Pearson's correlation coefficient (r)

Year 11
General Mathematics

www.maffsguru.com

Learning Objectives

By the end of the lesson I hope that you understand and can apply the following to a range of questions from the Unit 1 and 2 General Mathematics course.

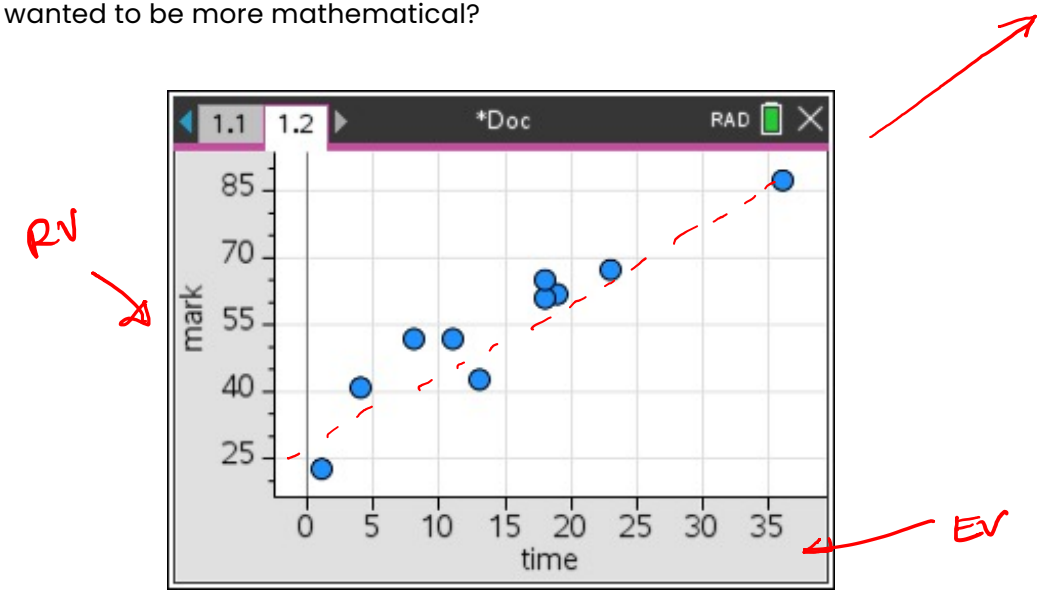
- Know what correlation means
- Know that we have a measure of the correlation between two variables
- Be able to use the correlation coefficient to describe linear relationships
- Understand what causation is and the three main types:
    - Common response
    - Confounding
    - Coincidence

Recap of past learning

In the previous lessons we have been looking at the very exciting topic of statisitcs! So much better than just having to find the mean, median, mode and range! At least now we're dealing with real world data and drawing pretty pictures.

If we look the following graph we can see that the data is pretty grouped together. It seems to be sort of a straight line and it seems to be heading up.

We now have a way of describing the points in terms of direction, form and strength (strong, moderate and weak) but what if we wanted to be more mathematical?



Pearson's correlation coefficient

I have no idea who Pearson was … but he came up with a "mathematical" way of describing how close the points of a scatter plot are to the line of best fit (or linear regression line).

To short cut the learning … we use his table (which is shown below):

strong positive linear association r between 0.75 and 0.99
moderate positive linear association r between 0.5 and 0.74
weak positive linear association r between 0.25 and 0.49
no linear association r between −0.24 and 0.24
weak negative linear association r between −0.25 and −0.49
moderate negative linear association r between −0.5 and −0.74
strong negative linear association r between −0.75 and −0.99

Pearson's correlation coefficient

We can see that the wording we use is directly dependent on the value of 'r' which we get from a set of data. This is generally worked out for us be the CAS (we we will see in the next video).

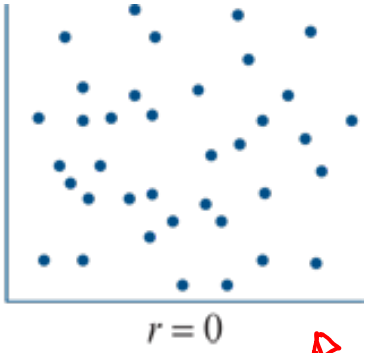
Once we have the value of 'r' we simply read the correct line from the table.

