Correlation

Specifications:

Calculation and interpretation of the product moment correlation coefficient.

Identification of response (dependent) and explanatory (independent) variables in regression.

Where raw data are given, candidates should be encouraged to obtain correlation coefficient values directly from calculators. Where summarised data are given, candidates may be required to use a formula from the booklet provided for the examination. Calculations from grouped data are excluded. Importance of checking for approximate linear relationship but no hypothesis tests. Understanding that association does not necessarily imply cause and effect.

Given formulae

Correlation and regression

For a set of n pairs of values (x_i, y_i)

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

The product moment correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\{\sum (x_i - \bar{x})^2\}\{\sum (y_i - \bar{y})^2\}}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

Previous chapters have concentrated on developing methods and models for a *single* random variable. This chapter focuses on methods suitable for use with two continuous or discrete random variables, usually denoted by X and Y, where the observations consist of a pair of values (x, y).

Often, all the data are collected at more or less the same time, as exemplified in the following table.

X	Y
Take-off speed of ski-jumper	Distance jumped
Number of red blood cells in blood sample	Number of white blood cells in the sample
Hand span	Foot length
Size of house	Value of house

However, sometimes data are collected on one variable later than the other variable, though the link (the same individual, the same plot of land, the same family, and so on) is clear:

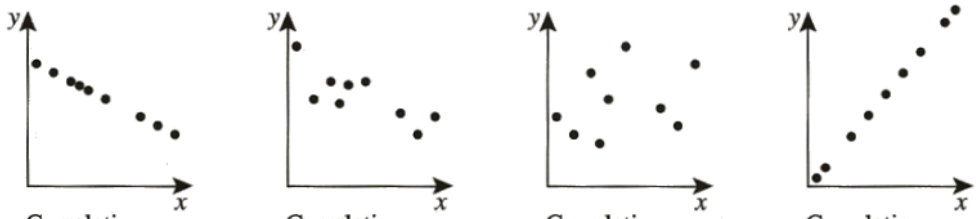
X	Y
Mark in mock examination	Mark in real examination (three months later)
Weight of fertilizer per hectare	Yield of crop in kilograms per hectare
Height of father	Height of son when aged 18

To quantify the extent to which there is an association between X and Y, you can use the idea of **correlation**.

The **product moment correlation coefficient**, r, is a measure of the extent to which any association between the observed values of the random variables X and Y is linear.

The product moment correlation coefficient is sometimes referred to simply as the **correlation coefficient**.

The value of r cannot be less than -1 nor greater than 1. These extremes are only attained when all the data points lie on a straight line, as shown in these diagrams. In other cases, $-1 < r < 1$.



In cases where increasing values of one variable are associated with generally decreasing values of the other variable, $r < 0$ and the variables are said to display **negative correlation**.

In cases where increasing values of one variable are associated with generally increasing values of the other variable, $r > 0$ and the variables are said to display **positive correlation**.

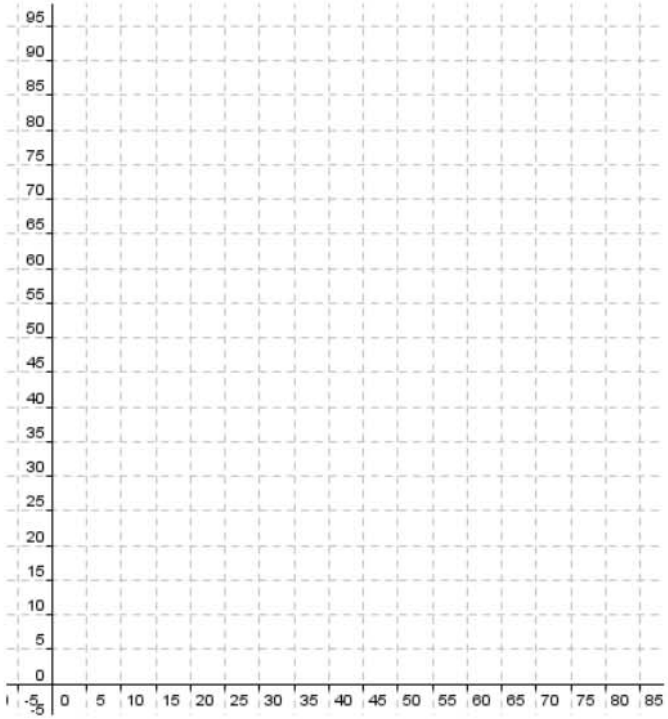
How do we work out the value of r ?

