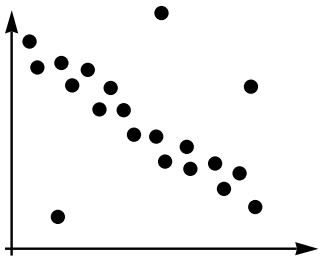
Bucking the trend



UNIT 13

This unit deals with scatter diagrams and correlation. It explores the relationship between correlation and cause, and whether behaviour is necessarily determined when there is a known correlation.

Using this unit

The unit is designed for students at Intermediate and Higher Tiers of GCSE and will last for about 3 hours. The unit could be either used to introduce scatter diagrams and correlation or as revision. The students need basic graphical skills and the ability to use algebraic formulae.

Students begin by looking at scatter diagrams and lines of best fit. Then the concept of correlation is introduced. The relation between correlation and causality is considered next. The unit introduces Spearman's correlation coefficient, as an example of a numerical measure of correlation. The unit ends with a discussion of how far behaviour can be explained, or excused, by correlation - should we "follow a trend"?

There is a small extension section on non-linear correlation.

Access to spreadsheets such as Excel would help students in this work.

Mathematical content

Handling Data (AT4)

- **Scatter diagrams and lines of best fit**
- **The concept of correlation**
- **Spearman's rank order correlation coefficient**

Spiritual and moral development

The aim of this unit is to encourage students to think about whether behaviour can ever be justified just because it is in line with a trend and to consider to what extent a trend may remove their personal responsibility. It is also hoped that they will reflect on whether rather intangible characteristics such as happiness can be measured.

TEACHER'S NOTES

TEACHER'S NOTES

Notes on the activities

It happens all the time

This is an introductory activity and could be addressed through a whole class discussion. At this stage, there is no need to discuss the validity of the statements, but instead just ensure meaning is clear. Discussion could focus on the seemingly less quantifiable categories such as happiness and success and how they might be measured.

Scatter diagrams and Line of best fit

These two sections introduce the ideas of scatter diagrams and lines of best fit. There are both explanations and practice examples. Students could work individually.

More data may be required for further practice. One way of getting a good quantity of data that offers possible correlation is to ask each member of the class to bring in a sheet with a number of their own details on it. These could include height, hand-span, shoe size, house number etc. This provides a source for a great number of scatter diagrams, and, eventually, correlation coefficients. It should be asked for as part of a homework assignment, in preparation for the lesson itself.

Correlation or cause?

This section addresses the interpretation of situations where correlation occurs. The connection between correlation and cause is a complex idea and students may need support.

Class discussion

It is worth bringing the class together to discuss their answers, including:

- which statements they think are true and why;
- for the true statements, whether one factor causes another or there is third factor.

Correlation coefficients

Although correlation coefficients do not appear in most GCSE syllabuses, it is worth introducing students to the idea that correlation can be quantified. Indeed, students may already have met correlation coefficients in Geography or Science (particularly Biology) and so it is worth consulting with these departments about the topic.

Spearman's rank order is chosen partly due to its accessibility. It should be noted that no explanation is given regarding tied ranks.

Will you buck the trend?

It is preferable that the whole class should work on this section at the same time. If some students have finished earlier work, they could try the extension activity - *Don't overdo it!* - before starting this section. The final core section encourages students to question what happens when there are genuine trends. In particular, is it necessary to follow the trend? Some of the examples within the section should be used sensitively, as they may well apply to members of the class, and there is a danger of a reinforcement of a negative self-image. The point is that it is possible to avoid the trend.

Class discussion

The students' responses to this final section would inform a productive class discussion and issues addressed could be:

- whether they agree with the new examples given in question 2;
- whether they know of real exceptions to genuine trends;
- whether they think there is pressure to follow trends in society.

