Association Between Variables

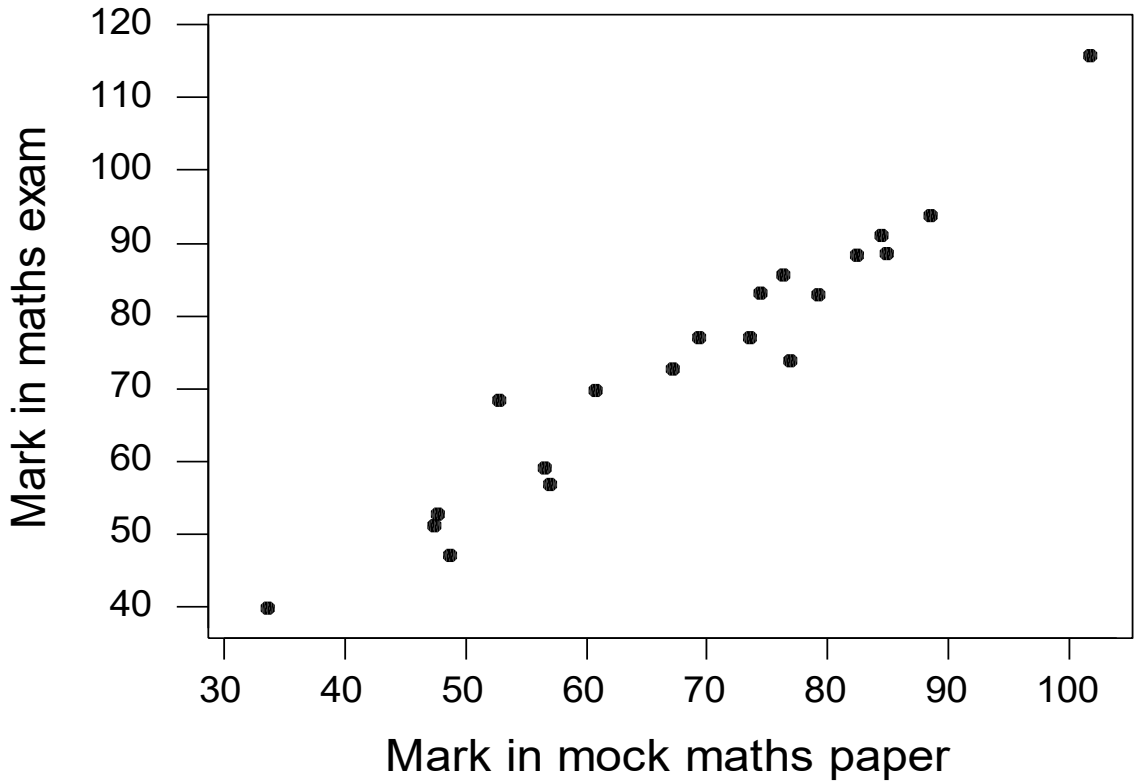
Jian Zhang

University of Kent

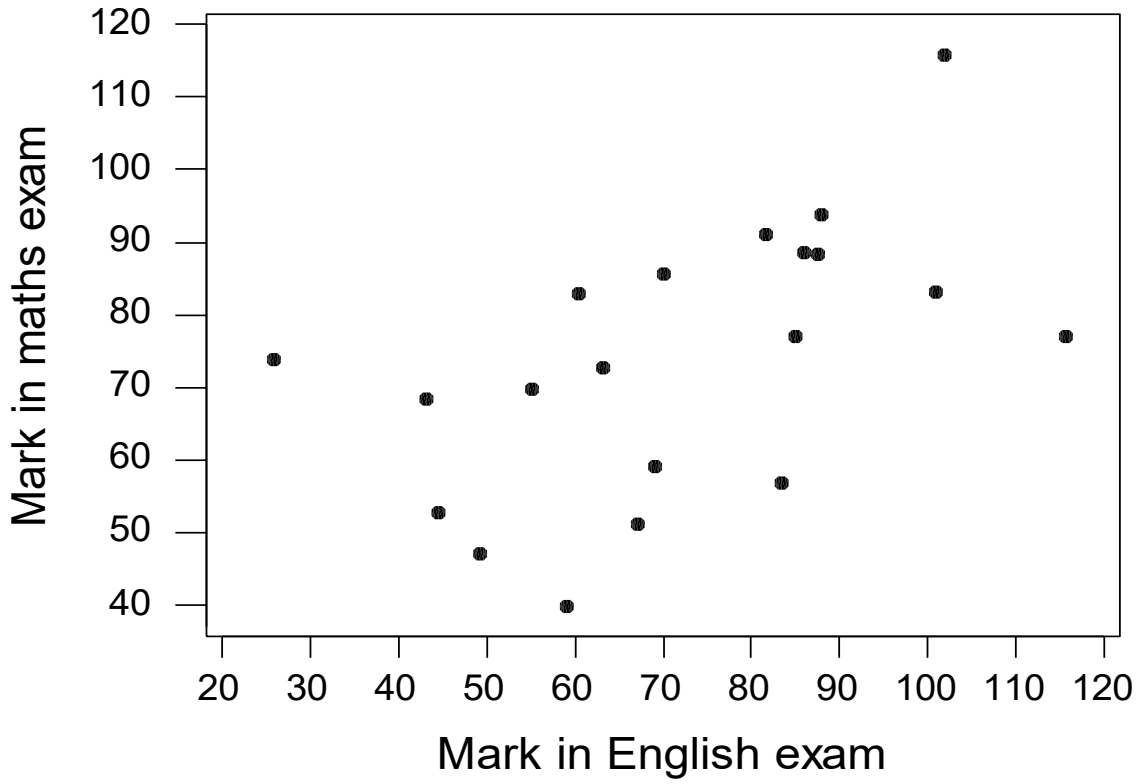
Association between variables

- Consider two random variables. These may be related to each other – for example, heights and weights of people are related.