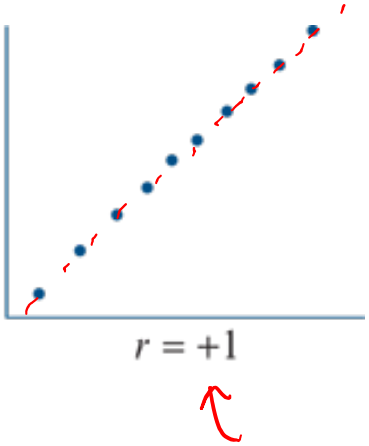
strong positive linear association r between 0.75 and 0.99
moderate positive linear association r between 0.5 and 0.74
weak positive linear association r between 0.25 and 0.49
no linear association r between -0.24 and 0.24
weak negative linear association r between -0.25 and -0.49
moderate negative linear association r between -0.5 and -0.74
strong negative linear association r between -0.75 and -0.99

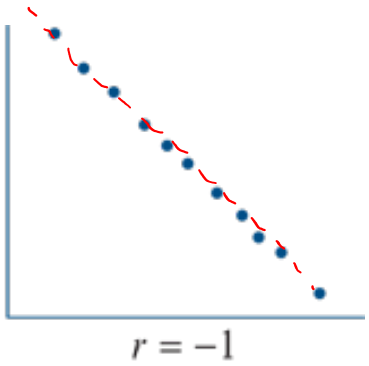
Examples have been extracted, with permission, from the Cambridge General Mathematics Units 1 and 2 Textbook

Some examples

Let's look at some examples of scatter plots and their 'r' values (gained from the CAS).







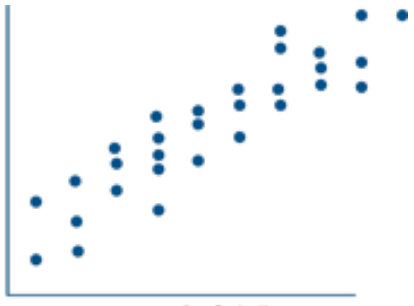
$r^2 \rightarrow r$

$\sqrt{r^2} = \pm r$

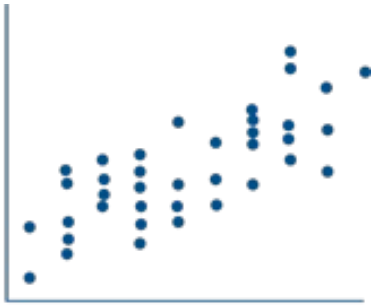
Examples have been extracted, with permission, from the Cambridge General Mathematics Units 1 and 2 Textbook

More likely examples …

The following are more likely!



$r = 0.915$



$r = 0.767$

$r = -0.874$

strong positive linear association r between 0.75 and 0.99
moderate positive linear association r between 0.5 and 0.74
weak positive linear association r between 0.25 and 0.49
no linear association r between −0.24 and 0.24
weak negative linear association r between −0.25 and −0.49
moderate negative linear association r between −0.5 and −0.74
strong negative linear association r between −0.75 and −0.99

Examples have been extracted, with permission, from the Cambridge General Mathematics Units 1 and 2 Textbook

Important information

The following is so important, it deserves a slide of its own!

The correlation coefficient (called 'r'):

- Measures the strength of a **linear** relationship.
- It has a value between -1 and 1
- Positive numbers mean a positive relationship
- Negative numbers mean a negative relationship

SACs try and trick you with this one next year. This can only be used on LINEAR relationships.

Non-linear relationships will also give you a value of 'r' on the CAS.

r assume linear

both variables numerical

Causation

This is, probably, one of the stupidest areas on the General Maths course (being honest).

We need to understand that most things generally have a cause and effect.

COVID-19 might cause exam results to go down as a result of people not liking home learning for example.

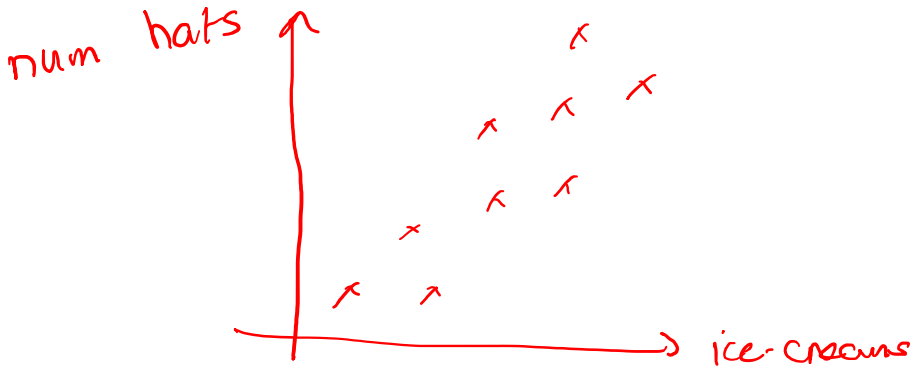
Not everything is linked though. Even though it might seem that it is!

For example, we know that there is a link between the temperature and the number of ice-creams which will be sold on a particular day.

This would have a **positive relationship** which might be quite strong!

Is there a link between the number of IKEAs a country has and the number of NOBEL prizes won? Absolutely not! Even though a graph might show otherwise. This is a complete coincidence.