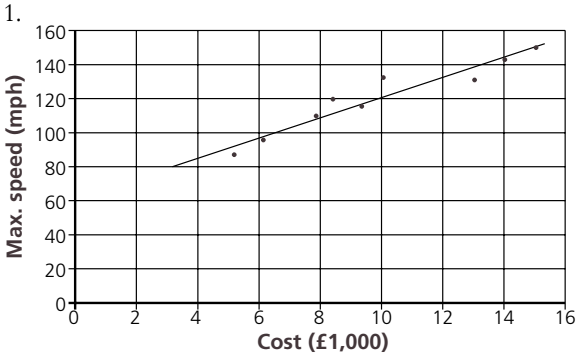
Don't overdo it! (extension activity)

This final section is an extension task which contains examples of trends which are not linear.

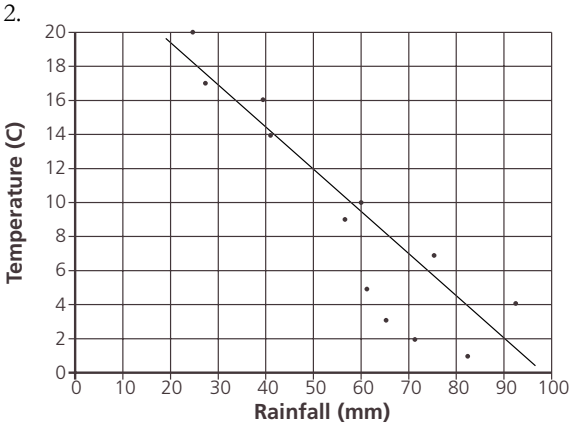
Answers

Task 1 and Task 2 (question 1):

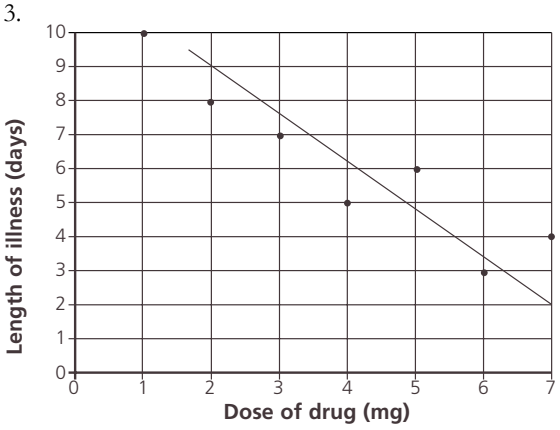
1.



2.



3.



Task 2:

2. a) 130 mph.
   b) 58 mm.

3. a) possibly about 2 days (uncertain due to extrapolation).
   b) extrapolation gives meaningless answer.

Task 3:

Discussion.

Task 4:

1. r = 1, positive correlation.
2. r = -1, negative correlation.
3. r = 19/21 = 0.905, positive correlation.
4. r = 29/30 = 0.997, positive correlation.
5. r = -128/143 = -0.895, negative correlation.
6. r = -13/14 = -0.929, negative correlation.

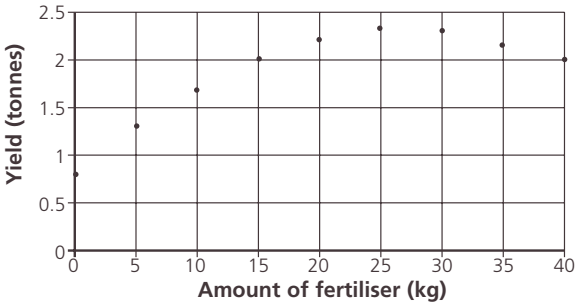
Task 5:

Class discussion. There are a number of successful dyslexics e.g. Susan Hampshire (actress), Michael Heseltine (politician). Duncan Goodhew (swimmer) and Ian Botham (cricketer) both had asthma.

Task 6:

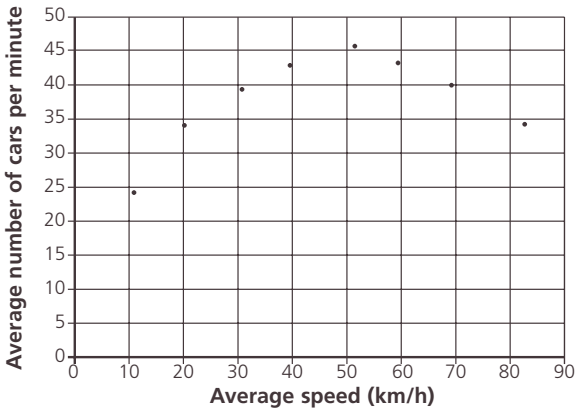
All three examples show non-linear correlation, where a straight line would be inappropriate, but where there is still evidence of a trend. In all three cases, too much of one quantity produces a decrease in the second.