- We now consider how to measure the strength of relationship between two random variables.

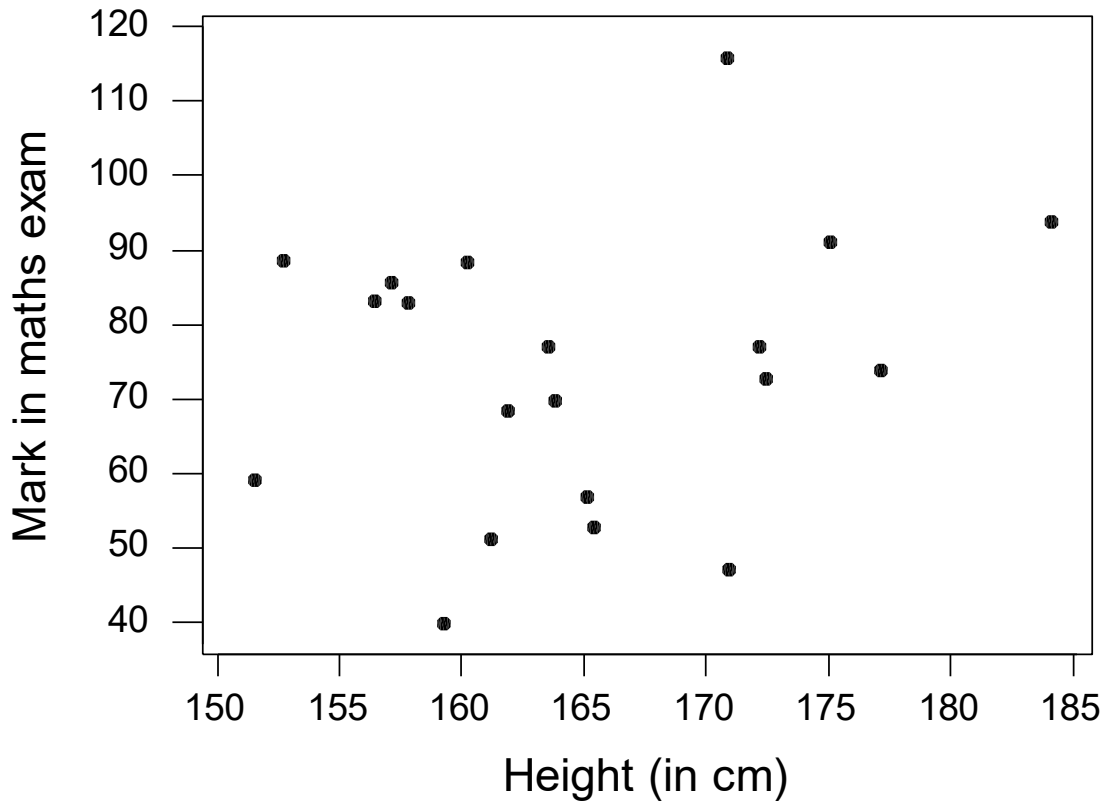
- Consider the following scatter plots showing (hypothetical) data from 20 school children



