Correlation (12 pages; 28/6/21)

(1) Formula for r

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

where $S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y})$,

$S_{xx} = \sum(x_i - \overline{x})^2$ and $S_{yy} = \sum(y_i - \overline{y})^2$

Notes

(i) The use of Pearson's product moment correlation coefficient (PMCC) (ρ for the population, and r for a sample) assumes that the data are drawn at random from a bivariate Normal distribution. This means that, for any value of one variable, the distribution of the values of the other is Normal.

If a bivariate Normal distribution applies (and there is correlation), then the scatter diagram is expected to show an elliptical pattern, with a concentration of data at the centre. The projection of the data on either of the x- or y- axes should exhibit a normal distribution.

Note that both variables have to be random. If a scatter diagram is plotted for 'random on non-random' data (eg for a scientific experiment, such as "extension for given load"), then a PMCC would not be appropriate.

(ii) Alternatively, $S_{xx} = \sum {x_i}^2 - n\bar{x}^2$ (and similarly for S_{yy});

$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$

(iii) If all points of the scatter diagram lie on a horizontal line, then r is undefined (as both the numerator and denominator of $\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ are zero, since all $y_i - \overline{y}$ are zero).

(iv) $\frac{S_{xy}}{n}$ is the sample covariance , so that $\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{Cov(x,y)}{\sqrt{s_x^2 s_y^2}}$,where $s_x^2 = \frac{S_{xx}}{n}$, and similarly for y.

(v) Values for r

$-1 \leq r \leq 1$

$0 - 0.2$ very weak

$0.2 - 0.5$ weak

$0.5 - 0.7$ moderate

$0.7 - 0.8$ fairly strong

$0.8 - 0.9$ strong

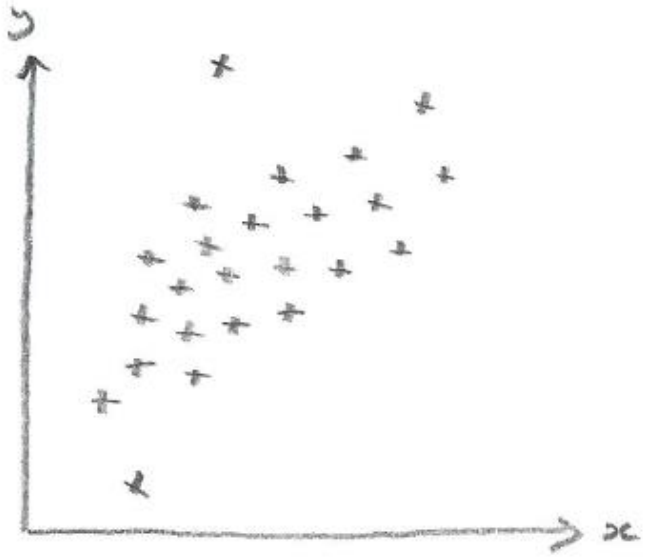
$0.9 - 1$ very strong

[These descriptions are fairly subjective.]

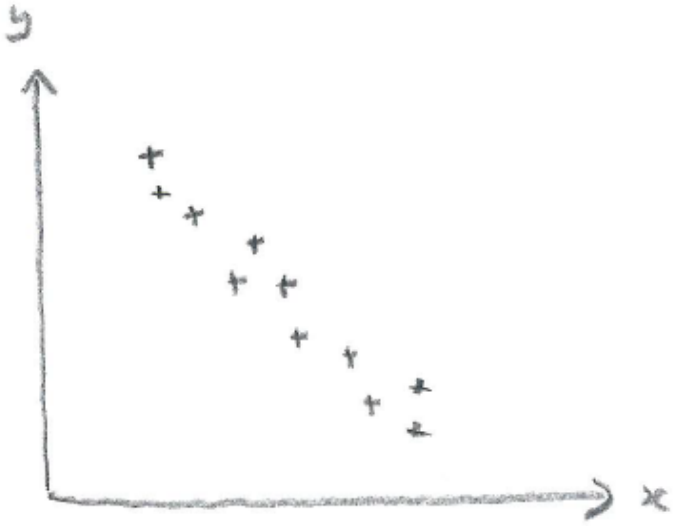
Note however that, in a hypothesis test for correlation, a low sample value for r could still provide sufficient evidence for correlation, if the sample is large enough.