1. a)



   b) No.
   c) Non-linear.
   d) Too much fertiliser can damage the crop.

2. a)

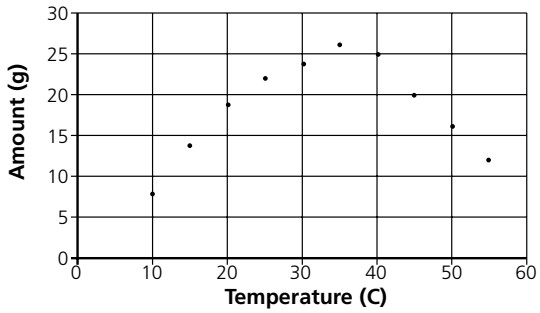


TEACHER'S NOTES

b) Low speeds probably occur during congestion when traffic moves very slowly and high speeds when there is little traffic and it is well spread out.

c) Non-linear.

3. a)

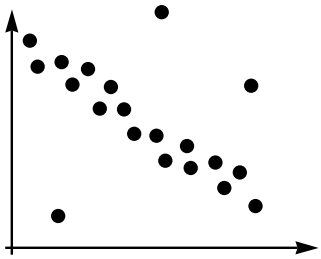


b) When temperature gets too hot growth is damaged, when it is too cold growth cannot start.

c) Non-linear.

TEACHER'S NOTES

Bucking the trend



UNIT 13

It happens all the time!

Here are some statements which may or may not be true. In each one it is claimed that one aspect of a person's (or team's) character or behaviour suggests a second one will also be the case. What do you think?

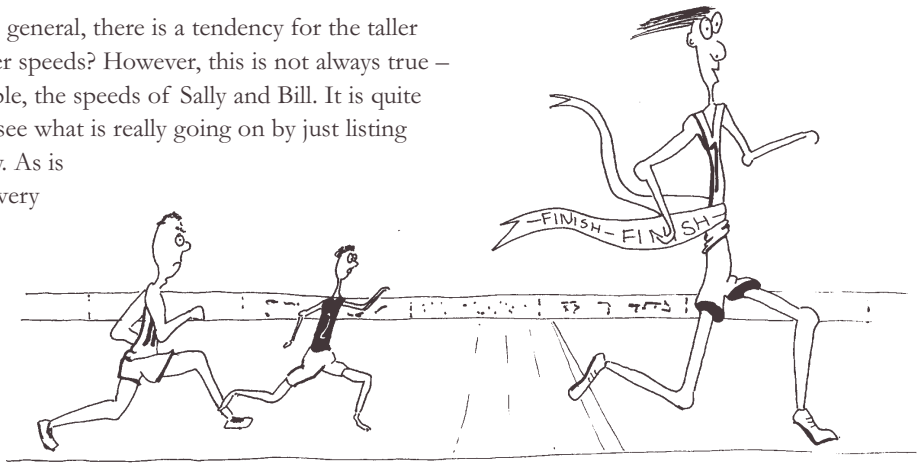
- The taller you are, the faster you run.
- The better you are at Maths, the worse you are at French.
- The more money you earn, the happier you are.
- The more goals a team scores, the higher up the League it will be.
- The more you smoke, the shorter your life will be.
- The cleverer you are, the more money you will earn.
- The better a boy is at sport, the more girl-friends he will have.
- The more you pray, the more successful you become.
- The more a car costs, the higher its top speed is.
- The harder you work, the better GCSE results you get.
- The more Vitamin C you take, the fewer colds you get.
- The further away you live from school, the longer it takes to get there.

