

Correlation

The objectives of this unit are to:

- * recap the ideas of scatter graphs and correlation met at GCSE;
- * introduce the Product-Moment Correlation Coefficient as a more formal way of describing the correlation between 2 variables.

1 Revision from GCSE: Scatter Graphs

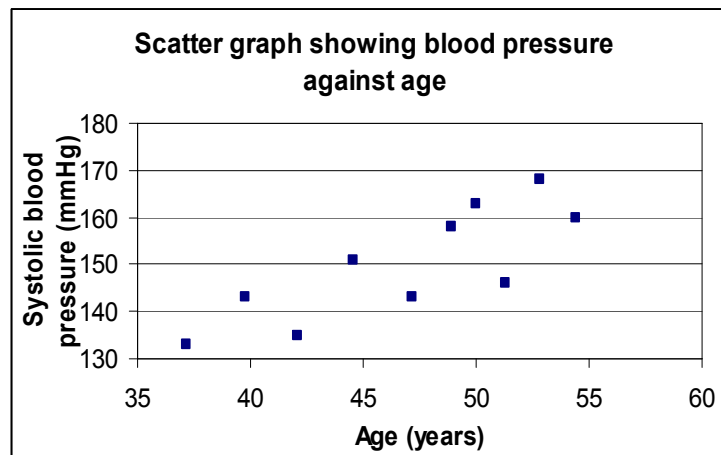
Scatter graphs are useful for assessing relationships between two variables.

- **Example 1:**

A company doctor is investigating the possible effect of stress upon the health of the company's management employees. She suspects that employees under stress will suffer from high systolic blood pressure. She takes a random sample of ten employees, aged between 35 and 55 years, and records their age and blood pressure:

Management employee	Age (x)	Systolic blood pressure (y)
A	37.2	133
B	39.8	143
C	42.1	135
D	44.6	151
E	47.2	143
F	48.9	158
G	50.0	163
H	51.3	146
I	52.8	168
J	54.4	160

A scatter graph showing the data is given below:



The scatter graph shows a relationship between blood pressure and age. Blood pressure appears to increase with age (i.e. older people tend to have higher blood pressure than young people). We say that the two variables are *positively correlated*.

The data in this example are a set of pairs of values for two variables, age and blood pressure. This is an example of **bivariate data**, where two variables are given for each member of the population.

Dependent and independent variables

The scatter graph for example 1 was drawn with age on the x -axis and blood pressure on the y -axis. This emphasises the fact that age might influence a person's BP but the reverse will not be the case. It is normal practice to plot the *dependent variable* (or *response variable*) on the vertical axis and the *independent variable* (or *exploratory variable*) on the horizontal axis.

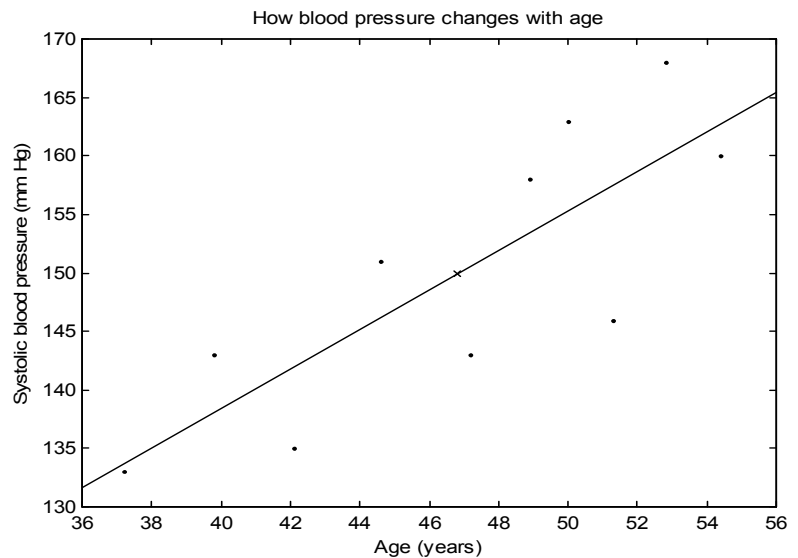
Here are some more examples of dependent and independent variables:

Independent variable	Dependent variable
Number of people in a lift	Total weight of passengers
The number of people visiting a bar in an evening	The volume of beer sold
Amount of fertiliser applied to a crop whilst growing	The crop yield produced

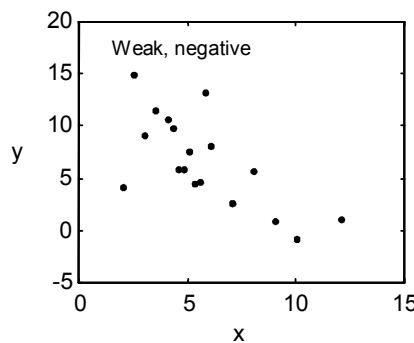
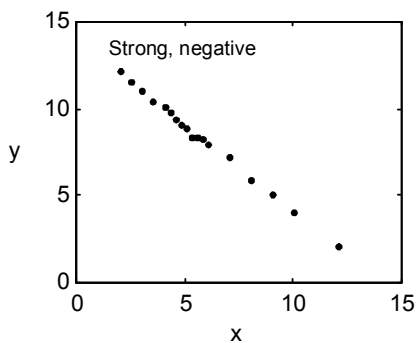
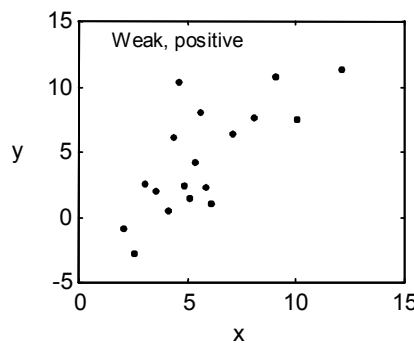
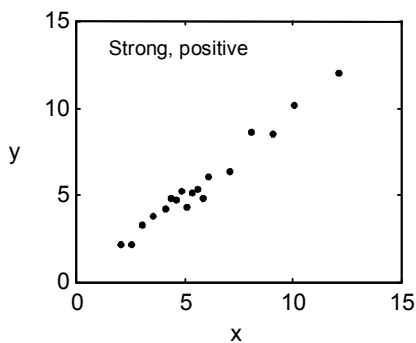
Lines of best fit

If your scatter graph suggests that there may be a linear association between the two variables, a line of best fit can reasonably be drawn on the scatter graph.

