26 Correlation

Introduction

Correlation analysis is concerned with measuring the degree of association between two variables, x and y. Initially, we assume that both x and y are **numerical**, e.g. height and weight.

Suppose we have a pair of values, (x, y), measured on each of the n individuals in our sample. We can mark the point corresponding to each individual's pair of values on a two-dimensional **scatter diagram** (Chapter 4). Conventionally, we put the x variable on the horizontal axis, and the y variable on the vertical axis in this diagram. Plotting the points for all n individuals, we obtain a scatter of points that may suggest a relationship between the two variables.

Pearson correlation coefficient

We say that we have a **linear relationship** between x and y if a straight line drawn through the midst of the points provides the most appropriate approximation to the observed relationship. We measure how close the observations are to the straight line that best describes their linear relationship by calculating the **Pearson product moment correlation coefficient**, usually simply called the **correlation coefficient**. Its true value in the *population*, ρ (the Greek letter, rho), is estimated in the *sample* by r, where

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

which is usually obtained from computer output.

Properties

- r ranges from -1 to $+1$.
- Its **sign** indicates whether one variable increases as the other variable increases (positive r) or whether one variable decreases as the other increases (negative r) (see Fig. 26.1).
- Its **magnitude** indicates how close the points are to the straight line. In particular if $r = +1$ or -1, then there is perfect correlation with all the points lying on the line (this is most unusual, in practice); if $r = 0$, then there is no **linear** correlation (although there may be a non-linear relationship). The closer r is to the extremes, the greater the degree of linear association (Fig. 26.1).
- It is dimensionless, i.e. it has no units of measurement.
- Its value is valid only within the range of values of x and y in the sample. Its absolute value (ignoring sign) tends to increase as the range of values of x and/or y increases and therefore you cannot infer that it will have the same value when considering values of x or y that are more extreme than the sample values.
- x and y can be interchanged without affecting the value of r.
- A correlation between x and y does not necessarily imply a 'cause and effect' relationship.
- r^2 represents the proportion of the variability of y that can be attributed to its linear relationship with x (Chapter 28).

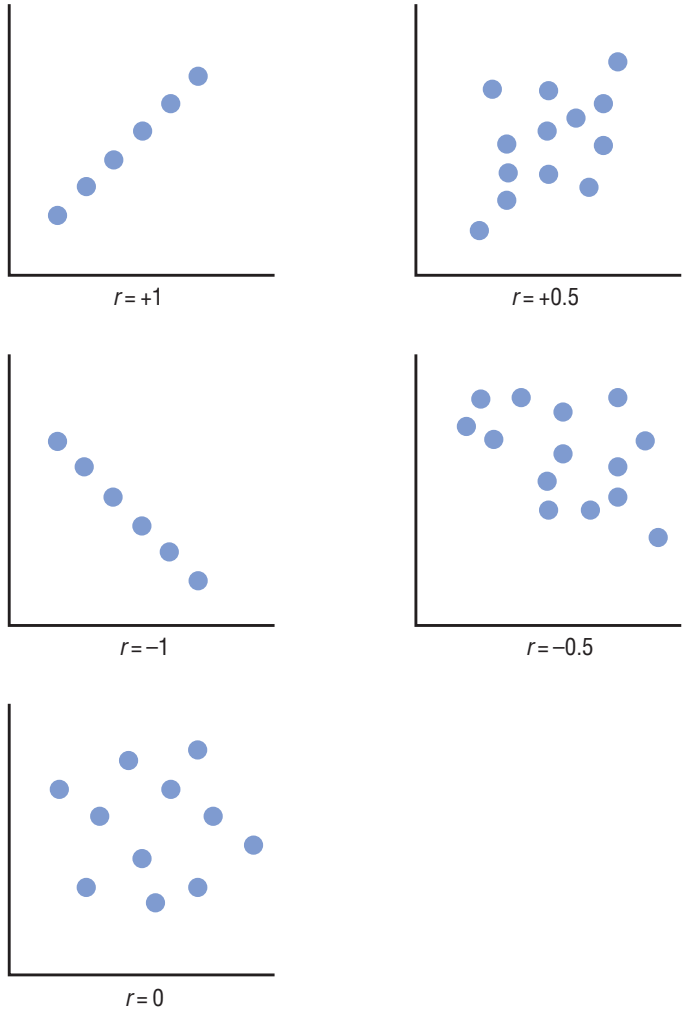

Figure 26.1 Five diagrams indicating values of r in different situations.

