21 Statistics

STARTING POINTS

- In scientific investigations, observations are recorded as **data** (singular, **datum**). There are different forms of data, but they are all potentially useful. No one type of data that is relevant to an enquiry is necessarily superior to any other type, provided the data have been accurately made and recorded.
- We can define the types of data we collect as:
  qualitative (or descriptive) observations such as, in behaviour studies, the feeding mechanism of honey bees visiting flowers (Figure 17.13, page 523) or nesting behaviour of a species of bird. Qualitative data may be recorded in written observations or notes, or by photography or drawings
  quantitative (or numerical) observations such as the size (numbers, length, breadth or area) of an organism, or of organs such as the leaves of a plant in shaded and exposed positions, or the pH values of soil samples in different positions (Figure 19.20, page 621).
- Quantitative data may be discrete or continuous:
  discrete data are whole numbers, such as the number of eggs laid in a nest – no nest ever contains 2.5 eggs!
  continuous data can be any values within some broad limit, such as the heights of the individuals of a population.
- In this chapter, a means of representing the **variability** of graphical data is illustrated, and statistical tests are examined. These concern the calculation of **means** and of **standard deviation**, and discussion of their usefulness, followed by application of the ***t*-test**.

Recording variability of data – error bars 1.1.1

In experimental science, the outcomes of investigations are checked to confirm they are reproducible. So, for example, when a leading laboratory makes an important discovery and publishes results in a paper, details of the experimental methods are given so that others may repeat the work. Incidentally, if other laboratories fail to confirm the results, then a controversy breaks out. The results are not accepted. Subsequent investigations on both sides of the ensuing argument eventually lead to a resolution of the difference.

Most often, initial results are confirmed because investigations are repeated several times, at the outset, before results are published.

In particular experiments that are part of your course, time may sometimes be too limited for you to repeat readings as you might wish. However, groups of fellow students carrying out the same experiment may be able to pool results. If so, then you may be able to see how variable or consistent a particular result is. So when you display data as part of your record of an investigation, using a graph for example, you can record the degree of variability in readings that the student group obtained in total. To do this you use **error bars**.

An example of error bars in use

A preliminary investigation of the effect of aerobic and anaerobic pre-treatment of tissue discs on their subsequent gain in mass is shown in Figure 21.1. Thin discs of plant tissue are often used because this technique allows all the cells in a sample to receive more or less identical conditions. (You can see the use of leaf tissue discs in an experiment in Figure 15.5, page 451.)

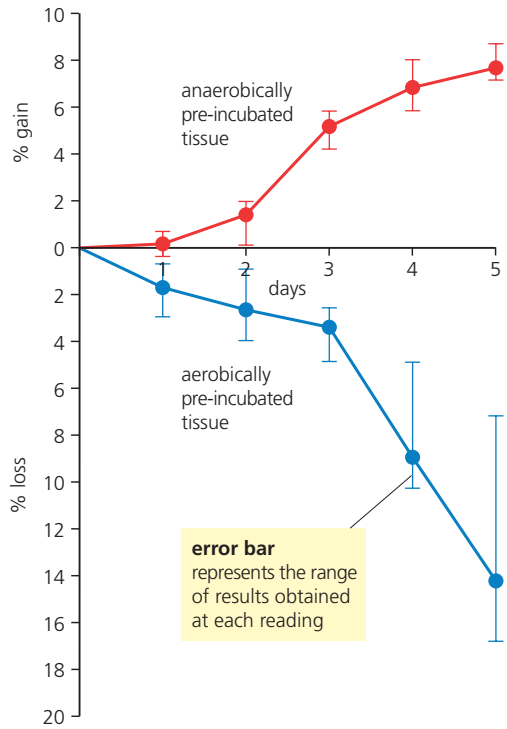


Figure 21.1 Change in fresh mass of tissue discs after aerobic and anaerobic pre-treatment

The results of this enquiry indicate that anaerobic pre-treatment leads to subsequent gain in mass, whereas aerobic pre-treatment leads to loss in mass. This experiment was based on five batches of ten discs for both treatments, and the variability of the results is recorded in error bars. Each error bar indicates the range of values (readings) from the highest to the lowest. In Figure 21.1, for example, the error bars draw attention to the much greater variation in results from aerobically pre-treated tissue.

Sometimes, the error bars shown in a graphical representation of data record variability as the standard deviation of a result. The calculation of standard deviations is discussed shortly.

Summarising data – the mean

Points on the curves in the graph in Figure 21.1 are based on five tissue batches (each of ten tissue discs) for each treatment. The value of each is the average or **arithmetic mean value** of individual batches. The value of means is that they convey the 'middleness' of the data.

How are the averages or means of the readings calculated?

To calculate a mean value, all the values from a particular treatment or observation are summed, and the total divided by the number of these values.

The formula for the arithmetic mean is:

$$\bar{x} = \frac{\Sigma x}{n}$$

where
$\bar{x}$ = arithmetic mean
Σx = sum of all the measurements
n = the total number of measurements

The value of means in other experimental situations

For example, let us imagine you have completed an investigation on the effects of the application of pesticide on the numbers of a common species of soil organism.