A



B



C

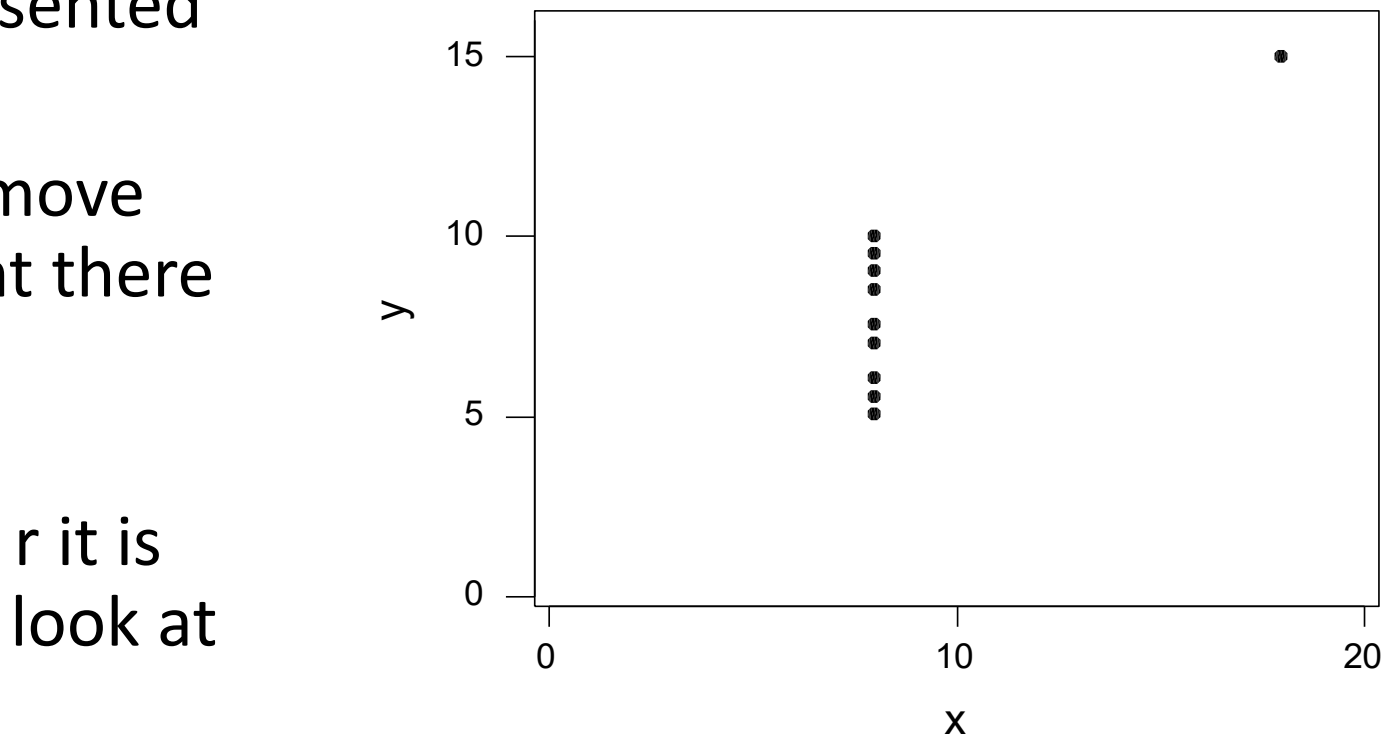
- In (A) there is a strong positive relationship between the mark in the maths exam and mark in the mock paper
- In (B) there is a weak positive relationship between English and maths marks
- In (C) there appears to be no relationship between height and maths mark

Product moment correlation coefficient

- The product moment correlation coefficient, r, gives a summary measure of the strength of (linear) association between two random variables. r can take values in the range [-1,1].
- If r is positive (negative), this indicates a positive (negative) relationship between the variables.
- The further r is from 0, the stronger the association between the two random variables.

