

CALCULATING THE CORRELATION COEFFICIENT

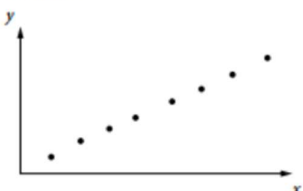
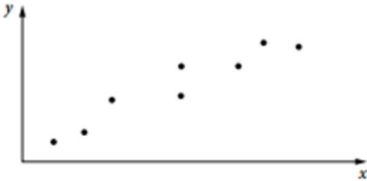
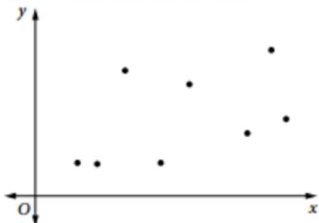
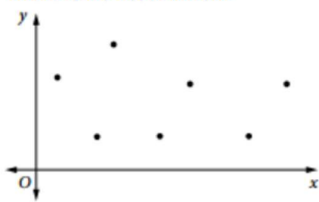
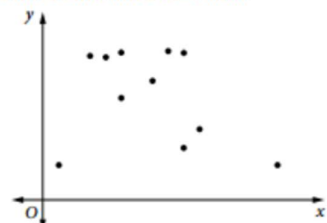
As mathematicians, the idea of using rather vague, undefined terms like 'how well the data fits a straight line' and having broad categories like weak, moderate and strong is quite unsettling.

Mathematicians, including young developing ones such as yourself, always prefer to have a number that can be used to describe the strength of the association.

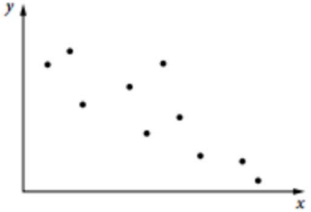
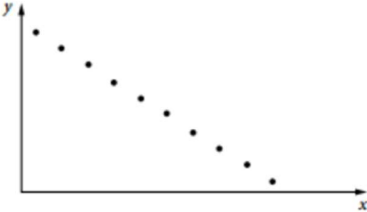
Correlation is the numerical measure to express how closely the individual coordinate pairs fit the line of best fit.

The correlation coefficient r

The correlation coefficient r is a number between -1 and 1 . The value of r is known as the product moment correlation coefficient; it is also sometimes called the Pearson correlation coefficient, after its developer Karl Pearson.

Correlation coefficient	Description
$0.75 \leq r \leq 1$	This correlates to a strong, positive, linear association with all points on the line of best fit. 
$0.5 \leq r \leq 0.75$	This correlates to a moderate, positive, linear association with the points spaced on either side of the line of best fit. 
$0.25 \leq r \leq 0.5$	This correlates to a weak, positive, linear association with the points spaced on either side of the line of best fit. 
$-0.25 \leq r \leq 0.25$	There is no association. 
$-0.5 \leq r \leq -0.25$	This correlates to a weak, negative, linear association with the points spaced on either side of the line of best fit. 

CALCULATING THE CORRELATION COEFFICIENT

$-0.75 \leq r \leq -0.5$	<p>This correlates to a moderate, negative, linear association with the points spaced on either side of the line of best fit.</p> 
$-1 \leq r \leq -0.75$	<p>This correlates to a strong, negative, linear association with all points on the line of best fit.</p> 

The formula for r is quite complex and often will require the use of digital technology. Two forms of the formula are given below. Here \bar{x} is the mean and s_x is the sample standard deviation of the independent variable x , while \bar{y} is the mean and s_y is the sample standard deviation of the dependent variable y . Also n is the number of points used.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{ns_x s_y} \quad \text{or} \quad r = \frac{\frac{1}{n} \sum xy - \bar{x}\bar{y}}{s_x s_y}$$

You can set up a spreadsheet to calculate the value of the correlation coefficient r . On scientific calculators, the correlation coefficient may be calculated after entering the data pairs into the statistics mode.

Outliers have a large impact on the calculation of r and must be removed before any calculation is done.

Example 9

Find the value of r , correct to 2 decimal places, for the following bivariate data set.