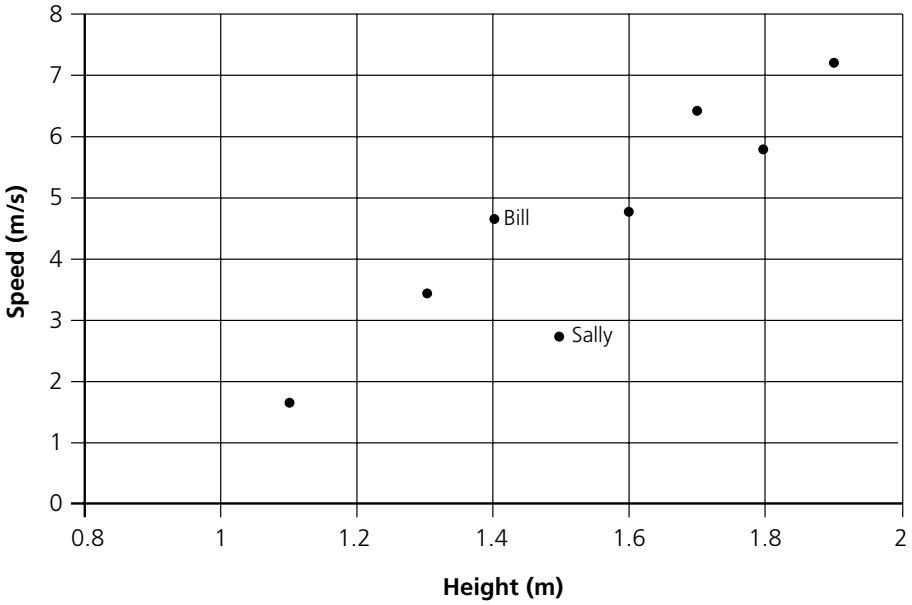
Scatter diagrams

We are now going to look at the first two statements in more detail. To do this, we would actually need some data. For example, for the first statement above, we would need data on heights and running speeds for a number of people. Some possible results are shown in the table.

Person	John	Fred	Ann	Chris	Sally	Bill	Lisa	Sarah
Height (m)	1.6	1.9	1.7	1.8	1.5	1.4	1.3	1.1
Speed (m/s)	4.8	7.2	6.4	5.8	2.8	4.6	3.4	1.6

Can you see that, in general, there is a tendency for the taller people to have faster speeds? However, this is not always true – compare, for example, the speeds of Sally and Bill. It is quite difficult, in fact, to see what is really going on by just listing the data numerically. As is often the case, it is very helpful to draw a graph. We need to use a scatter diagram.



