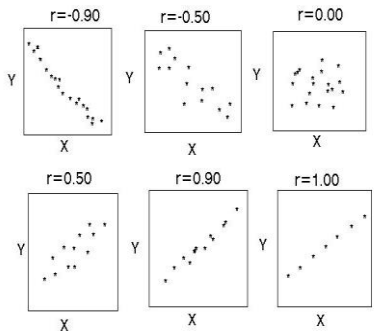
Week 3
Term 1
2021

HAWKER COLLEGE
Engage | Inspire | Achieve

Learning Brief

MA3



Goals

This week:

- calculate and interpret the correlation coefficient (p and *r*) to quantify the strength of a linear association
- use the coefficient of determination to assess the strength of a linear association in terms of the explained variation
- recognise that an observation association between two variables does not necessarily mean that there is a causal relationship between them

Theoretical Components

Resources:
For this week the theory work is in the *PDF file:* Week *3* Notes & Exercises

Correlation vs causation

https://www.youtube.com/watch?v=VMUQSMFGBDo

Coefficient of determination

https://www.youtube.com/watch?v=qQMAjsOihYc

Knowledge Checklist

- Scatter plots
- Calculating and interpretting correlation coefficients
- Correlation and causation
- Coefficient of determination
- Variation in a variable

Practical Components

There are questions to be answered in the booklet *Week 3 Notes & Exercises*

Investigation

See the end of the brief ☺

Quiz

No mathspace.co for this week.

MATHEMATICAL APPLICATIONS 3

WEEK 3 NOTES & EXERCISES

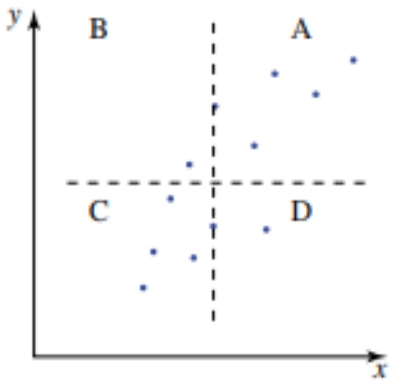
q-correlation coefficient

Last week, we looked at scatter plots and the idea of correlation. It is useful to have a quantitative measurement of the strength of the relationship shown between a pair of variables. This measurement is known as a correlation coefficient.

There are several different methods used for obtaining a correlation coefficient. One relatively simple method is called the q-correlation coefficient. It can be found by using the steps detailed below.

Step 1

Draw a vertical line on the scatter plot of the data to indicate the position of the median of the x values. There should be an equal number of points on each side of this line. Then draw a horizontal line to indicate the position of the median of the y values. Note that these lines may or may not pass through one or more of the data points depending on whether there is an even or odd number of points.



Step 2

The scatter plot is now divided into four quadrants that are labelled A, B, C and D as in the diagram above. The order of the labelling is important.

Step 3

Count the number of data points that lie in each of the quadrants. Points which lie on either of the median lines are omitted.

Step 5

The correlation coefficient is calculated using the formula:

$$q = \frac{(a + c) - (b + d)}{a + b + c + d}$$

where:

- a is the number of data points in A

- b is the number of data points in B
- c is the number of data points in C
- d is the number of data points in D.

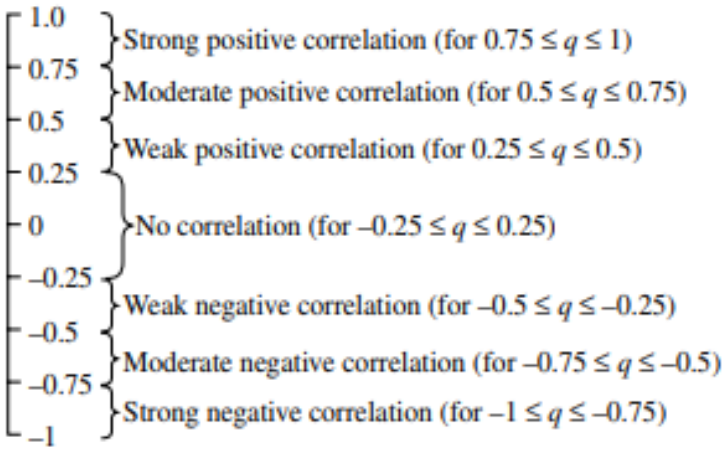
For these data: a = 5, b = 1, c = 5, d = 1