When not to calculate *r*

It may be misleading to calculate *r* when:

- there is a non-linear relationship between the two variables (Fig. 26.2a), e.g. a quadratic relationship (Chapter 33);
- the data include more than one observation on each individual;
- one or more outliers are present (Fig. 26.2b);
- the data comprise subgroups of individuals for which the mean levels of the observations on at least one of the variables are different (Fig. 26.2c);

Hypothesis test for the Pearson correlation coefficient

We want to know if there is any linear correlation between two numerical variables. Our sample consists of n independent pairs of values of x and y. We assume that at least one of the two variables is Normally distributed.

1 Define the null and alternative hypotheses under study

H_0: $\rho = 0$

H_1: $\rho \neq 0$

2 Collect relevant data from a sample of individuals

3 Calculate the value of the test statistic specific to H_0

Calculate r.

- If $n \leq 150$, r is the test statistic
- If $n > 150$, calculate $T = \sqrt{\frac{(n-2)}{(1-r^2)}}$

which follows a t-distribution with $n-2$ degrees of freedom.

4 Compare the value of the test statistic to values from a known probability distribution

- If $n \leq 150$, refer r to Appendix A10
- If $n > 150$, refer T to Appendix A2.

5 Interpret the *P*-value and results

Calculate a confidence interval for ρ. Provided *both variables are approximately Normally distributed*, the approximate 95% confidence interval for ρ is:

$$\left(\frac{e^{2z_1}-1}{e^{2z_1}+1}, \frac{e^{2z_2}-1}{e^{2z_2}+1}\right)$$

where $z_1 = z - \frac{1.96}{\sqrt{n-3}}$, $z_2 = z + \frac{1.96}{\sqrt{n-3}}$,

and $z = 0.5\log_e\left[\frac{(1+r)}{(1-r)}\right]$.

Note that, if the sample size is large, H_0 may be rejected even if r is quite close to zero. Alternatively, even if r is large, H_0 may not be rejected if the sample size is small. For this reason, it is particularly helpful to calculate r^2, the proportion of the total variance of one variable explained by its linear relationship with the other. For example, if $r = 0.40$ then $P < 0.05$ for a sample size of 25, but the relationship is only explaining 16% ($= 0.40^2 \times 100$) of the variability of one variable.

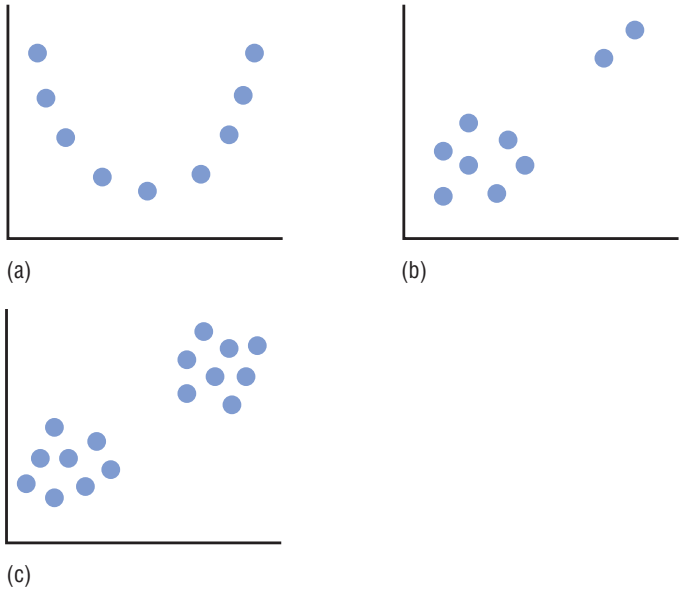

Figure 26.2 Diagrams showing when it is inappropriate to calculate the correlation coefficient. (a) Relationship not linear, $r = 0$. (b) In the presence of outlier(s). (c) Data comprise subgroups.

