Schoolworkout Maths

S1 Revision Notes: Correlation

Section 1: Product-moment correlation coefficient

Recap:
The product-moment correlation coefficient, *r*, measures how close the points on a scatter graph lie to a straight line (or more mathematically it measures the strength of the linear relationship between two variables).

The key points:

- ***r* always lies between -1 and 1**. (If you calculate a value of r that does not lie between -1 and 1 then you've made a mistake!!).
- The formula for calculating *r* is

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

where

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

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

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

Note: These formulae are in the formula book.

- You can use your calculator's *r* button for finding a correlation coefficient.

- **Example:**

In a school there are 10 students on the first year of an A-level Geography course who also study mathematics at A-level. The table shows the marks of each student in the mock examinations in Geography and Mathematics.

Student	Geography mark, x	Maths mark, y
A	79	63
B	72	58
C	57	78
D	78	80
E	73	75
F	56	63
G	83	77
H	81	92
I	13	11
J	62	90

$\sum x = 654,\ \sum x^2 = 46686,\ \sum y = 687,\ \sum y^2 = 52025,\ \sum xy = 48408$

a) Calculate the value of the product-moment correlation coefficient between x and y and interpret the value that you obtain.
b) Plot a scatter graph of these data.
c) Student I was absent from school for most of the year owing to illness. With reference to this and to your scatter diagram, discuss briefly whether your interpretation in a) should be amended.

- **Solution**

a) If you use the formulae:

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 46686 - \frac{(654)^2}{10} = 3914.4$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 52025 - \frac{(687)^2}{10} = 4828.1$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 48408 - \frac{654 \times 687}{10} = 3478.2$$

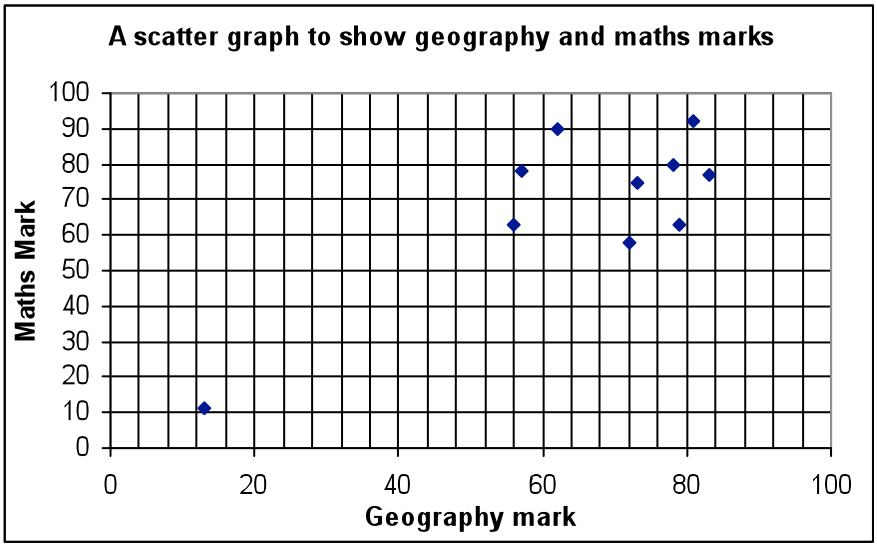
So, $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \dfrac{3478.2}{\sqrt{3914.4 \times 4828.1}} = 0.800$

This indicates a strong positive correlation suggesting that the better a student does in Geography, the better they also tend to do well in Maths.

Note: It is important that you interpret a value of the correlation coefficient in the context of the question.

Note 2: This question could also be completed by typing the data into your calculator.
The steps involved will vary slightly depending on the make of your calculator.

b) A scatter graph for these data is shown below:



Note: Remember to give the graph a title and to label the axes.

c) When you ignore student I, the scatter graph shows very little evidence of any correlation in the remaining 9 points. It is the outlier which forced the correlation coefficient so high.

Note 1: The above example illustrates why the product-moment correlation coefficient is not suitable to data with outliers.
Spearman's rank correlation coefficient (see below) is an alternative correlation coefficient which is more suited to data with outliers or to data which does not appear to follow a linear pattern.

Note 2: You are expected to know that the product-moment correlation coefficient is not affected by *linear* scalings of the variables. The correlation between x and y is the same, for example, as the correlation between u and v where $u = \frac{4x-2}{5}$ and $v = 7y + 2$ (because the transformations of the variables are *linear*).

Section 2: Spearman's rank correlation coefficient

Recap:
The Spearman's rank correlation coefficient measures the tendency of two variables to increase together (i.e. does y increase when x increases or does y decrease when x decreases).

The key points:

- **r_S always lies between -1 and 1**. (If you calculate a value of r_S that does not lie between -1 and 1 then you've made a mistake!!).
- **Always remember to rank the data.**
- The formula for calculating r_S is

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

where d is the **difference in the ranks**.

Note: This formula is in the formula book.

- **Example:**

The table shows the hardness of the water supply (measured in units of parts per million of calcium) and the cardiovascular disease death rate (in units of deaths per 10000 population) in a random sample of ten English towns.

Town	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Hardness	20	94	119	106	131	77	61	127	32	45
Death rate	77	62	65	58	46	70	51	44	74	82

Calculate the value of Spearman's rank correlation coefficient for these data and interpret what this tells you about the possible relationship between the variables.

Solution:
We first **rank** the data – it doesn't matter whether we rank from highest to lowest or from lowest to highest so long as we do the same for both variables.

Town	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Rank Hardness	1	6	8	7	10	5	4	9	2	3
Rank Death rate	9	5	6	4	2	7	3	1	8	10
d	-8	1	2	3	8	-2	1	8	-6	-7
d^2	64	1	4	9	64	4	1	64	36	49

So $\sum d^2 = 296$

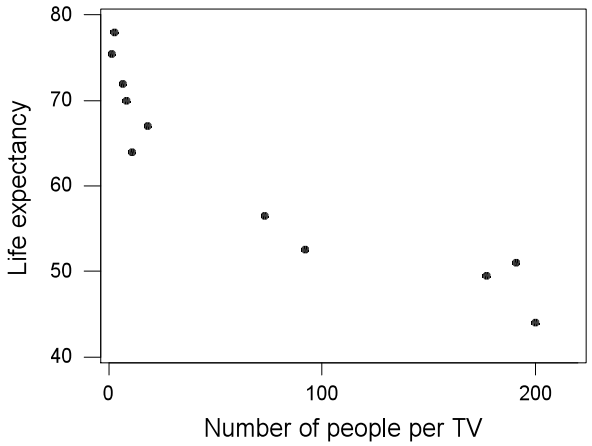
$$r_S = 1 - \frac{6 \times \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 296}{10 \times (100 - 1)} = -0.794 \quad \text{(to 3 sf).}$$

There is negative rank correlation between hardness and death rate. This means that the death rate tends to fall as the level of water hardness increases.

Note: Correlation coefficients do not show a causal relationship.

Example: Information about two variables (life expectancy and the number of people per television set) is available for 12 countries (as shown in the following diagram):

Life expectancy plotted against number of people per TV



It is clear that the two variables are negatively correlated meaning that people living in countries with more TVs tend to live longer. However, it clearly would be wrong to conclude that simply sending more televisions to countries with low life expectancies would cause their inhabitants to live longer.

This example illustrates the very important distinction between causation and association. Two variables may be strongly correlated without a cause-and-effect relationship existing between them.

Often the explanation is that both variables are related to a 3rd variable not being measured. In the example above for instance both life expectancy and the number of TVs in the population will both be related to the country's wealth.