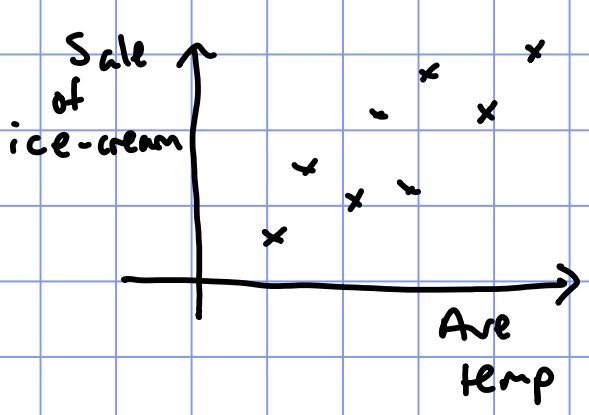
Correlation and Regression

Note Title 06/01/2006

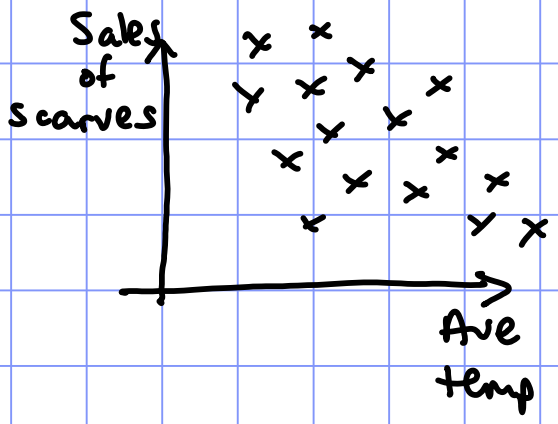
(a) Correlation

This is the strength of the relationship between two variables.

Graphically we can look this using a scattergraph.

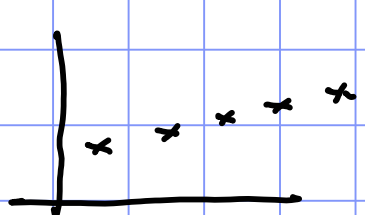


POSITIVE CORRELATION (STRONG)

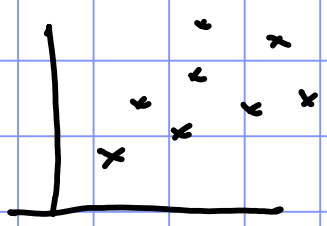


NEGATIVE CORRELATION (WEAKER)

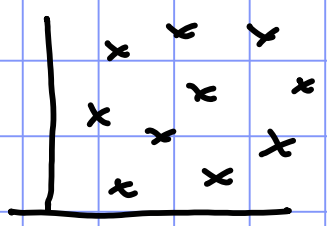
Numerically, we can measure the strength of the correlation by calculating the 'Product-Moment Correlation Coefficient' (PMCC). This always lies between -1 and 1 and is given the letter r.



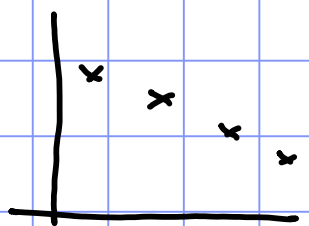
$r = 1$



$r \approx 0.7$



$r \approx 0$



$r = -1$

Example The heights and shoe sizes of a group of Y12 girls are as follows:—

Height (x)	Shoe size (y)
162	5.5
169	6.5
168	6.5
165	6
168	7.5
165	6
169	6
168	6
173	7
164	4.5
160	4
150	4.5
174	7
166	4
169	6.5

$\Sigma x = 2490$

$\Sigma y = 87.5$

$\Sigma x^2 = 413806$

$\Sigma y^2 = 527.75$

$\Sigma xy = 14588$

$n = 15$

We now need to calculate 3 'summary statistics':

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

$$= 413806 - \frac{2490^2}{15}$$

$$= 466$$

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

$$= 527.75 - \frac{(87.5)^2}{15}$$

$$= 17.\dot{3}$$

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

$$= 14588 - \frac{2490 \times 87.5}{15}$$

$$= 63$$

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

$$= \frac{63}{\sqrt{466 \times 17.\dot{3}}}$$

$$= 0.701 \quad (3\text{ dp})$$

This indicates quite a strong correlation between height and shoe size. This can be seen on the scattergraph.