Common response



Examples have been extracted, with permission, from the Cambridge General Mathematics Units 1 and 2 Textbook

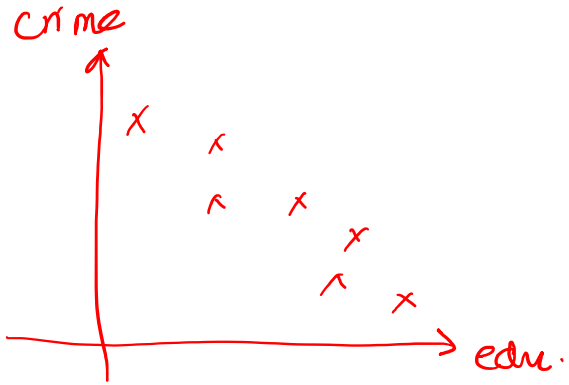
Common response

Let's get this over and done with an more onto something more interesting!

This is when two variables are linked by a third.

Question:
If data about the variable crime rates and unemployment in a range of cities were gathered, a high correlation would be found. Can it be inferred that high unemployment causes high crime rates?

Answer:
A common link (or common response) might actually be *level of education* ***rather than*** a direct link between crime rates and unemployment.



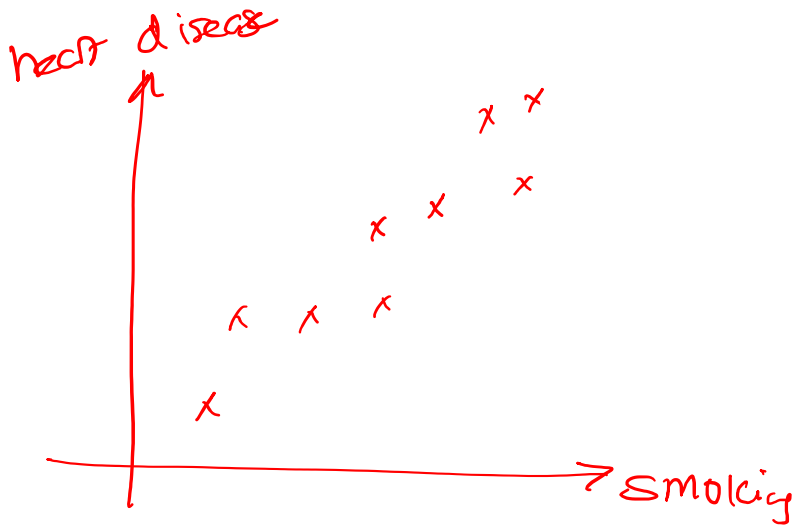
Examples have been extracted, with permission, from the Cambridge General Mathematics Units 1 and 2 Textbook

Confounding

There might be a link to one of them, but not both.

Question:
Suppose we were to find a high correlation between the smoking rate and heart disease across a group of countries. Can we conclude that smoking causes heart disease?

Answer:
One possible explanation is that people who smoke are also prone to neglect other lifestyle factors such as exercise and diet. It could well be that people who smoke also tend not to exercise regularly, and it is the lack of exercise which causes heart disease.



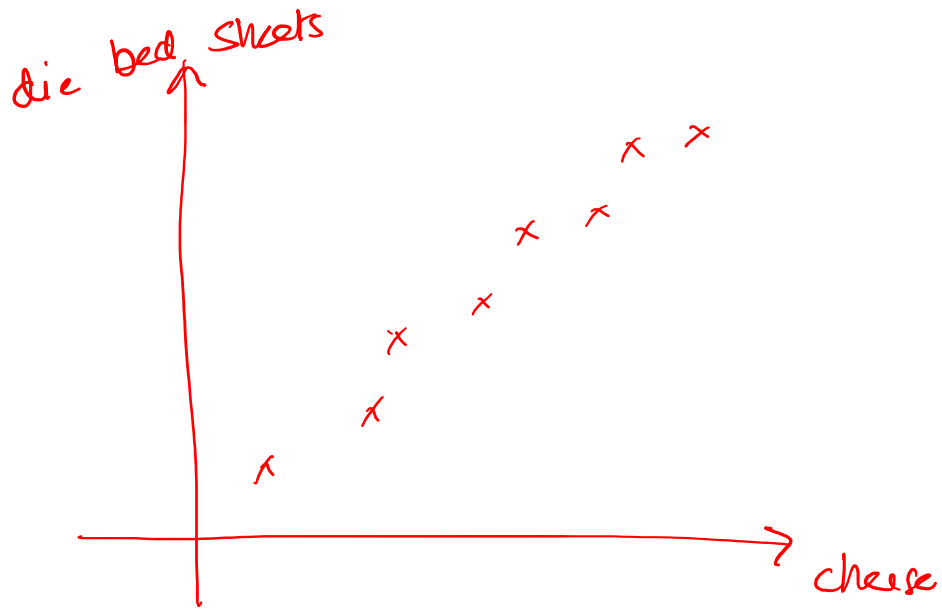
Examples have been extracted, with permission, from the Cambridge General Mathematics Units 1 and 2 Textbook

Coindidence

There is no actual relationship … it's just a coincidence

Question:
It turns out that there is a strong correlation (r=0.95) between cheese consumption and the number of people who died becoming tangled in their bed sheets.

Answer:
We will leave that one there!



Examples have been extracted, with permission, from the Cambridge General Mathematics Units 1 and 2 Textbook