$$q = \frac{(a + c) - (b + d)}{a + b + c + d}$$

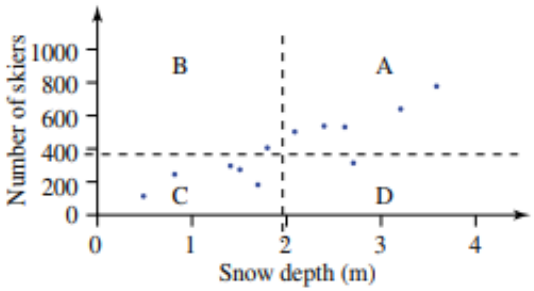
$$q = \frac{(5 + 5) - (1 + 1)}{5 + 1 + 5 + 1}$$

$$q = 0.67 \text{ (to 2 decimal places)}$$

The correlation coefficient may be interpreted by using the scale below.



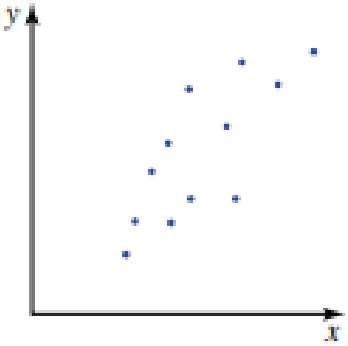
The data used in this calculation is from the data about the ski resort on the first page of the Week 1 Notes.



A q-correlation coefficient of 0.67 indicates a moderate positive correlation. In this case, we would conclude: There is evidence to show that the greater the depth of snow, the greater the number of skiers. Notice that the scale for correlation coefficients runs from -1 to 1. Every correlation coefficient must lie within this range. Any answers outside the range would indicate an error in workings.

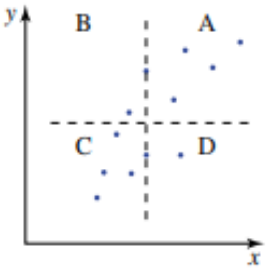
Example

For the graph below, a) calculate the q-correlation coefficient and b) state what type of correlation is involved.



Solution

a. First find the position of the median lines and label the quadrants. Second count the number of data points in each quadrant, ignoring those that lie on a line. Lastly use the formula to calculate the q-correlation coefficient.



a = 4, b = 1, c = 4, d = 1

$$q = \frac{(a+c)-(b+d)}{a+b+c+d}$$

$$q = \frac{(4+4)-(1+1)}{4+1+4+1}$$

$$q = 0.6$$

b. There is a moderate positive correlation.

Exercise 1

1. What type of correlation would be represented by scatter plots which have the following correlation coefficients?
2.
   a. 0.4

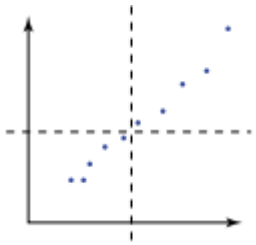
   b. 0.8

   c. 0.7

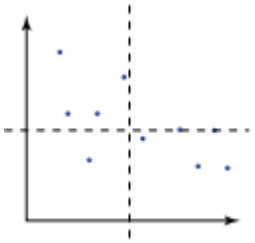
3. For each of the following graphs: i) calculate the q-correlation coefficient and ii) state what type of correlation is involved.

$$q = \frac{(a + c) - (b + d)}{a + b + c + d}$$

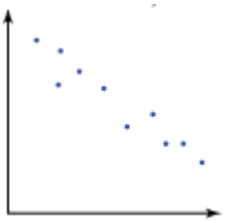
a.



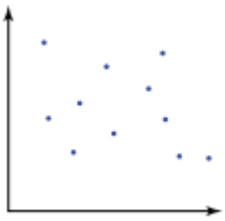
b.



c.



d.



Pearson's correlation coefficient (*r*)

There are limitations to the q-correlation coefficient. The q-correlation coefficient is useful because it is easy to apply to a scatter plot, but it is not the most sophisticated or reliable way of finding a measure of linear association. A much more rigorous coefficient *r*, which is found by the formula:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where:

- n is the number of pairs of data in the set
- s_x is the standard deviation of the x values
- s_y is the standard deviation of the y values
- $\bar{x}$ is the mean of the x values
- $\bar{y}$ is the mean of the y values

This can be complicated and tedious to calculate. Fortunately, we can use online calculators to do it for us. There are two important limitations on the use of r. The calculation of r is applicable to sets of bivariate data which are known to be linear in form and which do not have outliers.

