8.1 Spearman's rank correlation coefficient

Ranking data

For each of the following lists, rank the items in order of how much you like them, with a "1" being the one you like most and a "6" the one you like least.

You must rank them all, and ties are not allowed!

Subject	Rank
Maths	
English	
PE	
Art	
Science	
Drama	

Film genre	Rank
Sci-fi	
Romance	
Documentary	
Comedy	
Adventure	
Thriller	

Sport	Rank
Soccer	
Basketball	
Tennis	
Baseball	
American football	
Table tennis	

Now, find someone to compare your choices with.

In the following table, write down your ranks for the same items. Then, in the other columns, work out the difference between your ranks (d), and the square of the difference (d^2).

Finally, add up the totals in the last column (d^2).

Item	You	Them	d	d^2
			Total:	

Once you have this total, do the following:

- multiply it by 6
- divide the answer by 210
- subtract the answer from 1.

Your answer should be a number between −1 and +1.

Which other mathematical calculation in statistics has an answer between −1 and +1?

International-mindedness

Charles Pearson (1863–1945) was an English psychologist who developed rank correlation, usually shown as the Greek letter ρ (rho) or, r_S, as a tool for psychiatry.

What do you think an answer close to −1 means?

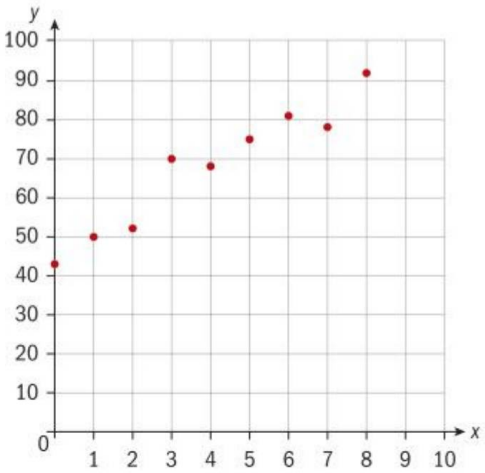
What do you think an answer close to +1 means?

What do you think an answer close to 0 means?

What can you say about your answer?

Consider the following sets of data:

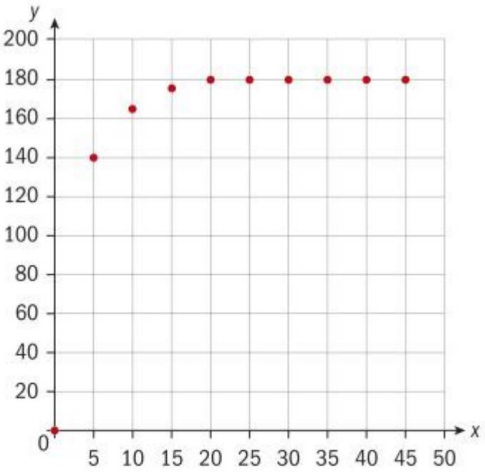
Hours of study, x	0	1	2	3	4	5	6	7	8
Test results, y	43	50	52	70	68	75	81	78	92



A scatter graph of the data is shown here, and Pearson's correlation coefficient is 0.97. So, there is a strong, positive relationship between the hours of study and the test results.

Now consider this data:

Number of minutes, x	0	5	10	15	20	25	30	35	40	45
Temperature of oven, y°C	0	140	165	175	180	180	180	180	180	180

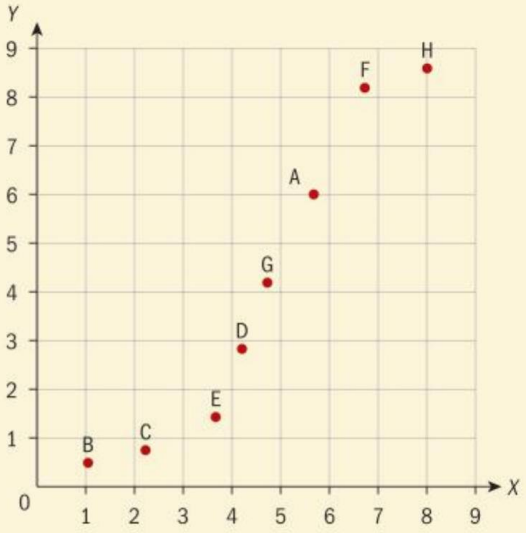


A scatter graph of this data is shown here, and Pearson's correlation coefficient is 0.73. So, there is only a moderate, positive relationship between the number of minutes and the temperature of the oven. However, the scatter graph indicates that there is an exponential relationship between the data points. Pearson's only measures for linear relationships. So, is there another test to find the strength of relationships that are not linear?