Spearman's rank correlation coefficient

We calculate **Spearman's rank correlation coefficient**, the non-parametric equivalent to Pearson's correlation coefficient, if one or more of the following points is true:

- at least one of the variables, x or y, is measured on an ordinal scale;
- neither x nor y is Normally distributed;
- the sample size is small;
- we require a measure of the association between two variables when their relationship is non-linear.

Calculation

To estimate the population value of Spearman's rank correlation coefficient, ρ_s, by its sample value, r_s:

1 Arrange the values of x in increasing order, starting with the smallest value, and assign successive *ranks* (the numbers 1, 2, 3, . . . , n) to them. Tied values receive the mean of the ranks these values would have received had there been no ties.

2 Assign ranks to the values of y in a similar manner.

3 r_s is the Pearson correlation coefficient between the *ranks* of x and y.

Properties and hypothesis tests

These are the same as for Pearson's correlation coefficient, replacing r by r_s, except that:

- r_s provides a measure of association (not necessarily linear) between x and y;
- when testing the null hypothesis that $\rho_s = 0$, refer to Appendix A11 if the sample size is less than or equal to 10;
- we do not calculate r_s^2 (it does not represent the proportion of the total variation in one variable that can be attributed to its linear relationship with the other).

Example

As part of a study to investigate the factors associated with changes in blood pressure in children, information was collected on demographic and lifestyle factors, and clinical and anthropometric measures in 4245 children aged from 5 to 7 years. The relationship between height (cm) and systolic blood pressure (mmHg) in a sample of 100 of these children is shown in the scatter diagram (Fig. 28.1); there is a tendency for taller children in the sample to have higher blood pressures. **Pearson's correlation coefficient** between these two variables was investigated. Appendix C contains a computer output from the analysis.

1 H_0: the population value of the Pearson correlation coefficient, ρ, is zero

H_1: the population value of the Pearson correlation coefficient is not zero.

2 We can show (Fig. 37.1) that the sample values of both height and systolic blood pressure are approximately Normally distributed.

3 We calculate r as 0.33. This is the test statistic since $n \leq 150$.

4 We refer r to Appendix A10 with a sample size of 100: $P < 0.001$.

5 There is strong evidence to reject the null hypothesis; we conclude that there is a linear relationship between systolic blood pressure and height in the population of such children. However, $r^2 = 0.33 \times 0.33 = 0.11$. Therefore, despite the highly significant result, the relationship between height and systolic blood pressure explains only a small percentage, 11%, of the variation in systolic blood pressure.

In order to determine the 95% confidence interval for the true correlation coefficient, we calculate:

$$z = 0.5\ln\left(\frac{1.33}{0.67}\right) = 0.3428$$

$$z_1 = 0.3428 - \frac{1.96}{9.849} = 0.1438$$

$$z_2 = 0.3428 + \frac{1.96}{9.849} = 0.5418$$

Thus the confidence interval ranges from

$$\frac{(e^{2\times 0.1438} - 1)}{(e^{2\times 0.1438} + 1)} \text{ to } \frac{(e^{2\times 0.5418} - 1)}{(e^{2\times 0.5418} + 1)}, \text{ i.e. from } \frac{0.33}{2.33} \text{ to } \frac{1.96}{3.96}.$$

We are thus 95% certain that ρ lies between 0.14 and 0.49.

As we might expect, given that each variable is Normally distributed, **Spearman's rank correlation coefficient** between these variables gave a comparable estimate of 0.32. To test H_0: $\rho_s = 0$, we refer this value to Appendix A10 and again find $P < 0.001$.

Data kindly provided by Ms O. Papacosta and Dr P. Whincup, Department of Primary Care and Population Sciences, Royal Free and University College Medical School, London, UK.