frac{-660}{\sqrt{105}\sqrt{4700}}=\frac{-660}{702.4956}=-0.9395$$

It shows that there is high degree of negative correlation between age and playing habits.

Example 15. From the data given below, calculate the coefficient of correlation by Karl Pearson's method between density of population and death rate:

Cities	Area in sq. miles	Population (in '000)	No. of deaths
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840
E	120	72	1224
F	80	24	312

Solution: First we calculate density of population and death rate by using the formula and denote them by X and Y.

$$\text{Density of Population}=\frac{\text{Population}}{\text{Area}}$$
$$\text{Death Rate}=\frac{\text{No. of Deaths}}{\text{Population}}\times1000$$

Cities	Area (in sq. mile)	Population ('000)	No. of deaths	Density (X)	Death rate (%) (Y)
A	150	30,000	300	$\frac{30,000}{150}=200$	$\frac{300}{30,000}\times1,000=10$
B	180	90,000	1440	$\frac{90,000}{180}=500$	$\frac{1,440}{90,000}\times1,000=16$
C	100	40,000	560	$\frac{40,000}{100}=400$	$\frac{560}{40,000}\times1,000=14$
D	60	42,000	840	$\frac{42,000}{60}=700$	$\frac{840}{42,000}\times1,000=20$
E	120	72,000	1224	$\frac{72,000}{120}=600$	$\frac{1224}{72,000}\times1,000=17$
F	80	24,000	312	$\frac{24,000}{80}=300$	$\frac{312}{24,000}\times1,000=13$

Calculation of Coefficient of Correlation

Cities	Density (X)	$\bar{X}=450$, $x=\frac{X-450}{50}$	x^2	Death Rate (Y)	$\bar{Y}=15$, $y=Y-15$	y^2	xy
A	200	−5	25	10	−5	25	25
B	500	+1	1	16	+1	1	1
C	400	−1	1	14	−1	1	1
D	700	+5	25	20	+5	25	25
E	600	+3	9	17	+2	4	6
F	300	−3	9	13	−2	4	6
$N=6$	$\Sigma X=2700$	$\Sigma x=0$	$\Sigma x^2=70$	$\Sigma Y=90$	$\Sigma y=0$	$\Sigma y^2=60$	$\Sigma xy=64$

$$\bar{X}=\frac{\Sigma X}{N}=\frac{2700}{6}=450,\qquad \bar{Y}=\frac{\Sigma Y}{N}=\frac{90}{6}=15$$

Since the actual means of X and Y are whole numbers, we should take deviations from actual means of X and Y to simplify the calculations:

$$r=\frac{\Sigma xy}{\sqrt{\Sigma x^2}.\sqrt{\Sigma y^2}}$$
$$=\frac{64}{\sqrt{70}\times\sqrt{60}}=\frac{64}{\sqrt{70\times60}}$$
$$=\frac{64}{64.81}=+0.9875$$

There is a high degree of positive correlation between density of population and death rate.

EXERCISE 1.4

1. Calculate the Correlation Coefficient from the following data of marks obtained in Commerce (X) and Economics (Y):

X:	50	60	58	47	49	33	65	43	46	68
Y:	48	65	50	48	55	58	63	48	50	70

[Ans. $r=0.61$]

2. Seven students obtained the following percentage of marks in the college test (X) and in the final examination (Y). Find out the coefficient of correlation between these variables:

X:	50	62	72	25	20	60	60
Y:	48	65	74	33	25	55	66

[Ans. $r=0.9$

3. Calculate Karl Pearson's coefficient of correlation following data:

X:	78	89	96	69
Y:	125	137	156	112

Assume 69 and 112 as the mean values for X and Y

4. From the following data, calculate the coefficie Y-series:

	X-series
Mean:	381.2
Assumed mean:	380
Standard deviation (σ):	16.79

Summation of products of corresponding deviati means ($\Sigma dxdy$) = 390 and no. of pairs of observati

5. The following table gives the distribution of i defective items among them, according to size between size and defect in quality.

Size group:	15—16	16—17
No. of items:	200	270
No. of defective items:	150	162

[Hint: See Example 52]

6. Find out coefficient of correlation from the foll

X:	300	350	400	450
Y:	800	900	1000	1100

[Hint: Let $dx=\frac{X-500}{50}$, $dy=\frac{Y-1200}{100}$]

7. Calculate the coefficient of correlation betw following data :

Age group :	0—20	20—40
Rate of mortality :	350	280

8. Calculate Karl Pearson's coefficient of correl data given below:

Age:	16	17
No. of students:	350	320
Regular players:	315	256

3. Calculate Karl Pearson's coefficient of correlation between the values of X and Y for the following data:

X:	78	89	96	69	59	79	68	61
Y:	125	137	156	112	107	136	123	108

Assume 69 and 112 as the mean values for X and Y respectively. [Ans. $r = +0.954$]

4. From the following data, calculate the coefficient of correlation between X-series and Y-series:

	X-series	Y-series
Mean:	381.2	24.5
Assumed mean:	380	25
Standard deviation (σ):	16.79	2.97

Summation of products of corresponding deviations of X and Y series from their assumed means $(\Sigma dxdy) = 390$ and no. of pairs of observations $= 10$. [Ans. $r = 0.794$]

5. The following table gives the distribution of items of production and also the relative defective items among them, according to size groups. Find the correlation coefficient between size and defect in quality.