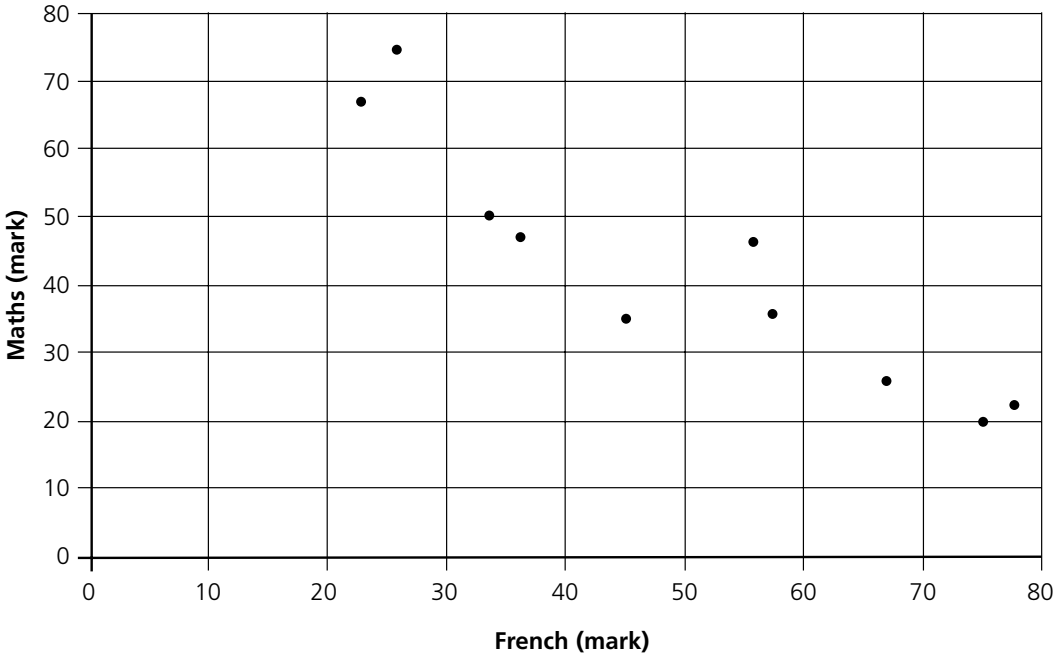


In this graph, each person is shown as one point, and we plot the two measurements for each person, their height and their speed, using the two axes. The graph shows much more clearly that for these people there is a general trend, the taller the people are the faster they run. We call this trend **correlation**. The graph also shows more clearly the people who do not follow the trend.

Next let us look at a table of results and a scatter diagram linked to the second statement. The results of some pupils in two tests, French and Maths, are shown.

Pupil	A	B	C	D	E	F	G	H	I	J
French	23	45	67	34	78	26	75	56	36	58
Maths	67	35	26	50	22	75	20	46	47	36

Again, there is a fairly definite trend, but it is different from the first example. In this case, as one feature gets bigger, the other one gets smaller. This is called negative correlation, in contrast to the previous case, which was positive correlation.



1 Plot scatter diagrams for the following sets of data. In each case write down whether there is any correlation between the data; if there is, say whether it is positive or negative.

1.

Type of car	A	B	C	D	E	F	G	H	I
Cost of car (£1000)	5.1	6.2	7.8	8.4	9.3	10.1	12.9	14.0	14.9
Top speed (m.p.h.)	86	95	110	120	115	132	131	144	150

2.

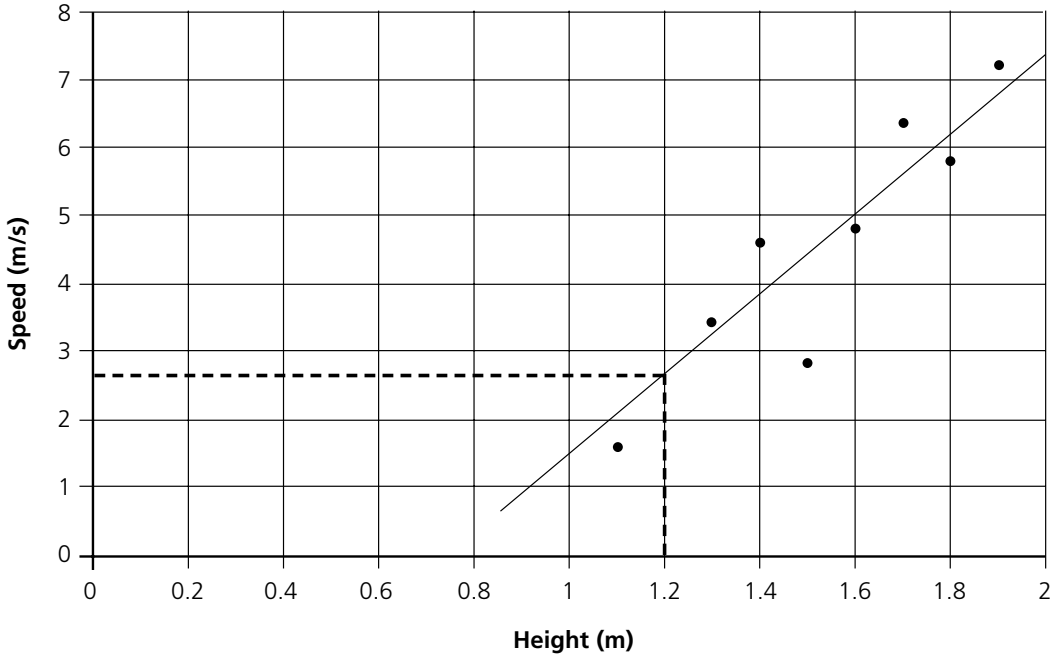
Month	J	F	M	A	M	J	J	A	S	O	N	D
Rainfall (mm)	71	82	93	75	60	41	27	24	39	56	61	65
Monthly temp. (°C.)	2	1	4	7	10	14	17	20	16	9	5	3

3.