(2) Example scatter diagrams

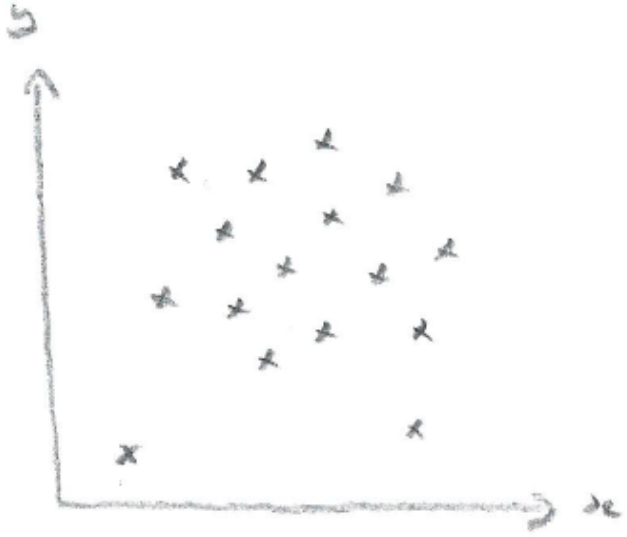
(A) Fairly strong positive correlation (with two outliers)



(B) Strong negative correlation



(C) No correlation

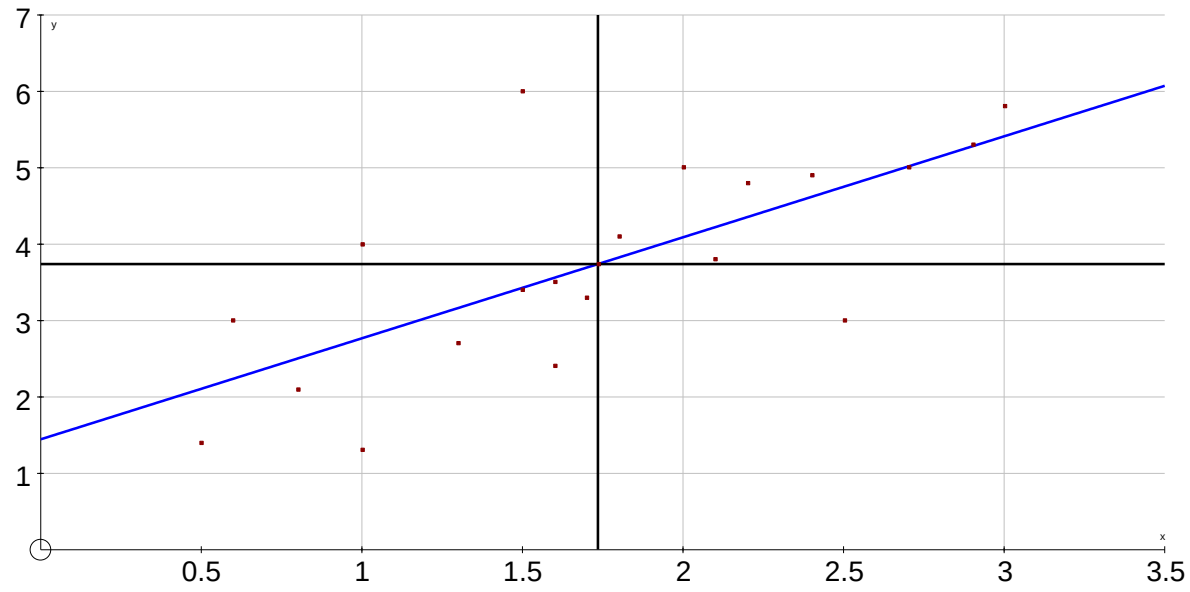


(3) Justification for the formula $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$,

where $S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y})$, $S_{xx} = \sum(x_i - \overline{x})^2$

and $S_{yy} = \sum(y_i - \overline{y})^2$

First of all, we can set up a new origin at $(\bar{x}, \bar{y})$. The point (x_i , y_i) then becomes $(x_i - \bar{x} , y_i - \bar{y})$ relative to this origin.