Investigation 1

Mould is grown in eight different petri dishes with different amounts of nutrients (X), and the area of the dish covered in mould after 48 hours (Y) is recorded. The results are given in the table and also shown on the graph.

X	Y
5.68	6.00
1.04	0.50
2.22	0.76
4.20	2.84
3.66	1.44
6.72	8.20
4.72	4.20
8.00	8.60



1 Use your GDC to graph these results.

2 Calculate the Pearson's product moment correlation coefficient (PMCC) for this data and comment on your results.

Now give each data point a rank, which is the position of the point if the data were listed in order of size for each of the variables. For example, H would be ranked 1 for both X and Y. (It does not matter if we rank from largest to smallest, like this, or from smallest to largest; the result will be the same.)

3 a Complete the following table showing the ranks for each of the data points.

	A	B	C	D	E	F	G	H
X rank								1
Y rank								1

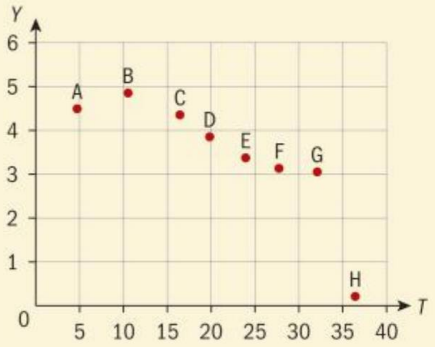
b Use your GDC to graph these ranks.

c Calculate the value of PMCC for these ranks.

d Comment on your result, relating it to the particular shape of the graph.

In another experiment the temperature (T) is varied and the area of the petri dish covered after 48 hours (Y) is recorded.

T	Y
4.95	4.50
10.49	4.86
16.40	4.36
19.80	3.86
23.90	3.38
27.70	3.14
32.30	3.06
36.40	0.22



4 **a** Use your GDC to graph these results.

b Calculate the value of PMCC for this data and comment on your results.

c Complete the following table showing the ranks for each of the data points.

	A	B	C	D	E	F	G	H
T rank								1
Y rank								8

d Use your GDC to graph these results.

e Calculate the value of PMCC for this data and comment on your results.

f Discuss the features of the data that led to this value.

The PMCC of the rank values is called Spearman's rank correlation coefficient.

5 **Factual** What type of data is used for Spearman's?

6 **Factual** What type of data is used for Pearson's?

7 **Conceptual** What do correlation coefficients tell you about the relationship between two variables?

The product moment correlation coefficient of the ranks of a set of data is called Spearman's rank correlation coefficient. The IB notation is r_s.

Spearman's correlation coefficient shows the extent to which one variable increases or decreases as the other variable increases.

An r_s value of 1 means the set of data is strictly increasing, and a value of -1 means it is strictly decreasing. Data that is only increasing or only decreasing is known as **monotonic**.

A value close to 0 suggests that the data is not consistently increasing or decreasing.

Example 1

1 Find Spearman's rank correlation coefficient for the following sets of data.

a

Time spent training, x hours	23	34	17	23	29	45
Time to run 2 km, y min	12	10	14	11	11	8

b

Number of pets, x	1	2	3	4	5
Time spent each week caring for them, y hours	6	7	8	8	16

Continued on next page

2 A student was asked to rank nine different makes of burger in terms of which she liked best to which she liked least. She put 1 for the one she liked best and 9 for the one she liked least. These rankings and the costs of the burgers are given in the table.