Patient	A	B	C	D	E	F	G
Dosage of drug (mg)	1	2	3	4	5	6	7
Length of illness (days)	10	8	7	5	6	3	4

Line of best fit

Where there is a clear trend, it can be useful to illustrate this by drawing a straight line through the data. Although there are some fairly complicated rules about finding precisely the right line, it is often good enough just to use your common sense in drawing it. One useful guide is to aim to have as many of the points below the line as above it. Here is how a best straight line for the first set of data might look.



Lines of best fit can be used to estimate new values. In the athletics example, we might wonder how fast someone who is 1.2 m tall might run. By reading up to the line of best fit and across, we can see that 1.2 m corresponds to 2.7 m/s, and so that would be our estimate of the person's speed.

2

1. Go back to the scatter diagrams you did in Task 1. Draw a line of best fit on each one. Make sure that it gives a good sense of the way the data is going.

2. Use the appropriate lines of best fit from question 1 to answer the following questions.
   a) A new car costs £10,900. Estimate its top speed.
   b) A meteorologist loses the rainfall records for March. Knowing that the temp is 10ºC, what should she estimate the rainfall as?

3. It is usually safe to use the line of best fit to fill in gaps between points (interpolation), but it can be unwise to use it to find values beyond the original data (extrapolation). Use the line of best fit from the medical case to estimate:
   a) How long would the illness be if 9mg of drug was given?
   b) How long would the illness be if 20 mg of the drug was given? Is this sensible?!

Correlation or cause?

Very often, two sets of data are correlated because one causes the other. Tall people are generally faster runners because their height helps them. People good at Maths are generally good at Physics because their mathematical skill helps in their Physics.

However, there are other situations where the characteristics are not directly related, but which both depend on a third item. For example, in the last ten years, there has been a positive correlation between the number of dishwashers bought each year and the number of video cameras bought each year. No-one would suggest, however, that buying a dishwasher creates an urge in a person to buy a video camera! Instead, both things have happened because better technology has made the items more available and cheaper, and increased earnings have enabled more people to buy them.

3

Look again at the statements you read at the beginning of the unit.

1. Which of these do you think are true?

2. For those which you think are true, which do you think occur because
   a) one factor causes the other (try to explain why) or
   b) because they are both caused by a common third factor (try to say what)?

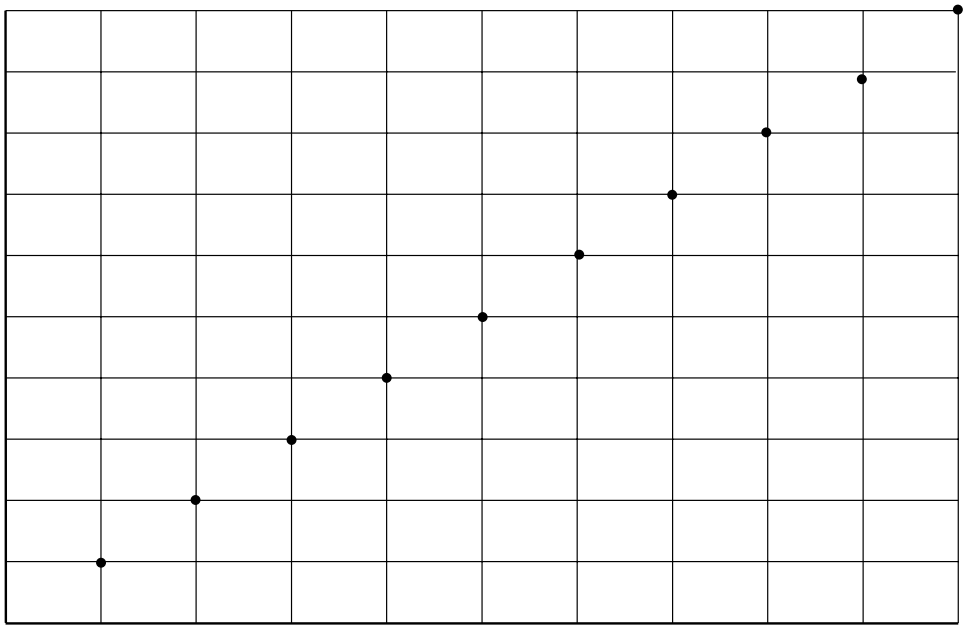
So correlation does not necessarily mean that one factor is causing another. We cannot put responsibility - credit or blame - onto one item, just because it is correlated with something else. For example, the trend in the second last statement might not prove that Vitamin C actually prevents colds. It might just be that people who bother to take Vitamin C are very health-conscious people anyway, who look after themselves in all sorts of ways, and so stay healthy. Use of Vitamin C could just be a reflection of their behaviour, not a cause of their health.