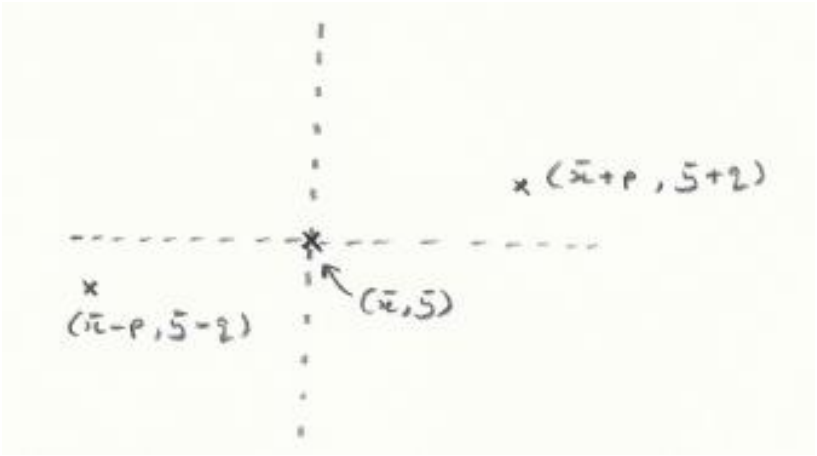
In the 1st and 3rd quadrants, $(x_i - \overline{x})(y_i - \overline{y})$ is positive, whilst it is negative in the 2nd and 4th quadrants.

Thus scatter diagrams where the points lie close to a line of best fit of positive gradient will have a high proportion of positive contributions to S_{xy} , leading to a large positive value of r.

Similarly, if the points tend to lie close to a line of best fit of negative gradient then there will be a high proportion of negative contributions to S_{xy}, leading to a large negative value of r.

The denominator has the effect of scaling down the value, to produce a value of r between -1 and 1.

It might be thought that the value of r should depend on the slope of the line of best fit, but this isn't in fact the case: for example, r can equal 1 whatever the gradient (provided it is positive). This can be seen from the following example, involving just 3 points.



$S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) = (-p)(-q) + 0 + pq = 2pq$

$S_{xx} = \sum(x_i - \overline{x})^2 = (-p)^2 + 0 + p^2 = 2p^2$

$S_{yy} = \sum(y_i - \overline{y})^2 = (-q)^2 + 0 + q^2 = 2q^2$

so that $r = \frac{2pq}{\sqrt{(2p^2)(2q^2)}} = 1$

Also, if the points are $(\overline{x} - p, \overline{y} + q), (\overline{x}, \overline{y})$ & $(\overline{x} + p, \overline{y} - q)$ instead, then

$S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) = (-p)q + 0 + p(-q) = -2pq$

$S_{xx} = \sum(x_i - \overline{x})^2 = (-p)^2 + 0 + p^2 = 2p^2$

$S_{yy} = \sum(y_i - \overline{y})^2 = (q)^2 + 0 + (-q)^2 = 2q^2$

so that $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-2pq}{\sqrt{(2p^2)(2q^2)}} = -1$

(Note though that the sign of r depends on whether the gradient is positive or negative.)

(4) Issues

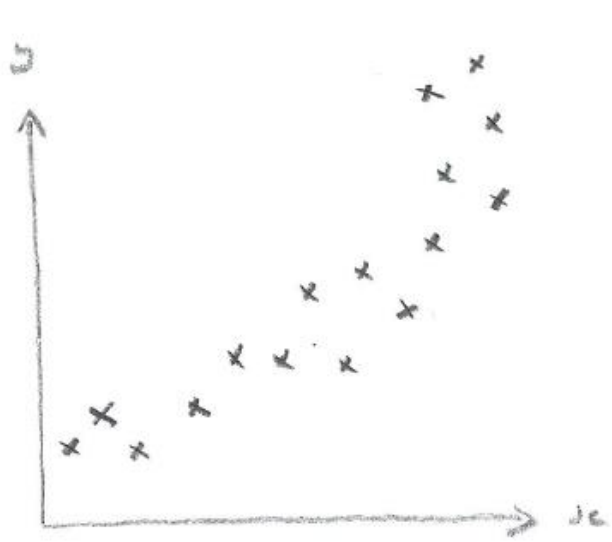
(4.1) Cause and effect

Correlation tests cannot establish a causal link between two variables; only that there is an ‘association’ (or connection).