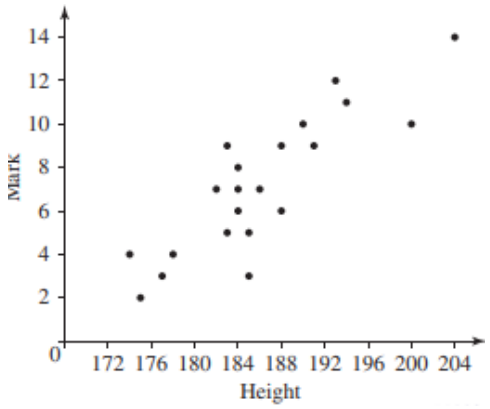
Example

The heights (in centimetres) of 21 football players were recorded against the number of marks they took in a game of football. The data are shown in the following table. Draw a scatter plot to show the data

Height (cm)	Number of marks taken
184	6
194	11
185	3
175	2
186	7
183	5
174	4
200	10
188	9
184	7
188	6

Height (cm)	Number of marks taken
182	7
185	5
183	9
191	9
177	3
184	8
178	4
190	10
193	12
204	14

Solution



Height is the independent variable, so it is plotted on the x axis.

Using the calculator on https://www.socscistatistics.com/tests/pearson/default2.aspx, we get a value of *r* to be 0.86. This indicates there is a strong positive linear association between the height of a player and the number of marks they take in a game. That is, the taller the player, the more marks we might expect them to take.

Correlation and causation

Just because variables are highly correlated, we cannot infer a casual relationship. Two variables *X* and *Y* may be highly correlated because:

1. *X* causes *Y*
2. *Y* causes *X*
3. Some third influence causes both *X* and *Y*

We cannot decide which of these applies simply based on a correlation coefficient.

Solution (for the above example)

From the example above about football player's height and number of marks taken, we concluded *r* = 0.86. While we can say that there is a strong association between the height of a football player and the number of marks he takes, we cannot state that the height of a football player causes him to take a lot of marks. Being tall might assist in taking marks, but there will also be many other factors which come into play; for example, skill level, accuracy of passes from team mates, abilities of the opposing team, and so on.

A famous example of correlation and causation concerns the high positive correlation between the number of human births and the stork population of European towns.

Storks generally nested in chimneys, so the larger towns with more chimneys had a larger stork population. Since larger towns also have larger populations of people, these towns also produced more babies. The third variable, the size of the town, was responsible for the high correlation, not the theory storks bring babies. We must be careful attributing causality is a logical problem not a statistical one.

The coefficient of determination (r^2)

The coefficient of determination is given by r^2. It is found simply by squaring r. its value also ranges from 0 to 1. The coefficient of determination tells us the proportion of variation in one variable that can be explained by the variation in the other variable.

Solution (for the above example)

From the example above about football player's height and number of marks taken, $r = 0.86$ and $r^2 = 0.74$. Thus 74% of the variation in the number of marks can be explained by the variation in the height of the football players.

Exercise 2

1. The yearly salary (x$1000) and the number of votes polled in the Brownlow medal count are given below for 10 footballers.
2.

Yearly salary (x$1000)	180	200	160	250	190	210	170	150	140	180
Number of votes	24	15	33	10	16	23	14	21	31	28

   a. Construct a scatter plot for the data.

   b. Calculate r using the link above and comment on the correlation between salary and number of votes

   c. Calculate the coefficient of determination for the data and interpret its value.

Thus _____ % of the variation in the number of _____ can be explained by the variation in the _____ of the football players.

3. A set of data, obtained from 40 smokers, gives the number of cigarettes smoked per day and the number of visits per year to the doctor. The Pearson's correlation coefficient for these data was found to be 0.87. Calculate the coefficient of determination for the data and interpret its value.

4. The data below shows the number of people in 9 households against weekly grocery costs.

Number of people in households	2	5	6	3	4	5	2	6	3
Weekly grocery costs ($)	60	180	210	120	150	160	65	200	90

   a. Construct a scatter plot for the data.

   b. Calculate *r*.

   c. Calculate the coefficient of determination and interpret its value.