Correlation coefficients

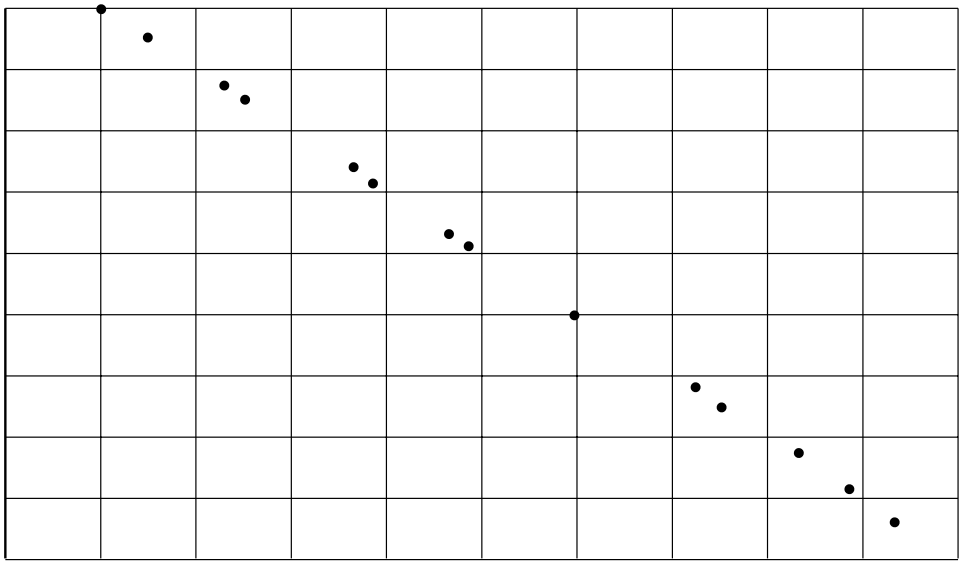
When a scatter diagram contains a large number of points, it can be difficult to judge whether or not there really is any correlation. One person might think that there is evidence of a trend, another might disagree. To help sort out these problems, statisticians have developed a way to produce a number as a measure of the correlation.

This number is called the correlation coefficient and ranges between -1 and +1. For perfect positive correlation the coefficient is +1 and for perfect negative correlation it is -1. Most situations, of course, give something in between -1 and +1. If, for example, we get a coefficient of 0.8, then that suggests that there is a quite strong positive correlation, even though it is not perfect.

Correlation of +1



Correlation of -1



There are, in fact, several alternative formulae for measuring correlation. We will look at one called Spearman's rank order correlation coefficient. The words *rank order* indicate that we do not use the actual data, but rather the positions, or ranks, of the numbers within the set of data.

Example

We will work on the Maths and French marks. We calculate the coefficient as follows:

Stage 1. Calculate the rank in French and Maths for each pupil (for example, pupil A was 10th in French, and 2nd in Maths).

Stage 2. Work out the difference between the two rankings for each pupil

Stage 3. Square the difference for each pupil

Stage 4. Use the following formula for calculating the coefficient

$$r = 1 - \frac{6 \times \sum d^2}{n(n^2 - 1)}$$

where r is the coefficient, $\sum d^2$ is the total of the differences squared and n is the number of pupils.