The outcome is that your field notebook now contains a large number of counts from randomly placed quadrats (page 601), some from treated soils, and some from untreated soil (see table in Figure 21.2).

In Figure 21.2, the data are also presented graphically, with the number of worms per quadrat on the *x*-axis, and the frequency of quadrats with each number of worms on the *y*-axis.

Look at the spread of data from both treatments.

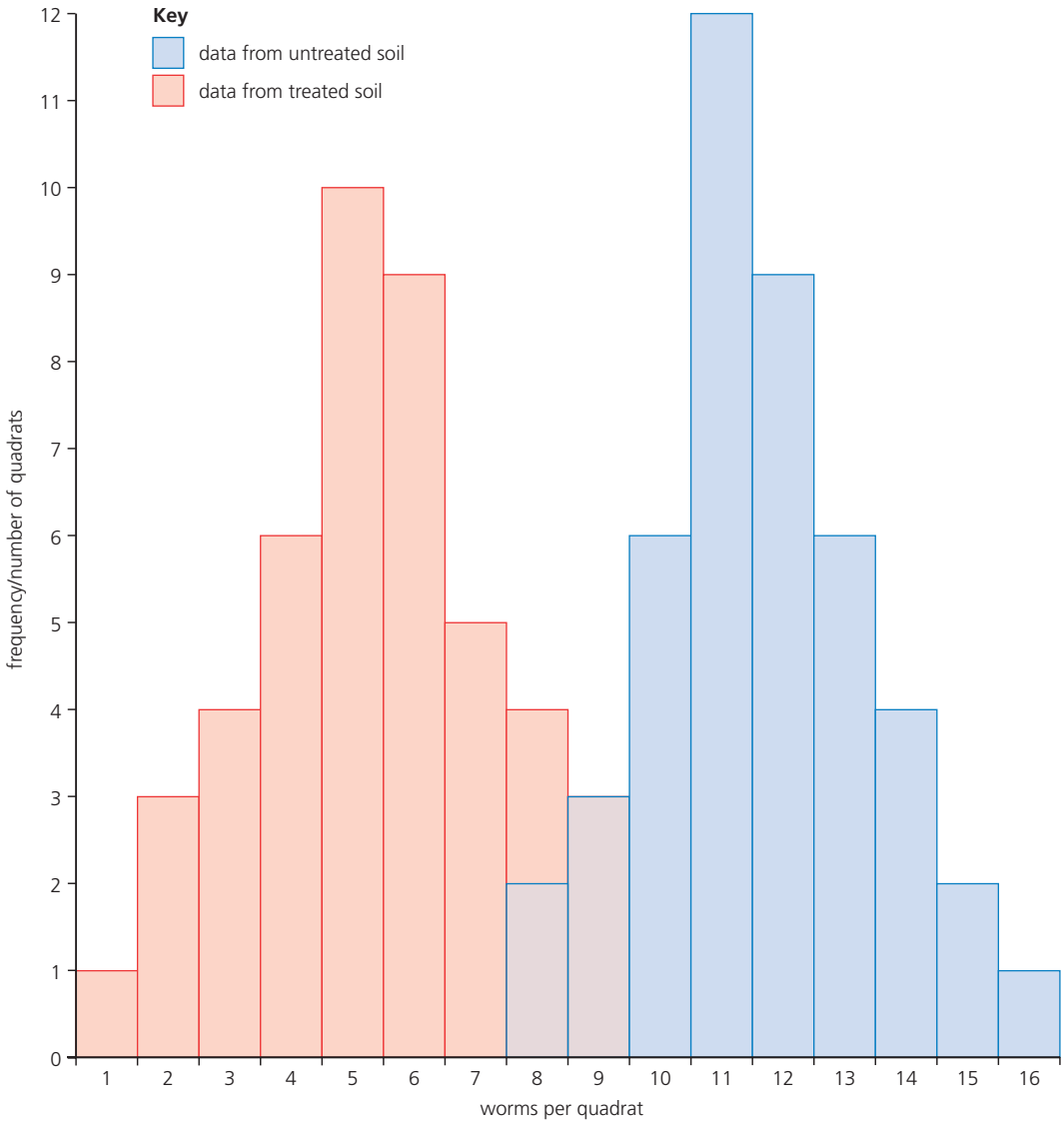
With these data presented as a graph, two characteristic bell-shaped curves result (which only slightly overlap). These are referred to as **normal distributions**. These arise, given a large enough number of observations or measurements, if the data may have an exactly symmetrical spread. However, data rarely exactly conform completely to a bell-shaped curve – an approximation is usual, given a finite number of observations. This is what we see in Figure 21.2.

