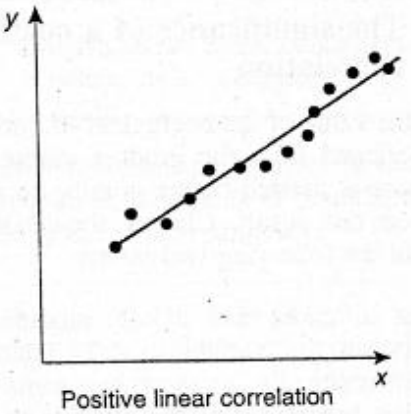
35

Linear correlation

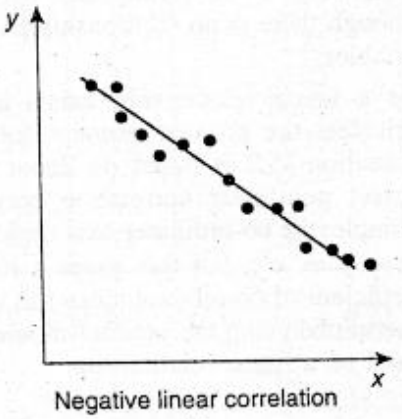
35.1 Introduction to linear correlation

Correlation is a measure of the amount of association existing between two variables. For linear correlation, if points are plotted on a graph and all the points lie on a straight line, then **perfect linear correlation** is said to exist. When a straight line having a positive gradient can reasonably be drawn through points on a graph **positive or direct linear correlation** exists, as shown in Fig. 35.1(a). Similarly, when a straight line having a negative gradient can reasonably be drawn through points on a graph, **negative or inverse linear correlation** exists, as shown in Fig. 35.1(b). When there is no apparent relationship between co-ordinate values plotted on a graph then no **correlation** exists between the points, as shown in Fig. 35.1(c). In statistics, when two variables are being investigated, the location of the co-ordinates on a rectangular co-ordinate system is called a **scatter diagram** — as shown in Fig. 35.1.



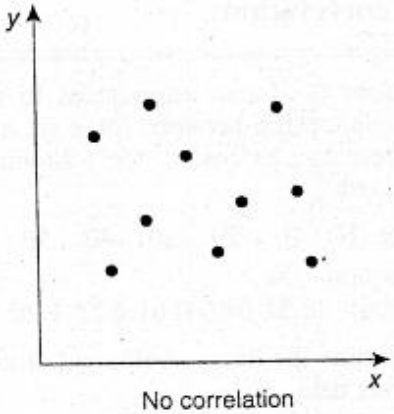
Positive linear correlation

(a)



Negative linear correlation

(b)



No correlation

(c)

Figure 35.1

35.2 The product-moment formula for determining the linear correlation coefficient

The amount of linear correlation between two variables is expressed by a **coefficient of correlation**, given the symbol r. This is defined in terms of the deviations of the co-ordinates of two variables from their mean values and is given by the **product-moment formula** which states:

coefficient of correlation,

$$r = \frac{\sum xy}{\sqrt{\{(\sum x^2)(\sum y^2)\}}} \qquad (1)$$

where the x-values are the values of the deviations of co-ordinates X from $\overline{X}$, their mean value and the y-values are the values of the deviations of co-ordinates Y from $\overline{Y}$, their mean value. That is, $x = (X - \overline{X})$ and $y = (Y - \overline{Y})$. The results of this determination give values of r lying between $+1$ and -1,

where +1 indicates perfect direct correlation, −1 indicates perfect inverse correlation and 0 indicates that no correlation exists. Between these values, the smaller the value of r, the less is the amount of correlation which exists. Generally, values of r in the ranges 0.7 to 1 and −0.7 to −1 show that a fair amount of correlation exists.