Working out the value of r

A class of students takes examinations in both mathematics and physics. The marks that they obtain are as follows.

Student	1	2	3	4	5	6	7	8	9	10
Mathematics	65	45	40	55	60	50	80	30	70	65
Physics	60	60	55	70	80	40	85	50	70	80

a) Plot the data on a scatter diagram.



b) Complete the table

Student	Maths x	Physics y	xy	x^2	y^2
1	65	60			
2	45	60			
3	40	55			
4	55	70			
5	60	80			
6	50	40			
7	80	85			
8	30	50			
9	70	70			
10	65	80			
Total	560	650	37 850	33 400	44 150
	$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum y^2$

c) Substitute in these formulae and work out

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 37\,850 - \frac{(560 \times 650)}{10} = 1450$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 33\,400 - \frac{560^2}{10} = 2040$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 44\,150 - \frac{650^2}{10} = 1900$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{1450}{\sqrt{2040 \times 1900}} = 0.736\,51 = 0.737 \text{ (to 3 dp)}$$

Formulae (given)

For n pairs of (x_i, y_i) values, the x-values and y-values having means $\bar{x}$ and $\bar{y}$ respectively
The quantities S_{xx}, S_{yy} and S_{xy} are defined as follows:

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

The product moment correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$$

Comment about the values of r

Some suggested descriptions of correlation are:

- When r is between -0.2 and 0.2, the correlation might be described as *weak* or *very weak*.
- When r is between 0.2 and 0.7, or between -0.2 and -0.7, the correlation might be described as *moderate*.
- When r is between 0.7 and 0.9, or between -0.7 and -0.9, the correlation might be described as *strong*.
- When r is above 0.9 or below -0.9 the correlation might be described as *very strong*.

Calculator practice

Many calculators have in-built routines for calculating the value of r. If you have a calculator of this type, make sure that you know how to find the value of r. You could start with the exam mark data in Example 1. Note that these calculators, when calculating r, usually store the values of n, $\sum x$, $\sum x^2$, $\sum y$, $\sum y^2$ and $\sum xy$ in accessible memories. As well as $\bar{x}$ and $\bar{y}$, they often also give $s_x = \sqrt{\frac{S_{xx}}{n-1}}$ and $s_y = \sqrt{\frac{S_{yy}}{n-1}}$, the sample standard deviations of the x-values and the y-values.

Exercises:

5 Given the following summary data,

$\sum x = 367 \quad \sum y = 270 \quad \sum x^2 = 33\,845 \quad \sum y^2 = 12\,976 \quad \sum xy = 17\,135 \quad n = 6$

calculate the product moment correlation coefficient (r) using the formula:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

6 The ages, a years, and heights, h cm, of seven members of a team were recorded. The data were summarised as follows:

$\sum a = 115 \quad \sum a^2 = 1899 \quad S_{hh} = 571.4 \quad S_{ah} = 72.1$

a Find S_{aa}.

b Find the value of the product moment correlation coefficient between a and h.

c Describe and interpret the correlation between the age and height of these seven people based on these data.

5 0.202

6 **a** 9.71
b 0.968
c There is positive correlation. The greater the age, the taller the person.

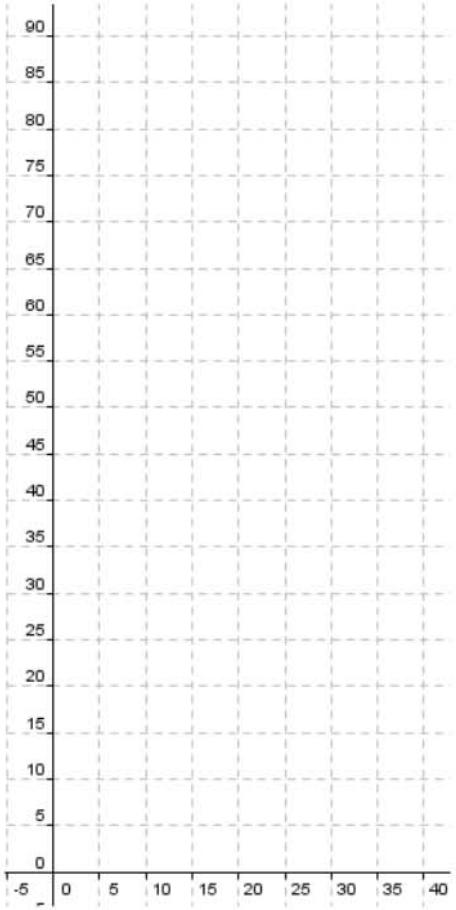
Presenting your workings out

The moisture content, m, of core samples of mud is measured as a percentage. It is believed that m is related to the depth, d metres, at which the core is collected. The results for eight samples are given in the table.