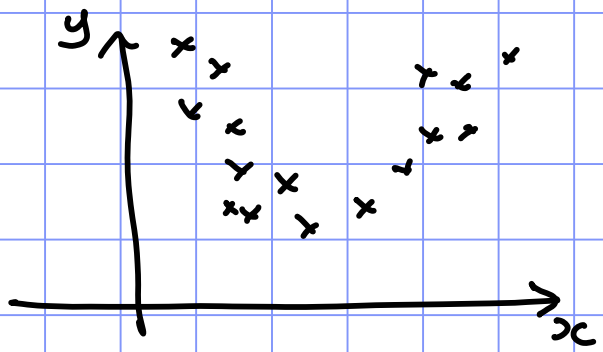
<u>Coding</u> We may wish to simplify the numbers using a coding formula such as $t_i = \frac{x_i - c}{a}$

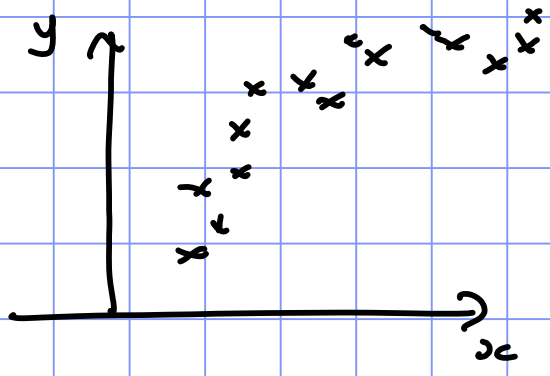
We can code one or both variables. If we use coding, there is NO NEED for ANY DECODING at the end of the calculations.

Interpreting correlation

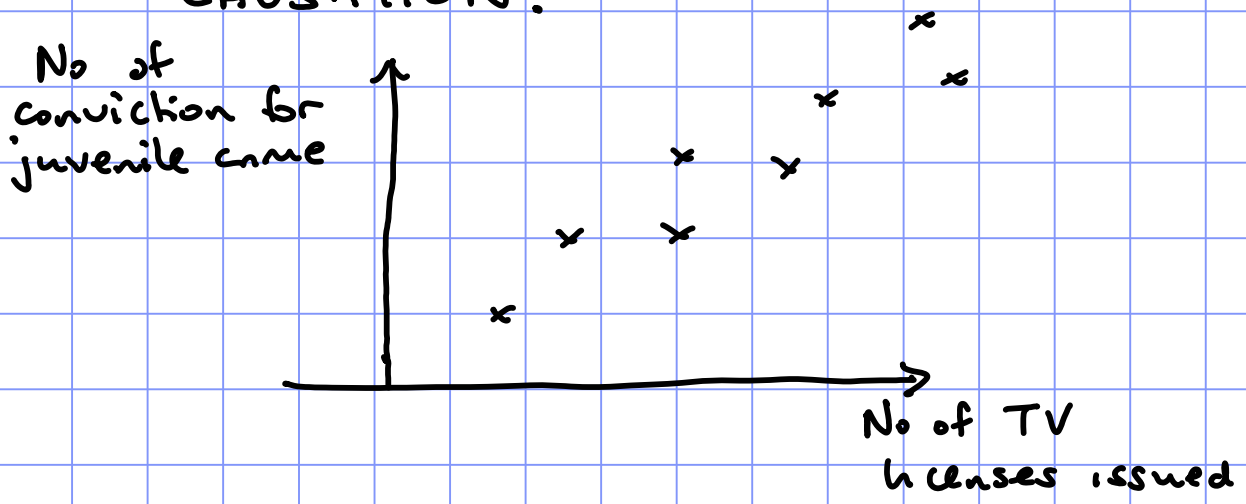
① r is a measure of LINEAR correlation (ie it measures how close the points lie to a straight line).

It is possible for data to exhibit other forms of correlation, but r is not designed to measure this. Measuring this is beyond the scope of S1 - S2.





② Correlation does not necessarily imply CAUSATION.