Burger	A	B	C	D	E	F	G	H	I
Taste rank	7	3	4	6	1	9	2	5	8
Cost, US $	3.50	7.45	6.50	4.50	8.50	2.65	3.95	4.35	1.45

a Explain why you cannot use Pearson's in this example.

b Find Spearman's rank correlation coefficient for this data and comment on your answer.

1 a The ranks are

x	4.5	2	6	4.5	3	1
y	2	5	1	3.5	3.5	6

$r_s = -0.956$

So, there is a strong, negative correlation. The more hours that you train the faster you can run the 2 km.

If we order the values of x by their size, we get their rank.

When more than one piece of data have the same value the rank given to each is the average of the ranks. For example, the two values of x equalling 23 here would have ranks 4 and 5; hence, each is given a rank of $\frac{4+5}{2} = 4.5$.

The ranked data is put into a GDC and the PMCC obtained.

b The ranks are

x	5	4	3	2	1
y	5	4	2.5	2.5	1

$r_s = 0.975$

So, there is a strong, positive correlation. The more pets you have the more hours it takes each week to look after them.

Often, when one of the variables increases at a fixed rate, for example measurements taken at one minute intervals, the order of the ranks will be the reverse of the order of the data.

2 a Because the ranks are given rather than quantifiable data.

b Ranking the costs, you get:

Taste	7	3	4	6	1	9	2	5	8
Cost	7	2	3	4	1	8	6	5	9

and $r_s = 0.8$. So, there is a moderately strong relationship between the taste and the cost of the burgers.

Spearman's rank correlation coefficient is only valid for the data given in the question. If some data points are similar then any small changes could affect the value of r_s.

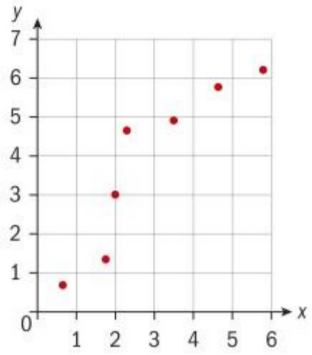
TOK

What practical problems can or does mathematics try to solve?

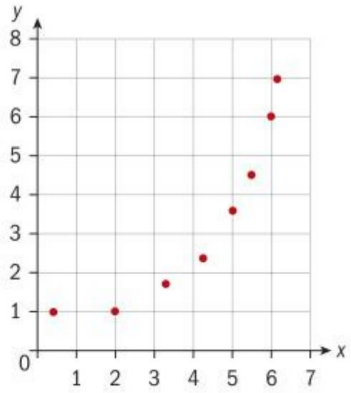
Exercise 8A

1 Write down the value of Spearman's rank correlation coefficient for each of the sets of data shown.

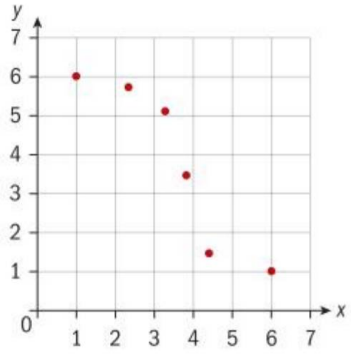
a



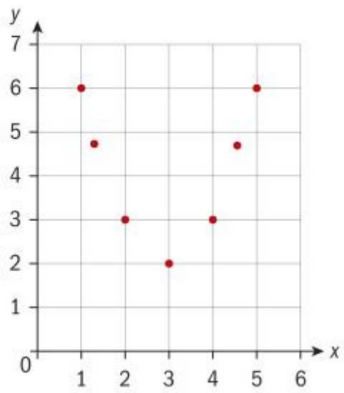
b



c



d



2 A group of students is asked to rank six snack foods by taste and value for money. The ranks are averaged and recorded in the following table.

Calculate Spearman's rank correlation coefficient for the data and comment on your results.