d	0	5	10	15	20	25	30	35
m	90	82	56	42	30	21	21	18

a) Calculate the value of r.

b) Draw a scatter diagram, and comment, in the context of the data, on your value of r.



Using my calculator, I worked out

$\sum x =$ [redacted] $\quad \sum y =$ [redacted] $\quad \sum xy =$ [redacted]

$\sum x^2 =$ [redacted] $\quad \sum y^2 =$ [redacted]

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = \text{[redacted]} - \frac{\text{[redacted]}}{\text{[redacted]}} = \text{[redacted]}$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = \text{[redacted]} - \frac{\text{[redacted]}}{\text{[redacted]}} = \text{[redacted]}$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = \text{[redacted]} - \frac{\text{[redacted]}}{\text{[redacted]}} = \text{[redacted]}$$

And Product moment correlation coefficient is $r = \dfrac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \dfrac{\text{[redacted]}}{\sqrt{\text{[redacted]} \times \text{[redacted]}}} = -0.95$

Comment:

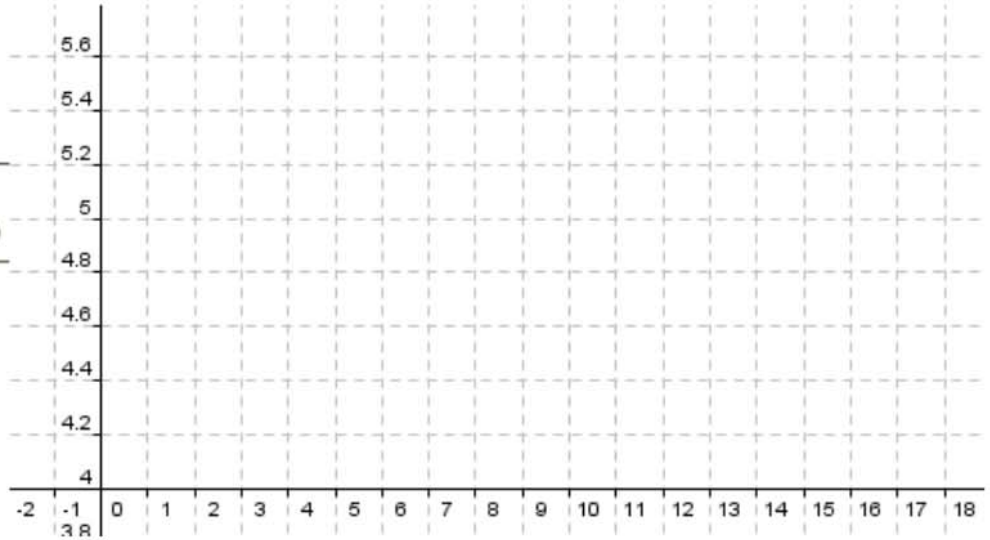
Strong evidence that the moisture content decreases with depth

The data in the following table relates the average temperature (in degrees Celsius) and the average butterfat content for a group of cows (expressed as a percentage of the milk).

Temperature (°C)	17	16	13	4	8	14	16	3	3	16
Butterfat (%)	4.65	4.83	4.55	5.44	4.69	4.65	4.65	4.95	4.66	4.60

a) Draw a scatter diagram.

b) Calculate the value of the product moment correlation coefficient, and comment on its value.



Hide points

Show points

b) Using my calculator, I worked out

$\sum x =$ $\sum y =$ $\sum xy =$

$\sum x^2 =$ $\sum y^2 =$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} =$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} =$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} =$$

And Product moment correlation coefficient is $r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} =$

Comment:

Exercises:

3 A survey of common garden birds in Great Britain gave figures for each species, comparing the total number recorded with the percentage of gardens where each species was seen. The figures for nine species are given in the following table.