Size group:	15—16	16—17	17—18	18—19	19—20	20—21
No. of items:	200	270	340	360	400	300
No. of defective items:	150	162	170	180	180	114

[Hint: See Example 52] [Ans. $r = -0.95$]

6. Find out coefficient of correlation from the following data:

X:	300	350	400	450	500	550	600	650	700
Y:	800	900	1000	1100	1200	1300	1400	1500	1600

[Hint: Let $dx = \frac{X - 500}{50}$, $dy = \frac{Y - 1200}{100}$] [Ans. $r = +1$]

7. Calculate the coefficient of correlation between age group and mortality rate from the following data :

Age group :	0—20	20—40	40—60	60—80	80—100
Rate of mortality :	350	280	540	760	900

[Ans. $r = 0.947$]

8. Calculate Karl Pearson's coefficient of correlation between age and playing habits from the data given below:

Age:	16	17	18	19	20	21	22
No. of students:	350	320	280	240	180	120	50
Regular players:	315	256	182	132	63	18	4

[Ans. $r = -0.994$]

9. Following figures give the rainfall in inches and production in '00 tons for Rabi and Kharif crops for number of years. Find the coefficient of correlation between rainfall and total production:

Rainfall:	20	22	24	26	28	30	32
Rabi production:	15	18	20	32	40	39	40
Kharif production:	15	17	20	18	20	21	15

[**Ans.** $r = +0.917$]

10. With the following data in 4 cities, calculate the coefficient of correlation by Pearson's method between the density of population and the death rate:

Cities	Area in sq.km.	Population ('000)	No. of deaths
A	200	40	480
B	150	75	1200
C	120	72	1080
D	80	20	280

[**Ans.** $r = +0.821$]

11. Calculate 'r' from the following data:
$\Sigma X = 225, \Sigma Y = 189, N = 10, \Sigma(X-22)^2 = 85, \Sigma(Y-19)^2 = 25$ and $\Sigma(X-22)(Y-19) = 43$
[**Hint:** See Example 53 Aliter] [**Ans.** $r = 0.96$]

(3) Method Based on the Use of Actual Data

This method is also known as **Product moment method**. When number of observations are few, correlation coefficient can also be calculated without taking deviations either from actual mean or from assumed mean *i.e.* from actual X and Y values. In this method, the correlation coefficient can be determined in the following way:

(1) First of all, values of the variables X and Y series are summed up to get ΣX and ΣY.

(2) The values of the variables of X and Y series are squared up and added to get ΣX^2 and ΣY^2.

(3) The values of X variable and Y variable are multiplied and the product is added up to get ΣXY.

(4) Finally, the following formula is used to get the correlation coefficient:

$$r = \frac{\Sigma XY - \dfrac{\Sigma X . \Sigma Y}{N}}{\sqrt{\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}}\sqrt{\Sigma Y^2 - \dfrac{(\Sigma Y)^2}{N}}}$$

or

$$r = \frac{N . \Sigma XY - \Sigma X . \Sigma Y}{\sqrt{N . \Sigma X^2 - (\Sigma X)^2}\sqrt{N . \Sigma Y^2 - (\Sigma Y)^2}}$$

Example 16. From the following data, find Karl Pearson

X:	2	3	1
Y:	4	5	3

Solution:

Calculation of Coeffici

X	X^2	Y
2	4	4
3	9	5
1	1	3
5	25	
6	36	
4	16	
$N = 6, \Sigma X = 21$	$\Sigma X^2 = 91$	ΣY

Applying the formula:

$$r = \frac{N . \Sigma X}{\sqrt{N . \Sigma X^2 - (\Sigma X}$$

$$= \frac{6 \times 90 - (}{\sqrt{6 \times 91 - (21)^2}}$$

$$= \frac{36}{\sqrt{105}\sqrt{60}} = \frac{0}{\sqrt{6}}$$

Example 17. Calculate product moment correlation c

X:	−5	−10
Y:	50	40

Solution: In this question the mean of X and Y ser will pose a problem in computing dev actual values will be used.

Calculation of C

X	X^2
−5	25
−10	100
−15	225
−20	400
−25	625
−30	900
$\Sigma X = -105$ $N = 6$	$\Sigma X^2 = 2275$