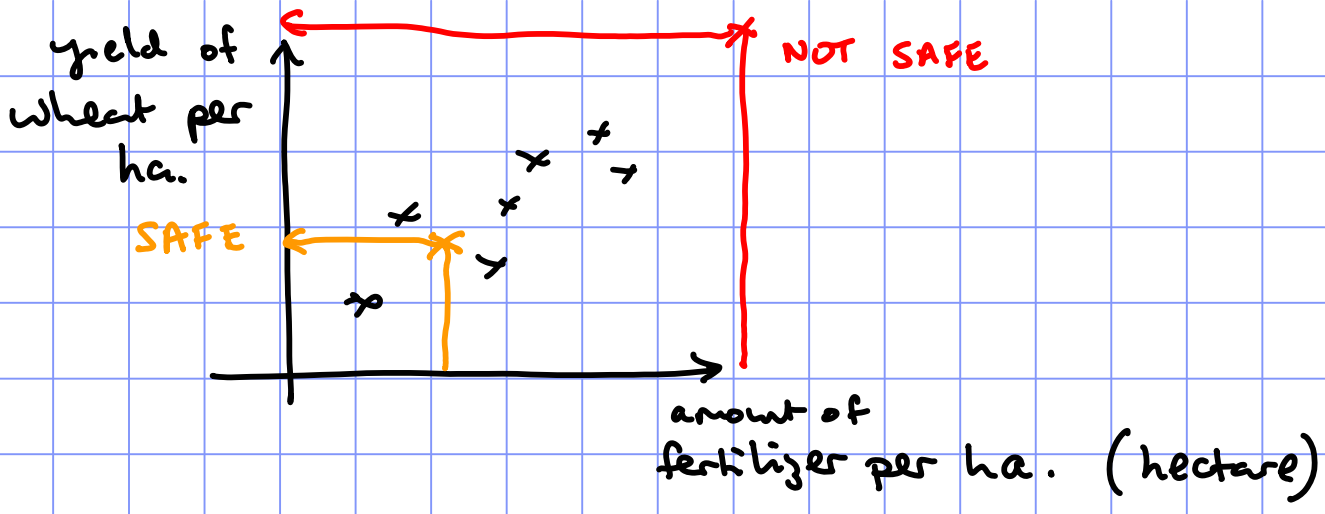


In this case we can say that the more TV licenses there are, the more juvenile there is. HOWEVER, we cannot conclude that watching more TV is causing more juvenile crime. We can say that it may be worth investigating this link.

If we find there is no correlation, we would be justified in saying that there is no causation.

③ We must be careful when extrapolating beyond the range of our data.

e.g an experiment with fertilizer gives the following data :-



We cannot assume that this trend will continue if we use larger and larger amounts of fertilizer. - see ×

[Interpolation (making estimates within the range of the given data) is however generally safe.] - see ×

Regression lines

If there is a correlation between our two variables, we may wish to use one variable (the EXPLANATORY or INDEPENDENT variable) to estimate or predict values of the other (the RESPONSE or DEPENDENT variable)

The EXPLANATORY variable is plotted on the HORIZONTAL axis, and is often (but not always) called x.
The RESPONSE variable (often called y) is plotted on the vertical axis.

We can draw a REGRESSION LINE (line of best fit) and use it to estimate values of the response variable. Or we may calculate them using the equation of the regression line.

A regression always passes through the point $(\bar{x}, \bar{y})$.
Its gradient is given by the formula $\frac{S_{xy}}{S_{xx}}$.

So the equation of the regression line is

$$y - \bar{y} = \frac{S_{xy}}{S_{xx}}(x - \bar{x})$$

We can rearrange this:

let $b = \frac{S_{xy}}{S_{xx}}$

So $y - \bar{y} = b(x - \bar{x})$

$y - \bar{y} = bx - b\bar{x}$

$y = (\bar{y} - b\bar{x}) + bx$

$y = a + bx$

where a (the y-intercept) is $\bar{y} - b\bar{x}$

and b (the gradient) is $\frac{S_{xy}}{S_{xx}}$

Example We may assume that a person's shoe size (y) depends on their height (x). Find, using the data collected previously, the regression line of y on x, and use it to estimate the shoe size of a girl (a) 155 cm tall (b) 185 cm tall

$$\text{gradient} = \frac{S_{xy}}{S_{xx}} = \frac{63}{466} = 0.135$$

$$\begin{aligned} y\text{-intercept} &= \bar{y} - b\bar{x} \\ &= \frac{87.5}{15} - 0.135 \times \frac{2490}{15} \\ &= -16.6 \end{aligned}$$

Equation of regression line is

$$y = 0.135x - 16.6$$

(a) If $x = 155$cm, $y = 4.35$ ($\approx 4\frac{1}{2}$ shoe size)

(b) If $x = 185$cm, $y = 8.375$ ($\approx 8\frac{1}{2}$ shoe size).

Comment on the reliability of these estimates.

(a) should be reliable as 155cm is within the range of the data collected.

(b) 185cm is outside the range of the data collected, so we are EXTRAPOLATING which may <u>not</u> be reliable as the observed trend

may not continue.

p139 Ex 7A Q 4, 7, 13, 14, 15, 9

Interpreting the gradient and y-intercept

The y-intercept gives an approximate value for "y" (the response variable) when "x" (the explanatory variable) is equal to 0.

This may or may not have any real meaning depending on the data we are using
- for example using the data above, it tells us that a person 0cm tall has a shoe size of -16.6!

The gradient gives an approximate value for the increase in "y" for every increase of 1 in "x".
In our example we see that for every 1cm increase in height, shoe size increases by 0.135.

Coding Suppose we have two variable x and y. We code our data using formulae $p = \frac{x-100}{5}$

$$q = \frac{y-50}{2}.$$

We then find the regression line of p on q to be

$$p = 15 + 2.5q$$

To encode this back into terms of x and y, simply substitute in the formulae for p and q and then rearrange

$$\frac{x-100}{5} = 15 + 2.5\left(\frac{y-50}{2}\right)$$

$$\frac{x-100}{5} = 15 + 1.25y - 62.5$$

$$x - 100 = 75 + 6.25y - 312.5$$

$$\underline{\underline{x = 6.25y - 137.5}}$$