Number recorded (thousands)	702	673	308	156	455	411	370	230	196
Percentage of gardens	53.9	64.0	47.4	50.3	81.8	89.0	59.2	57.4	80.0

a) Draw a scatter diagram for the data.

b) Calculate the value of r.

c) Comment, in the context of the question, on your value of the correlation coefficient.

4 The following table gives the numbers of households living in their own properties, and the numbers that are not doing so (both in thousands), for selected regions of East Anglia.

Region	Southend	Colchester	Kings Lynn	St Albans	Ipswich
Owner-occupied	51.6	46.1	41.8	40.6	32.4
Other	19.4	17.6	16.5	12.1	17.5
Region	**Cambridge**	**Hertsmere**	**Harlow**	**Brentwood**	**Maldon**
Owner-occupied	22.8	28.5	19.8	22.8	19.5
Other	19.8	9.4	13.4	6.0	4.7

Source: *Office of National Statistics*

a) Display these data on a scatter diagram.

b) Determine the value of r, the correlation coefficient.

c) Interpret your value of r in the context of the data.

d) Suppose that the correlation coefficient had been equal to 1. What would this imply about the households?

5 The following table contrasts the rates of two types of crime (violence against the person and robbery) for various London districts. The rates, per 1000 population, apply to the year starting April 2000.

District	Westminster	Islington	Hounslow	Brent	Croydon
Violence	36.6	30.9	25.5	22.9	18.7
Robbery	10.3	8.6	3.4	7.7	4.6
District	**Kingston**	**Bexley**	**Havering**	**Richmond**	
Violence	16.6	14.6	12.1	10.4	
Robbery	1.6	1.9	1.7	1.2	

Source: *Office of National Statistics*

a) Display these data on a scatter diagram.

b) Calculate the value of r.

c) Interpret your value of r in the context of the data.

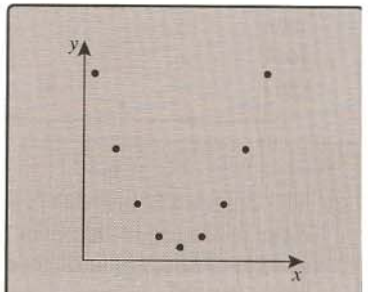
3) b) $r = 0.049$ *Little relation between...*
4) b) $r = 0.570$ *Moderate relation between...*
5) b) $r = 0.906$ *Strong evidence that...*

Limitations of correlation

• Non-linear relations

Although the value of the correlation coefficient is a measure of the extent to which the variables X and Y are linearly associated, its interpretation often needs care.

- A near-zero value of r should not be interpreted (without reference to the data) as implying that the variables are unrelated.
- A value of r near 1 or -1 should not be interpreted (without reference to the data) as implying that the variables are exactly linearly related, though clearly a straight-line relationship will be a good approximation

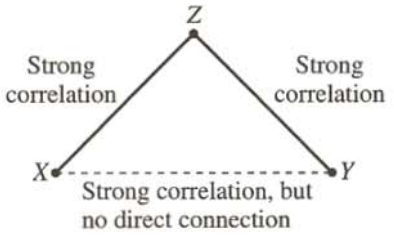


This scatter diagram (also shown on page 147) indicates a simple non-linear relation connecting X and Y, and yet the correlation is zero.

• Influence of a background variable

It is often the case that the two variables of interest, X and Y, are each strongly correlated with an unmeasured background variable, Z. The result is that any changes in Z affect each of X and Y, which are therefore strongly correlated with each other, even though there may be no direct connection.

Often this is incorrectly interpreted as meaning that X 'causes' Y (or vice-versa) when in fact it is Z that 'causes' both.



Example: The table records (in years) the life expectancies at birth of males (x) and females (y), based on data supplied by the Office for National Statistics.

	1972–76	1977–81	1982–86	1987–91	1992–96	1997–99
All males	69.2	70.0	71.4	72.3	73.9	75.0
All females	75.1	76.3	77.1	77.9	79.3	79.7

Source: *Longitudinal Study, Office for National Statistics*

The scatter diagram suggests that the near-linear relationship is perfectly genuine. Even so, it is not reasonable to claim that either variable influences the other. What has happened is that improved medical facilities and living conditions have resulted in longer lives for all people, so that both males and females show a steady rise in life expectancy. The background variable in this case is the unmeasured state of human welfare.