5. A set of data was obtained from a large group of women with children under 5 years of age. They were asked the number of hours they worked per week and the amount of money they spent on childcare. The results were recorded, and the value of Pearson's correlation coefficient was found to be 0.92.
Classify each of the following statements as True or False.

    a. There is a positive correlation between the number of working hours and the amount of money spent on childcare.

    b. The correlation between the number of working hours and the amount of money spent on childcare can be classified as strong.

    c. As the number of working hours increases, the amount spent on childcare increases as well.

    d. The increase in the number of hours worked causes the increase in the amount of money spent on childcare.

    e. The coefficient of determination is about 0.85.

    f. The number of working hours is the major factor in predicting the amount of money spent on childcare.

    g. About 85% of the variation in the number of hours worked can be explained by the variation in the amount of money spent on childcare.

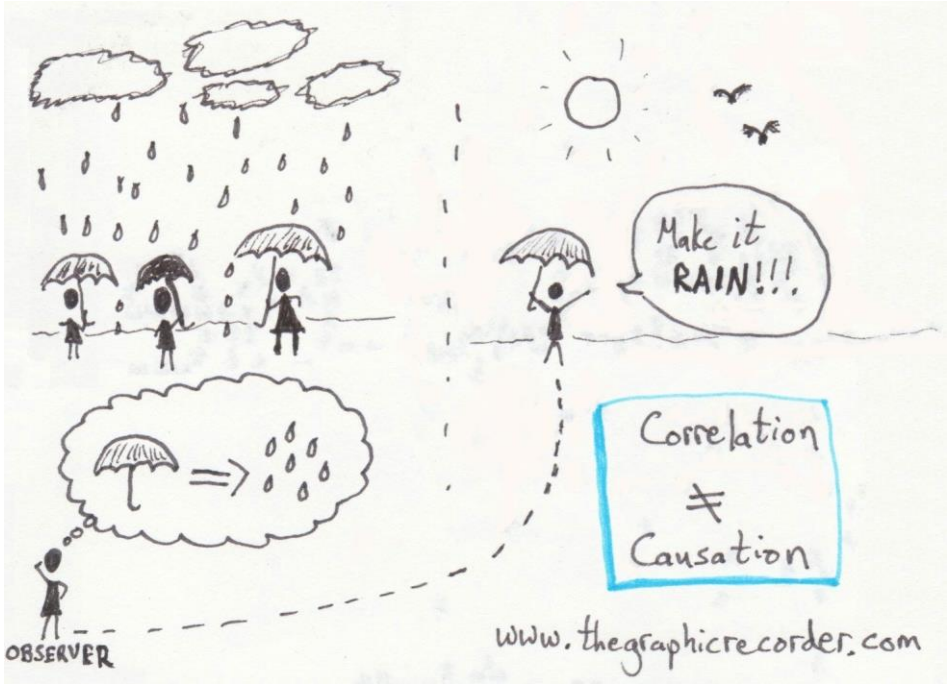
    h. Apart from number of hours worked, there could be other factors affecting the amount of money spent on childcare.

6. The data below shoes the annual advertising budgets (x$1000) and the yearly profit increases (%) of 8 companies. Calculate the coefficient of determination and interpret the results.

Annual advertising budget (x$1000)	11	14	15	17	20	25	25	27
Yearly profit increase (%)	2.2	2.2	3.2	4.6	5.7	6.9	7.9	9.3

Investigation Week 3

Describe/explain the following cartoon in relation to causation and correlation.



Marking Rubric **Name:**

CRITERIA	EXPECTATIONS	POSS	MULT	GIVEN	TOTAL
Practical	Student completes practical work, including exercises and Mathspace task, of the brief to an acceptable standard set by the teacher.	**2**	**3**		**/6**
Investigation Task	Student completes the investigation task of the week to an acceptable standard set by the teacher.	**2**	**2**		**/4**
Reasoning and Communications	Student responses are accurate and appropriate in presentation of mathematical ideas, with clear and logical working out shown.	**4**	**-**		**/4**
Concepts and Techniques	Student submitted work selects and applies appropriate mathematical techniques to solve practical problems and demonstrates proficiency in the use of mathematical facts, techniques, and formulae.	**4**	**-**		**/4**
	Submission Guidelines				
Timeliness	Student submits the exercises, Mathspace/online task and investigation by the set deadline. See scoring guidelines for specific details.	**2**	**-**		**/2**
				FINAL	**/20**

Student Reflection: How did you go with this week's work? What did you learn? What did you find easy?
What do you need to work on?