Pupil	A	B	C	D	E	F	G	H	I	J	
Rank in French	10	6	3	8	1	9	2	5	7	4	
Rank in Maths	2	7	8	3	9	1	10	5	4	6	
Difference in rank (d)	8	1	5	5	8	8	8	0	3	2	
Difference squared (d^2)	64	1	25	25	64	64	64	0	9	4	Total 320

$n = 10$, so $r = 1 - \frac{6 \times 320}{10 \times (100 - 1)} = \frac{64}{33} = -0.94$

As this is near -1, we can say that there is strong negative correlation between the French and Maths marks.

For each of the following situations, work out the value of Spearman's rank order correlation coefficient and describe in words the correlation.

1.

Pupil	A	B	C	D	E	F	G	H
Rank in Physics	1	2	3	4	5	6	7	8
Rank in Maths	1	2	3	4	5	6	7	8

2.

Pupil	A	B	C	D	E	F	G	H
Rank in English	1	2	3	4	5	6	7	8
Rank in Maths	8	7	6	5	4	3	2	1

3. The runners' height and speed data (First example).

4. The cars' cost and top speed data (Task 1, question 1).

5. The months' rainfall and average temperature data (Task 1, question 2).

6. The patients' drug dosage and length of illness data (Task 1, question 3).

Will you buck the trend?

We saw earlier that sometimes correlation is due to a cause-and-effect process, and at other times there is not such a clear link. However, even when correlation is due to a genuine causal effect, we must remember that correlation is normally never perfect, and that some of the data may not follow the trend. For example, it is sometimes said that there is a positive correlation between unemployment and crime – in areas with higher unemployment, there tends to be more crime.

Do you think that this is true?
If so, why does it happen?

One possible cause is that unemployment leads to people having lots of spare time, with not much to do, and with little spending money, so crime provides both a form of activity and material gain. However, this correlation is not perfect - it does not have a correlation coefficient of +1, and this means that there are many people who do not fit the pattern, just as some tall people are not particularly fast runners.

Even when there is correlation, we still have the moral choice of whether we are going to fit the pattern - go along with the crowd - or whether we are going to do our own thing, based on our thoughts and ideas.

In example 6, a very clever person might choose to work as an Oxfam worker or missionary, in a Third World country, even though the correlation suggests says they could get a high-paid job in more comfortable surroundings.

Should they? What would you do?

5

1. Look back again at the original examples at the start of the Unit. In the cases where you think there is a trend, ask the question:

   "do you have to follow the trend?"

2. Here are some more examples, which may be true or false:
   - The more you are affected by dyslexia, the worse you will do at school.
   - The more you suffer from asthma, the worse you will do at sport.
   - If you come from a one-parent family, you won't succeed in life.

   a) In each case say if you think there is a real trend or not.
   b) If you think there is a trend, can you also think of exceptions to the trend.
   c) Say whether you would want to buck the trend.

Don't overdo it!

6

1. The following data was collected from an agricultural research station, after an experiment to find out how much crop, per hectare, could be produced by adding fertilizer.

Amount of fertiliser (kg)	0	5	10	15	20	25	30	35	40
Yield (tonnes)	0.8	1.3	1.7	2.0	2.2	2.3	2.3	2.2	2.0

a) Plot a scatter diagram of yield against fertilizer.
b) Does it make sense to draw a line of best fit?
c) Is there correlation?
d) How would you account for the pattern?

2. Some transport research students did a large number of one hour long surveys. For each hour they calculated the average speed of cars passing and the average number of cars passing per minute. Some of the results are shown in the table.

Average speed (km/h)	31	52	83	20	69	39	11	90	59
Average no. of cars per minute	39	46	34	34	40	43	24	20	43

a) Plot a scatter diagram of this data.
b) Try to account for any pattern you see.
c) Is there any correlation?

3. In a biological experiment to produce penicillin, the amount produced in one hour was measured, as was the temperature at which the experiment was conducted. The results were:

Temperature (C)	10	15	20	25	30	35	40	45	50	55
Amount (g)	8	14	19	22	24	26	25	20	16	12

a) Plot a scatter diagram of this data.
b) Try to explain any pattern you see.
c) Is there any correlation?