For example, it can probably be shown that there is a correlation between consumption of champagne and long life expectancy. This doesn’t imply that champagne improves life expectancy (in fact the reverse may be true): the result has probably arisen because wealthy people (who can afford lots of champagne) are more likely to be in good health than poorer people.

As another example, it may be shown that there is a negative correlation between the time a child spends on social media and their reading ability. This might suggest the conclusion that a child's reading ability suffers when they spend too long on social media. An alternative explanation is that children of greater reading ability may tend to have more self-discipline. Or that well-educated parents (who encourage their children to read) may limit the time spent on social media.

(4.2) The PMCC provides a test of linear correlation only - ie there may still be a non-linear pattern, as in the examples below.

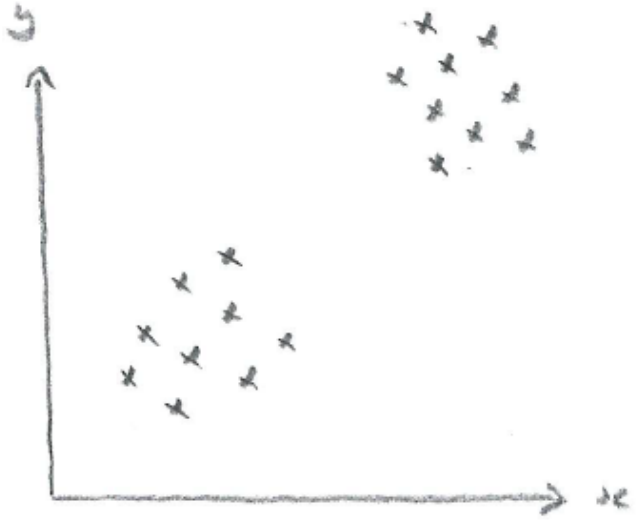


(a)

(b)

(4.3) Misleading data

The data in a scatter diagram may be so grouped as to give rise to an unreasonably high (absolute) value of r.

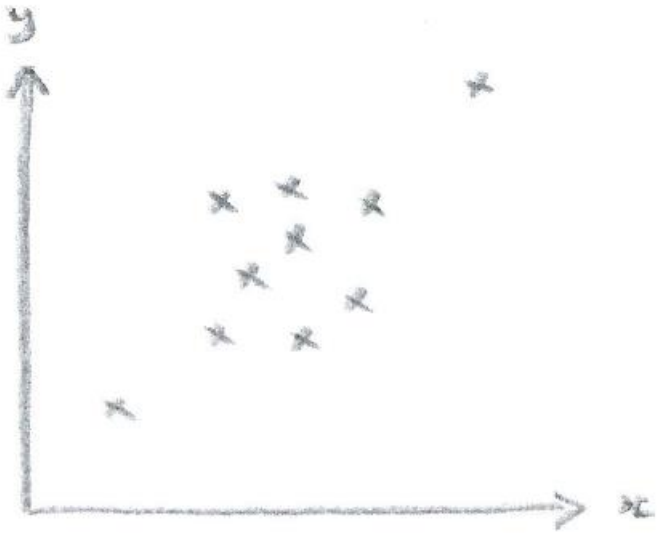


For example, in the diagram above there would appear to be two distinct regions of the data, neither of which exhibits any correlation.

In general, a clear lack of correlation in a scatter diagram will override a high value of r.

(4.4) Outliers

It may be appropriate to exclude outliers; either because data errors are suspected, or because the items in question are thought not to be typical of the population.