Study time (hours)	1	6	4	12	3	2	9
Number of Facebook friends	900	40	100	20	650	1000	40000

Solution

Identify any outliers and remove them from the data set: The point (9, 40 000) does not fit the other six points so it should be removed.

Identify the independent variable, if there is one: In this case consider the following two options.

- Having more friends could mean the person becomes more distracted and less study will be done.
- Alternatively, increasing your study time is unlikely to increase the number of Facebook friends you have.

Independent variable is the number of Facebook friends.

Use digital technology to calculate the value of r : $r = -0.77$

CALCULATING THE CORRELATION COEFFICIENT

Association and causation—not the same thing

Consider the following associations of pairs of variables.

Variable 1	Variable 2	r	r^2
Hours of structured practice	Piano-playing ability	0.8	0.64
Hours of training	Tiredness after a 5 km run	0.75	0.56
Hours of productive study	Exam result	0.9	0.81
Maths exam result	English exam result	0.95	0.90
Number of newsagencies	Number of hospitals	0.9	0.81

Each of these associations has a strong positive correlation, but can you say that the change in one variable can be caused by a change in the other?

Simply knowing that two variables are associated, no matter how strongly, is not sufficient evidence to conclude that the two variables are causally related.

A causal relationship is demonstrated when a change in the independent variable causes a change in the dependent variable.

Other reasons for an association may be:

- Both variables may be responding to a third variable. This is known as a common response as both variables share a common response to a third variable.
- There may be causation, but the change may also be caused by one or more uncontrolled variables whose effects cannot be disentangled from the effect of the independent variable. This is known as **confounding**.
- There also remains the possibility that what is considered to be the dependent variable is actually the independent variable and is causing the change.
- It may also simply be a coincidence.

Statisticians calculate a **coefficient of determination**, r^2 , to give the level of association, but you should not assume that this implies the level of causation.

For the first three associations in the table above, there seems to be a clear cause-and-effect situation. A change in variable 1 (the independent variable) causes a responsive change in variable 2, the dependent variable.

In fact, the value of r^2 can be used to be even more precise about this:

- 64% of the change in piano-playing ability is due to a change in the hours of structured practice. 36% of the change is due to other factors.
- 56% of the change in tiredness after a 5 km run is due to the change in the hours of training. 44% of the change is due to other factors.
- 81% of the change in exam results is due to the change in the hours of productive study. 19% of the change is due to other factors.

The strong correlation in Maths and English results suggest that it is most likely that both of these results are due to other uncontrolled variables such as intelligence, completion of work and hours of study. This would be an example of confounding causation.

The association between the number of hospitals and newsagencies is a good example of a common response. Both of the variables are responding to the population of the town.

CALCULATING THE CORRELATION COEFFICIENT

Example 10

Determine if there is a causal relationship between each pair of variables:

- (a) hours of coaching versus tennis ability: $r = 0.8$ and $r^2 = 0.64$.
- (b) number of hours of sunlight versus temperature: $r = 0.5$ and $r^2 = 0.25$

Where causation is observed, complete the following statement.

___% of the change in variable 2 is due to the change in variable 1. ___% of the change is due to other factors.

Where causation cannot be verified discuss the reason.

Solution

- (a) Decide whether one variable directly influences the other: More hours of coaching would increase a person's tennis ability so a causal relationship is established.
64% of the change in tennis ability is due to the change in hours of coaching. 36% of the change is due to other factors.
- (b) Decide whether one variable directly influences the other: These factors are more likely to respond to a common response in this case such as season or location.
Hours of sunlight and temperature both respond to the season and location so they both respond to a common response.

Note: If, for example, you studied the link between the number of phone calls made and the amount of donations received, and found r^2 to be 0.36, the causation is still there. The amount of donations received has many other factors like the persuasiveness of the caller, the financial situation at the time, the areas the calls are made to and if there is a current high profile emergency situation.

Causation is not an exact science and whilst there are fairly obvious examples of each type, in the real world the situation is often somewhere in between and open to some interpretation or further research. In the following exercise only more obvious examples have been chosen.