35.3 The significance of a coefficient of correlation

When the value of the coefficient of correlation has been obtained from the product moment formula, some care is needed before coming to conclusions based on this result. Checks should be made to ascertain the following two points:

(a) that a 'cause and effect' relationship exists between the variables; it is relatively easy, mathematically, to show that some correlation exists between, say, the number of ice creams sold in a given period of time and the number of chimneys swept in the same period of time, although there is no relationship between these variables;

(b) that a linear relationship exists between the variables; the product-moment formula given in Section 35.2 is based on linear correlation. Perfect non-linear correlation may exist (for example, the co-ordinates exactly following the curve $y = x^3$), but this gives a low value of coefficient of correlation since the value of r is determined using the product-moment formula, based on a linear relationship.

35.4 Worked problems on linear correlation

Problem 1. In an experiment to determine the relationship between force on a wire and the resulting extension, the following data is obtained:

Force (N)	10	20	30	40	50	60	70
Extension (mm)	0.22	0.40	0.61	0.85	1.20	1.45	1.70

Determine the linear coefficient of correlation for this data.

Let X be the variable force values and Y be the dependent variable extension values. The coefficient of correlation is given by:

$$r = \frac{\sum xy}{\sqrt{\{(\sum x^2)(\sum y^2)\}}}$$

where $x = (X - \overline{X})$ and $y = (Y - \overline{Y})$, $\overline{X}$ and $\overline{Y}$ being the mean values of the X and Y values respectively. Using a tabular method to determine the quantities of this formula gives:

X	Y	$x = (X - \overline{X})$	$y = (Y - \overline{Y})$
10	0.22	−30	−0.699
20	0.40	−20	−0.519
30	0.61	−10	−0.309
40	0.85	0	−0.069
50	1.20	10	0.281
60	1.45	20	0.531
70	1.70	30	0.781

$\sum X = 280, \quad \overline{X} = \frac{280}{7} = 40$

$\sum Y = 6.43, \quad \overline{Y} = \frac{6.43}{7} = 0.919$

xy	x^2	y^2
20.97	900	0.489
10.38	400	0.269
3.09	100	0.095
0	0	0.005
2.81	100	0.079
10.62	400	0.282
23.43	900	0.610
$\sum xy = 71.30$	$\sum x^2 = 2800$	$\sum y^2 = 1.829$

Thus $r = \frac{71.3}{\sqrt{[2800 \times 1.829]}} = \mathbf{0.996}$

This shows that a **very good direct correlation exists** between the values of force and extension.

Problem 2. The relationship between expenditure on welfare services and absenteeism for similar periods of time is shown below for a small company.

Expenditure (£'000)	3.5	5.0	7.0	10	12	15	18
Days lost	241	318	174	110	147	122	86

Determine the coefficient of linear correlation for this data.

Let X be the expenditure in thousands of pounds and Y be the days lost.

The coefficient of correlation,

$$r = \frac{\sum xy}{\sqrt{\{(\sum x^2)(\sum y^2)\}}}$$

where $x = (X - \overline{X})$ and $y = (Y - \overline{Y})$, $\overline{X}$ and $\overline{Y}$ being the mean values of X and Y respectively. Using a tabular approach:

X	Y	$x = (X - \overline{X})$	$y = (Y - \overline{Y})$
3.5	241	−6.57	69.9
5.0	318	−5.07	146.9
7.0	174	−3.07	2.9
10	110	−0.07	−61.1
12	147	1.93	−24.1
15	122	4.93	−49.1
18	86	7.93	−85.1

$\sum X = 70.5, \quad \overline{X} = \frac{70.5}{7} = 10.07$

$\sum Y = 1198, \quad \overline{Y} = \frac{1198}{7} = 171.1$

xy	x^2	y^2
−459.2	43.2	4886
−744.8	25.7	21580
−8.9	9.4	8
4.3	0	3733
−46.5	3.7	581
−242.1	24.3	2411
−674.8	62.9	7242
$\sum xy = -2172$	$\sum x^2 = 169.2$	$\sum y^2 = 40441$

Thus

$$r = \frac{-2172}{\sqrt{[169.2 \times 40441]}} = -0.830$$

This shows that there is **fairly good inverse correlation** between the expenditure on welfare and days lost due to absenteeism.