Note: A line of best fit will always pass through the mean point (\bar{x}, \bar{y}) .



Positive/ negative correlation



Notice that when variables are correlated almost all the observation points are contained within an ellipse. The narrower the elliptical profile, the greater the correlation.

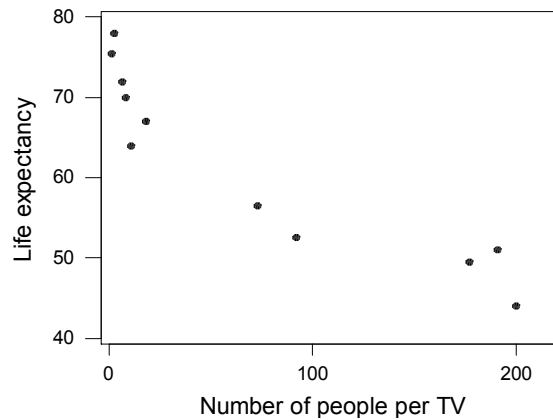
Correlation does not imply causation

It is important to realise that scatter plots point to associations between variables. They do not necessarily show a *causal* relationship.

❖ Example 2:

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. 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.

2 Product-moment Correlation Coefficient

- ✓ The *product-moment correlation coefficient*, r , (or *Pearson's correlation coefficient*) tells you the strength of (linear) association between two random variables (i.e. how close the points on a scatter graph lie to a straight line).
- ✓ The value of r can lie anywhere between -1 and +1.
- ✓ If r is positive, this indicates a positive relationship between the variables. If r is negative, it indicates a negative relationship. The further r is from 0, the stronger the association between the two random variables.
- ✓ $r = +1 \Rightarrow$ exact straight line relationship with positive slope.
- ✓ $r = -1 \Rightarrow$ exact straight line relationship with negative slope
- ✓ $r = 0 \Rightarrow$ complete random scatter (no correlation).

One formula used to find the correlation coefficient is:

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

where:

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

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

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

Note: You **don't** need to learn these formulae as they appear in the formula book!

Example 1 (continued)

For the data of Example 1, the value of r can be seen calculated below:

The mean age of the employees is: $\bar{x} = \frac{37.2 + 39.8 + \dots + 54.4}{10} = 46.83$.

The mean systolic blood pressure is: $\bar{y} = \frac{133 + 143 + \dots + 160}{10} = 150$.

Employee	Age (x)	B P (y)	xy	x ²	y ²
A	37.2	133	4947.6	1383.84	17689
B	39.8	143	5691.4	1584.04	20449
C	42.1	135	5683.5	1772.41	18225
D	44.6	151	6734.6	1989.16	22801
E	47.2	143	6749.6	2227.84	20449
F	48.9	158	7726.2	2391.21	24964
G	50	163	8150	2500	26569
H	51.3	146	7489.8	2631.69	21316
I	52.8	168	8870.4	2787.84	28224
J	54.4	160	8704	2959.36	25600
Total	468.3	1500	70747.1	22227.39	226286

$$\sum x \quad \sum y \quad \sum xy \quad \sum x^2 \quad \sum y^2$$

So, $S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 70747.1 - \frac{468.3 \times 1500}{10} = 502.1$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 22227.39 - \frac{468.3^2}{10} = 296.901$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 226286 - \frac{1500^2}{10} = 1286$$

$$\text{Therefore, } r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{502.1}{\sqrt{296.901 \times 1286}} = 0.813 \text{ (to 3 sf).}$$

This value of r tells you that there is quite strong positive linear correlation between the two variables, i.e. that the points lie fairly close to a straight line with positive slope.

Further example

The table shows the temperature and the relative humidity at one place at regular intervals during one day:

Temperature °F, x	65	68	68	70	72	74	78	81	79	78	77	75
Relative humidity %, y	52	52	53	45	42	33	32	28	30	31	32	32

Find the correlation coefficient.

Solution:

We find that

$$\begin{aligned} \sum x &= 65 + \dots + 75 = 885 \\ \sum y &= 52 + \dots + 32 = 462 \\ \sum x^2 &= 65^2 + \dots + 75^2 = 65557 \\ \sum y^2 &= 52^2 + \dots + 32^2 = 18812 \\ \sum xy &= (65 \times 52) + \dots + (75 \times 32) = 33552 \end{aligned}$$

So,

$$\begin{aligned} S_{xx} &= 65557 - \frac{885^2}{12} = 288.25 \\ S_{yy} &= 18812 - \frac{462^2}{12} = 1025 \\ S_{xy} &= 33552 - \frac{885 \times 462}{12} = -520.5 \end{aligned}$$

So,

$$r = \frac{-520.5}{\sqrt{288.25 \times 1025}} = -0.958$$

Note: In examinations, you are usually given the values of $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$ and $\sum xy$. This saves you time.