MM104 / MM107
Statistics and Data Presentation Semester 2

Project 6: Correlation

Ainsley Miller
ainsley.miller@strath.ac.uk

Lecture Overview

In this lecture we will

- learn about correlation.
- learn the definition of correlation and the range of values the correlation coefficient can take.
- learn how to calculate the correlation coefficient by hand and using Minitab.
- learn how to interpret the correlation coefficient.
- learn about the importance of correlation not causation.

Correlation Motivation

Sometimes we are interested in finding out the relationship between variables, more precisely, how the value of one variable changes when the value of another variable changes. We express this via the **correlation coefficient**.

More formally we can say that:

> Correlation is the degree of **linear association** between two **numerical variables**.

Correlation Best Practises

- Remember, it is extremely bad statistical practise to use the word correlation when you are not carrying out a test for correlation ! But this week you can finally say the word correlation as you will be testing for correlation.
- Correlation is best visualised as a scatterplot.

Correlation coefficient

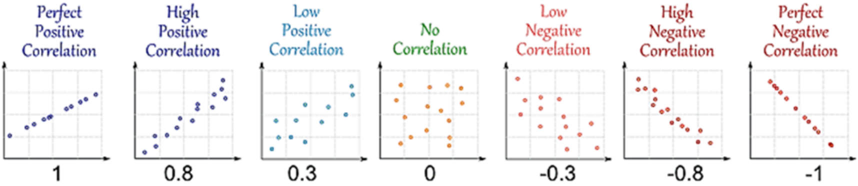
The first step is to visualise your data in terms of a scatterplot. This gives you a **subjective interpretation** as to whether or not the data is correlated.

This is then confirmed by a formal statistical test - Pearson's correlation coefficient. We use the letter r when discussing correlation. r denotes the sample correlation coefficient.

A correlation coefficient is a numerical index that reflects the relationship between two variables.

Correlation Coefficient cont...

- The value of the correlation coefficient ranges between -1 and 1.
- The absolute value of coefficient reflects the strength of the correlation.
- $r = 1$ Perfect positive correlation.
- $r = -1$ Perfect negative correlation.
- $r = 0$ No correlation



Interpreting a Correlation Coefficient

In the previous slide we used low, high and perfect to describe correlation. The following table has more categories and is therefore more informative.

Absolute Value of the Correlation Coefficient	General Interpretation
0.00	No relationship
0.01 – 0.19	Very weak relationship
0.20 – 0.39	Weak Relationship
0.40 – 0.59	Moderate Relationship
0.60 – 0.79	Strong Relationship
0.80 – 0.99	Very strong relationship
1.00	Perfect relationship

Calculating the correlation coefficient

The formula for the Pearson's correlation coefficient, $r_{x,y}$, between your x variable and your y variable is:

$$r_{x,y} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - (\sum x)^2\right)\left(n\sum y^2 - (\sum y)^2\right)}},$$

where n is the size of the sample, and x, y and $r_{x,y}$ have already been defined.

Consider an Example

Calculate the correlation coefficient of the following set of data.

x	y
2	3
4	2
5	6
6	5
4	3
7	6
8	5
5	4
6	4
7	5

Solution

The easiest way to do this by hand is to extend our table:

	x	y	x^2	y^2	xy
	2	3	4	9	6
	4	2	16	4	8
	5	6	25	36	30
	6	5	36	25	30
	4	3	16	9	12
	7	6	49	36	42
	8	5	64	25	40
	5	4	25	16	20
	6	4	36	16	24
	7	5	49	25	35
$\sum$	54	43	320	201	247

Solution cont...

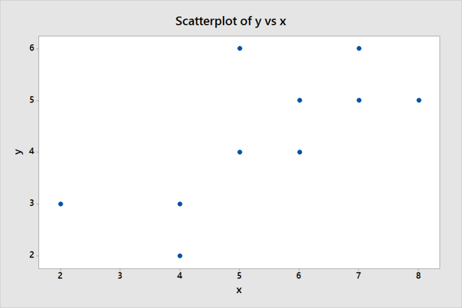
$$
\begin{aligned}
r_{x,y} &= \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - (\sum x)^2\right)\left(n\sum y^2 - (\sum y)^2\right)}} \\
&= \frac{(10\times 247) - (54\times 43)}{\sqrt{\left(10\times 320 - (54)^2\right)\left(10\times 201 - (43)^2\right)}} \\
&= \frac{2470 - 2322}{\sqrt{(3200-2916)(2010-1849)}} \\
&= \frac{148}{\sqrt{284\times 161}} \\
&\underline{\underline{= 0.692}}
\end{aligned}
$$

There is strong positive association between x and y.

Correlation in Minitab

Firstly let's produce a scatter-plot of the data in the previous example.

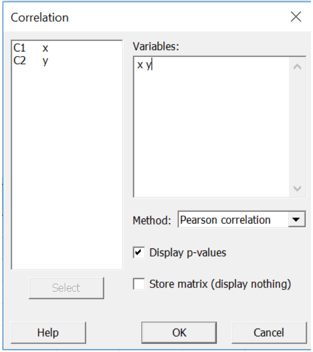
Graph > Scatterplot > Simple.
Then select the appropriate x and y.



Correlation in Minitab cont...

Now let’s find the Pearson correlation coefficient.

Stat > Basic Statistics > Correlation



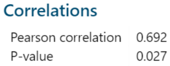
Variables: x y

Method: Pearson correlation

☑ Display p-values

□ Store matrix (display nothing)

Correlation in Minitab cont...



```
Correlations
Pearson correlation   0.692
P-value               0.027
```

We can see that the Pearson correlation co-efficient agrees with our hand calculation.

But why is there a p-value ?

Hypothesis Testing - Correlation

- H_0: $\rho = 0$ i.e. no association
- H_1: $\rho \neq 0$ i.e. a linear association

Here ρ (Greek letter rho - pronounced row) is the population correlation coefficient, and since our p-value was 0.027 which is less than our significance level (0.05) we reject the null hypothesis and conclude that there is a linear association between x and y.

This hypothesis test simply confirms the result from calculating the Pearson correlation coefficient r and does not tell you the strength of the linear association.

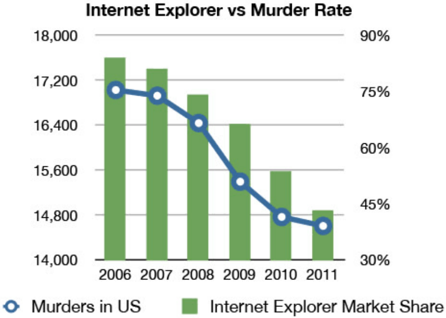
Correlation not causation

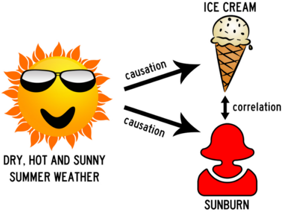
Whilst causation and correlation can exist at the same time, correlation does not mean causation.

It is important to note that, although a high r value indicates a strong linear positive or negative relationship between the variables it does not not necessarily mean that **changes in one variable cause changes in another**. Correlation is simply a relationship, whereas causation explicitly applies to cases where action A causes action B.

The relationship may also be entirely coincidental this is called **spurious correlation**.

Correlation not causation cont...


Internet Explorer vs Murder Rate
Murders in US
Internet Explorer Market Share


ICE CREAM
causation
correlation
causation
DRY, HOT AND SUNNY
SUMMER WEATHER
SUNBURN