Sometimes outliers may give a false impression of a linear pattern, as in the diagram above. Generally though, one or two such points shouldn't have too much effect on the value of r.

(5) Hypothesis testing

Layout of test:

H_0: $\rho = 0$ (where ρ is the population correlation coefficient)

H_1: $\rho > 0$ (one-tailed)

eg 5% significance level

test statistic: r (sample pmcc)

critical value from pmcc table (based on sample size)

Give two versions of the conclusion:

(a) in technical language

(b) in layman's language (avoid use of the word "correlation"); referring to the context

Give a 'non-assertive conclusion'; eg "there is insufficient evidence of an association", rather than "there is no association"

Notes

(i) As for any hypothesis test, fresh data should be used (rather than data that prompted the test).

(ii) The data have to be random, and come from a bivariate normal distribution (exhibiting an elliptical pattern), in order for the test to be valid.

(iii) If the sample is large enough, H_0 may be rejected, even though r is small. In this case, the size of the correlation (the 'effect size') will clearly be important.

(iv) Critical Values for r

	5%	2½%	1%	½%
	10%	5%	2%	1%
n				
1	–	–	–	–
2	–	–	–	–
3	0.9877	0.9969	0.9995	0.9999
4	0.9000	0.9500	0.9800	0.9900
5	0.8054	0.8783	0.9343	0.9587
6	0.7293	0.8114	0.8822	0.9172
7	0.6694	0.7545	0.8329	0.8745
8	0.6215	0.7067	0.7887	0.8343
9	0.5822	0.6664	0.7498	0.7977
10	0.5494	0.6319	0.7155	0.7646
11	0.5214	0.6021	0.6851	0.7348
12	0.4973	0.5760	0.6581	0.7079
13	0.4762	0.5529	0.6339	0.6835
14	0.4575	0.5324	0.6120	0.6614
15	0.4409	0.5140	0.5923	0.6411
16	0.4259	0.4973	0.5742	0.6226
17	0.4124	0.4821	0.5577	0.6055
18	0.4000	0.4683	0.5425	0.5897
19	0.3887	0.4555	0.5285	0.5751
20	0.3783	0.4438	0.5155	0.5614

(6) Linear Coding

To establish the effect on r of the transformation $y = ax + b$:

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\sum(ax_i+b)}{n} = \frac{(a\sum x_i)+nb}{n} = a\bar{x} + b$$

$$S_{yy} = (\sum {y_i}^2) - n\bar{y}^2 = (\sum(ax_i + b)^2) - n(a\bar{x} + b)^2$$

$$= (a^2 \sum {x_i}^2) + (2ab \sum x_i) + nb^2 - na^2\bar{x}^2 - 2nab\bar{x} - nb^2$$

$$= a^2\{(\sum {x_i}^2) - n\bar{x}^2\}, \text{ since } \sum x_i = n\bar{x}$$

$= a^2 S_{xx}$

Suppose that we are considering the correlation between x_i and z_i.

Then $r_{xz} = \frac{S_{xz}}{\sqrt{S_{xx}S_{zz}}}$ and $r_{yz} = \frac{S_{yz}}{\sqrt{S_{yy}S_{zz}}}$

Now $S_{yz} = (\sum y_i z_i) - n\bar{y}\bar{z}$

$= (\sum(ax_i + b)z_i) - n(a\bar{x} + b)\bar{z}$

$= (a\sum x_i z_i) + (b\sum z_i) - na\bar{x}\bar{z} - nb\bar{z}$

$= a\{(\sum x_i z_i) - n\bar{x}\bar{z}\}$, since $\sum z_i = n\bar{z}$

$= aS_{xz}$

Hence $r_{yz} = \frac{aS_{xz}}{\sqrt{a^2 S_{xx}S_{zz}}} = \frac{S_{xz}}{\sqrt{S_{xx}S_{zz}}} = r_{xz}$

ie the correlation coefficient is unaffected by a linear transformation of one (or both) of the variables.

As an example, clearly the correlation coefficient shouldn't change when the temperature scale is taken to be Centigrade instead of Fahrenheit.