	Pop-corn	Crisps	Choco-late bar	Chews	Chocolate-chip cookie
Taste	2	4	1	5	3
Value	5	3	2	4	1

3 Find Spearman's rank correlation coefficient for the following data sets.

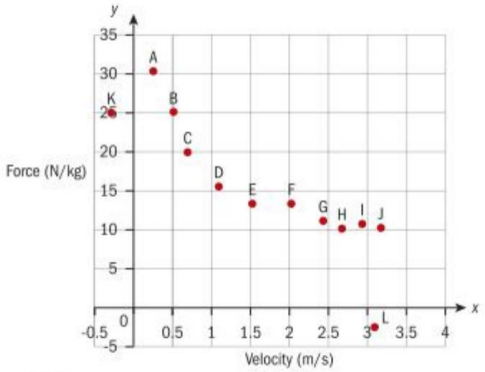
a

x	0	5	10	15	20	25	30
y	23	18	10	9	7	7	7

b

x	10	12	9	6	3	14	8
y	12	11	8	5	7	14	9

4 A sports scientist is testing the relationship between the speed of muscle movement and the force produced. In 10 tests the following data is collected.



Point

- A = (0.25, 30.4)
- B = (0.51, 25.1)
- C = (0.69, 20)
- D = (1.09, 15.6)
- E = (1.52, 13.4)
- F = (2.02, 13.4)
- G = (2.43, 11.2)
- H = (2.67, 10.2)
- I = (2.93, 10.8)
- J = (3.17, 10.3)
- K = (-0.29, 25.07)
- L = (3.09, -2.44)

a Explain why it might not be appropriate to use the PMCC in this case.

b Calculate Spearman's rank correlation coefficient (r_s) for this data.

c Interpret the value of r_s and comment on its validity.

5 A class took a mathematics test (marked out of 80) and an English test (marked out of 100), and the results are given in the following table.

Maths	15	25	37	45	60	72	74	78	78	79	79
English	44	47	42	49	52	44	54	59	69	78	89

a Calculate the PMCC for this data and comment on the result.

b Use graphing software to plot these points on a scatter diagram and comment on your result from **a**.

c Calculate Spearman's rank correlation coefficient for this data and comment on your result.

d State which is the more valid measure of correlation, and give a reason.

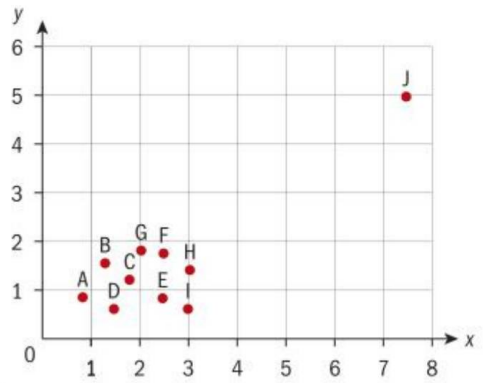
6 In a blind tasting, customers are asked to rank six different brands of coffee in terms of taste. These rankings and the costs of the different brands are given in the following table.

Brand	A	B	C	D	E	F
Taste rank	1	2	3	4	5	6
Cost	450	360	390	320	350	300

a Explain why you cannot use PMCC in this case.

b Find Spearman's rank correlation coefficient for this data and comment on your answer.

7 Consider the following data set:



Point

- A = (0.82, 0.86)
- B = (1.28, 1.56)
- C = (1.78, 1.22)
- D = (1.46, 0.62)
- E = (2.46, 0.84)
- F = (2.48, 1.76)
- G = (2.02, 1.82)
- H = (3.02, 1.42)
- I = (2.98, 0.62)
- J = (7.46, 4.98)

a For this data, calculate the PMCC:

i with the outlier J

ii without the outlier J.

b Calculate Spearman's rank correlation coefficient:

i with the outlier J

ii without the outlier J.

c Comment on the results.

The advantages of Spearman's rank correlation coefficient over the PMCC are:

- It can be used on data that is not linear.
- It can be used on data that has been ranked even if the original data is unknown or cannot be quantified.
- It is not greatly affected by outliers.

Developing inquiry skills

Can you use the PMCC or Spearman's rank correlation coefficient to compare the data in the opening scenario of this chapter, which looked at tree heights in different forest areas?

Why, or why not?

Developing your toolkit

Now do the Modelling and investigation activity on page 418.