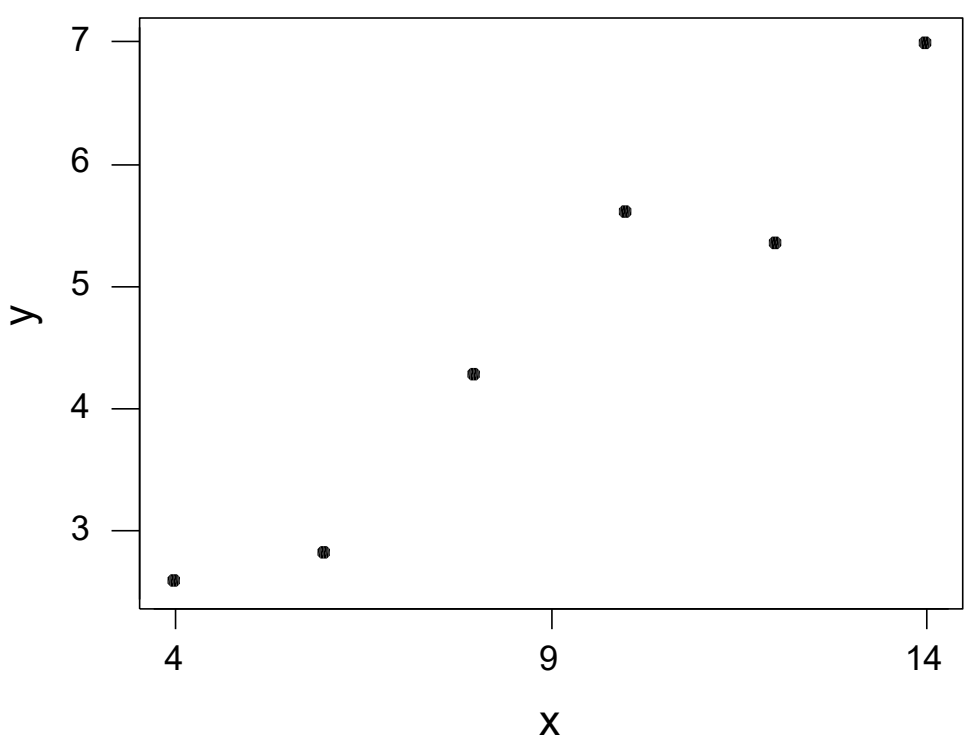
Importance of visualising data

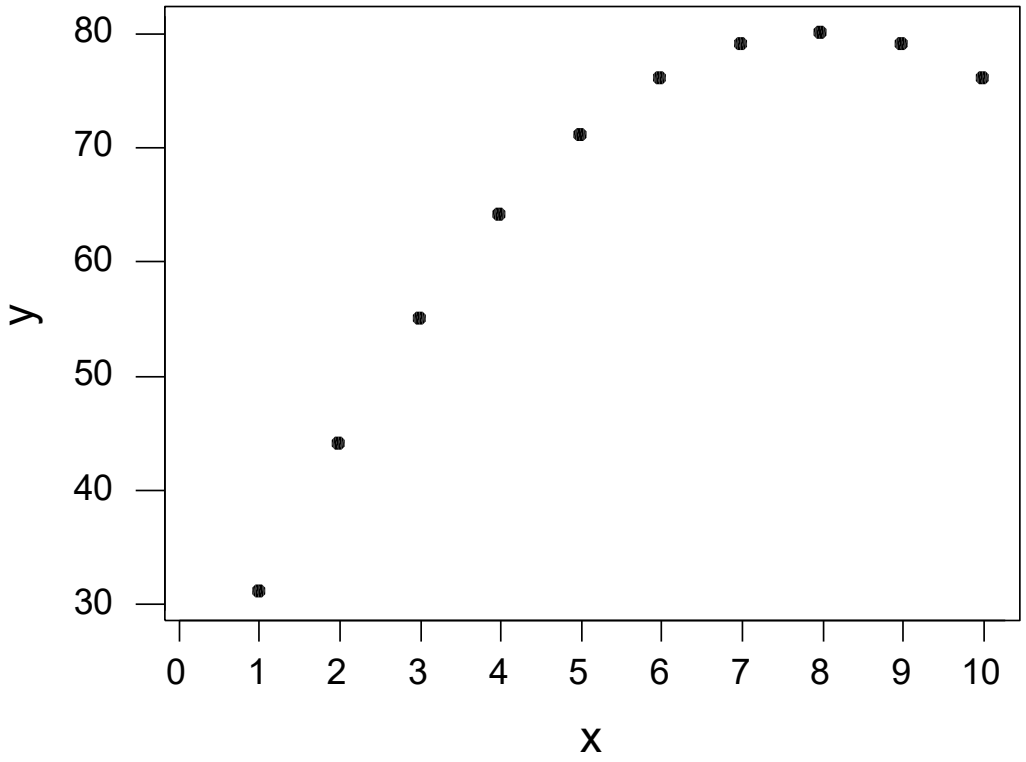
- Consider the data presented on this plot.
- r = 0.81, but if you remove the outlying data point there is in fact no linear relationship.
- So, when interpreting r it is always a good idea to look at a plot of the data.



Spearman rank correlation coefficient

- The Spearman rank correlation coefficient, r_s, is more general than the product moment correlation coefficient as it measures the strength of the monotonic (i.e. always moving in a consistent direction) association.





- The product-moment correlation coefficient would have problems with (B) but Spearman's correlation can handle both

Obtaining correlations in R

- In R to calculate the product moment correlation coefficient

```
> cor.test(data1,data2)
```

- In R to calculate the Spearman rank correlation coefficient

```
> cor.test(data1,data2,method="spearman")
```

Correlation…not Causation

- Just because two variables are correlated you cannot deduce causality