Problem 3. The relationship between monthly car sales and income from the sale of petrol for a garage is as shown:

Cars sold 2 5 3 12 14 7 3 28 14 7 3 13

Income from petrol sales (£'000) 12 9 13 21 17 22 31 47 17 10 9 11

Determine the linear coefficient of correlation between these quantities.

Let X represent the number of cars sold and Y the income, in thousands of pounds, from petrol sales. Using the tabular approach:

X	Y	$x = (X - \overline{X})$	$y = (Y - \overline{Y})$
2	12	−7.25	−6.25
5	9	−4.25	−9.25
3	13	−6.25	−5.25
12	21	2.75	2.75
14	17	4.75	−1.25
7	22	−2.25	3.75
3	31	−6.25	12.75
28	47	18.75	28.75
14	17	4.75	−1.25
7	10	−2.25	−8.25
3	9	−6.25	−9.25
13	11	3.75	−7.25

$\sum X = 111, \quad \overline{X} = \frac{111}{12} = 9.25$

$\sum Y = 219, \quad \overline{Y} = \frac{219}{12} = 18.25$

xy	x^2	y^2
45.3	52.6	39.1
39.3	18.1	85.6
32.8	39.1	27.6
7.6	7.6	7.6
−5.9	22.6	1.6
−8.4	5.1	14.1
−79.7	39.1	162.6
539.1	351.6	826.6
−5.9	22.6	1.6
18.6	5.1	68.1
57.8	39.1	85.6
−27.2	14.1	52.6
$\sum xy = 613.4$	$\sum x^2 = 616.7$	$\sum y^2 = 1372.7$

The coefficient of correlation,

$$r = \frac{\sum xy}{\sqrt{\{(\sum x^2)(\sum y^2)\}}}$$

$$= \frac{613.4}{\sqrt{\{(616.7)(1372.7)\}}} = \mathbf{0.667}$$

Thus, there is **no appreciable correlation** between petrol and car sales.

Now try the following exercise.

Exercise 133 Further problems on linear correlation

In Problems 1 to 3, determine the coefficient of correlation for the data given, correct to 3 decimal places.

1.

X	14	18	23	30	50
Y	900	1200	1600	2100	3800

[0.999]

2.

X	2.7	4.3	1.2	1.4	4.9
Y	11.9	7.10	33.8	25.0	7.50

[−0.916]

3.

X	24	41	9	18	73
Y	39	46	90	30	98

[0.422]

4. In an experiment to determine the relationship between the current flowing in an electrical circuit and the applied voltage, the results obtained are:

Current (mA)	5	11	15	19	24	28	33
Applied voltage (V)	2	4	6	8	10	12	14

Determine, using the product-moment formula, the coefficient of correlation for these results. [0.999]

5. A gas is being compressed in a closed cylinder and the values of pressures and corresponding volumes at constant temperature are as shown:

Pressure (kPa)	Volume (m^3)
160	0.034
180	0.036
200	0.030
220	0.027
240	0.024
260	0.025
280	0.020
300	0.019

Find the coefficient of correlation for these values. [−0.962]

6. The relationship between the number of miles travelled by a group of engineering salesmen in ten equal time periods and the corresponding value of orders taken is given below. Calculate the coefficient of correlation using the product-moment formula for these values.

Miles travelled	Orders taken (£'000)
1370	23
1050	17
980	19
1770	22
1340	27
1560	23
2110	30
1540	23
1480	25
1670	19

[0.632]

7. The data shown below refers to the number of times machine tools had to be taken out of service, in equal time periods, due to faults occurring and the number of hours worked by maintenance teams. Calculate the coefficient of correlation for this data.

Machines out of service:	4	13	2	9	16	8	7
Maintenance hours:	400	515	360	440	570	380	415

[0.937]