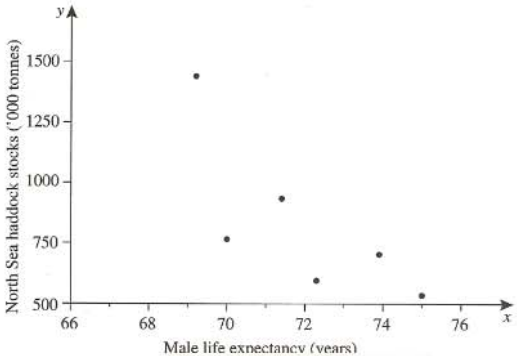
• Spurious Correlation - Nonsense Correlation

Sometimes the variables X and Y make a very curious or comical pair, as the next example illustrates. In this case, the correlation between X and Y may be referred to as being a **spurious correlation** (also called a **nonsense correlation**).

Example: The table records the life expectancies at birth of males (x years) and the North Sea stocks of haddock (y thousands of tonnes averaged over the time periods given).

Time period	1972–76	1977–81	1982–86	1987–91	1992–96	1997–99
Male life expectancy (x)	69.2	70.0	71.4	72.3	73.9	75.0
Haddock stocks (y)	1439	765	933	595	704	535

Sources: *Longitudinal Study, Office for National Statistics; Centre for Environment, Fisheries and Aquaculture Science*



$r = -0.770$ (to 3 dp).

The rather strong negative correlation between the variables might be interpreted as suggesting either that the haddock stocks would be increased by decreasing male life expectancy, or that the elimination of haddock from the North Sea would be accompanied by a further increase in male life expectancy.

In fact, as is often the case, the third variable is time. Over time, there have been improvements in medicine leading to increased life expectancy. Meanwhile, consistent over-fishing of the North Sea has resulted in reduced fish stocks. The activities of the health service and the fishing industry are quite unrelated, but both have changed with time. These simultaneous but unconnected changes have led to the high spurious correlation obtained here.

Identification of response and explanatory variables

In the measurement of correlation, the two variables are treated in the same way, which is reflected by the formula for r, which would be unaltered if x and y were interchanged.

Often, however, it is interesting to find out about the way that the values of one variable (usually Y) change as the values of the other variable change. This is emphasized by referring to Y and X as the **response variable** and the **explanatory variable**, respectively.

Here are some examples.

Explanatory variable	Response variable
Time for which a chemical reaction is allowed to proceed	Weight of chemical compound produced
Weight of chemical compound required	Time taken to produce this weight
An interval of time	Number of cars passing during this interval
Number of cars passing a junction	Time taken for these cars to pass

Former terminology referred to the response variable as the **dependent variable**, and the explanatory variable as the **independent variable**.

To decide which variable is which requires some knowledge of how and why the data were collected. Often, the explanatory variable takes some specified value, while the value taken, as a consequence, by the response variable is subject to random variation. The response variable is, therefore, denoted by Y, rather than y, since it is a random variable. Conversely, x rather than X is used for the explanatory variable to emphasize that it is *not* a random variable. Here are some examples.

x	Y
Number of bricks in pile	Weight of pile of bricks
Price of a commodity	Number sold
Capacity of a car engine	Average miles/gallon

The first example is a particularly obvious one. A lot of piles could contain precisely 10 bricks. However, each of those piles would have a different weight (because of the variations in the weights of individual bricks). Thus, while x is a fixed quantity (for example, 10) for each pile, y is an observation on a random variable Y.

Key point summary

1 A scatter diagram should be drawn to judge whether correlation is present.

2 The product moment correlation coefficient,

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left\{\sum x^2 - \frac{(\sum x)^2}{n}\right\}\left\{\sum y^2 - \frac{(\sum y)^2}{n}\right\}}} \quad \text{or} \quad \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Remember, this can be found directly using a calculator.

r is a measure of **linear** relationship only and $\mathbf{-1 \leq r \leq +1}$

Do not refer to r if a scatter diagram clearly shows a non-linear connection.

3 $r = +1$ or $r = -1$ implies that the points all **exactly** lie on a **straight line**.

$r = 0$ implies **no** linear relationship is present.

But ... no linear relationship between the variables does not necessarily mean that $r = 0$.

4 Even if r is close to $+1$ or -1, **no causal link** should be assumed between the variables without thinking very carefully about the nature of the data involved.

Remember the feet stretching! Will it really help you to get better at maths?