Figure 21.2 An investigation of the effects on soil worm populations of pesticide treatment

experimental results

Quadrats on soil treated with pesticide			Quadrats on untreated soils		
Worms per quadrat	Frequency	Total	Worms per quadrat	Frequency	Total
0		**0**	7		**0**
1	|	**1**	8	||	**2**
2	|||	**3**	9	|||	**3**
3	||||	**4**	10	𝍸 |	**6**
4	𝍸 |	**6**	11	𝍸 𝍸 ||	**12**
5	𝍸 𝍸	**10**	12	𝍸 ||||	**9**
6	𝍸 ||||	**9**	13	𝍸 |	**6**
7	𝍸	**5**	14	||||	**4**
8	||||	**4**	15	||	**2**
9	|||	**3**	16	|	**1**
10		**0**	17		**0**

graph of frequency against numbers of earthworms per quadrat



How can the data best be summarised so that a comparison of the effect of the alternative treatments can be made?

The answer is in different ways, one of which is to find the **mean value** of soil worms for each treatment. To do this, all the values from quadrats on treated soil are summed, and the total divided by the number of values. In this way, we have an average value for the effect of this soil treatment. The mean will convey the 'middleness' of the data. The counts from quadrats on untreated soil are treated in the same way too, of course. *You could calculate the means for the data in Figure 21.2, for both treated soil and untreated soil.*

Extension: Mean, median and mode

In the handling of experimental data, you may come across two other terms, namely **mode** and **median**. These are not alternatives to the mean value; rather they have different meanings:

- **mode** is the most frequent value in a set of values;
- **median** is the middle value in a set of values arranged in ascending order.

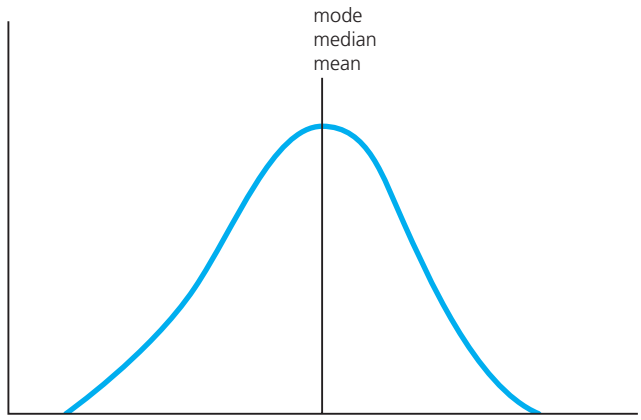
The graphs in Figure 21.3 illustrate how mean, mode and median relate in a **normal distribution** and in **skewed data**. You can see that it is in skewed data that they have particular significance.

Figure 21.3 Frequency distributions of symmetrical and skewed data

Normal distribution curve

Most biological data shows variability, but with values grouped symmetrically around a central value.

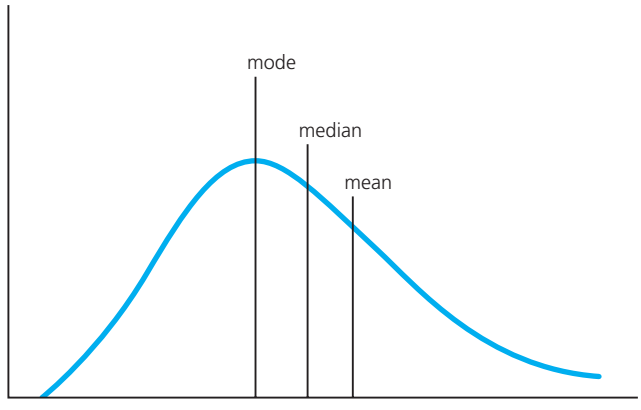
Here the mode, median and mean coincide.



Skewed distribution

Values reduce in frequency more rapidly on one side of the most frequently obtained value than the other.

Here the difference between the mean and mode is a measurement of 'skewness' of the data.



Calculating standard deviations

1.1.2–1.1.4

The **standard deviation** (**SD**, **s** or σ) of the mean tells us how spread out are the readings (the 'spreadoutness' of the data).

A small standard deviation indicates that the data is clustered closely around the mean value. A large standard deviation indicates a wider spread around the mean.

So, standard deviations are a measure of the variation in the data from the mean value of a set of values.

The five steps to calculating the standard deviation of a data set are:

1 Calculate the mean ($\bar{x}$)
2 Measure the deviations ($x - \bar{x}$)
3 Square the deviations $(x - \bar{x})^2$
4 Add the squared deviations $\sum(x - \bar{x})^2$
5 Divide by the number of samples (n).

Calculating standard deviations – an example

An ecologist investigated the reproductive capacity of two species of buttercup, *Ranunculus acris* (meadow buttercup) and *R. repens* (creeping buttercup).

Figure 21.4 Data on achene (fruit) production in two species of *Ranunculus*

Number of achenes	Frequency	
	R. repens	*R. acris*
15	1	0
16	1	0
17	1	0
18	2	1
19	4	1
20	4	1
21	8	1
22	7	1
23	9	3
24	10	4
25	16	4
26	9	5
27	10	5
28	4	6
29	5	8
30	3	14
31	1	12
32	1	10
33	2	7
34	1	3
35	1	2
36	0	3
37	0	2
38	0	3
39	0	2
40	0	2
41	0	0

The latter species spreads vegetatively via strong and persistent underground stems. Would this investment be reflected in a lowered production of fruit (the product of sexual reproduction) compared with fruit production by the meadow buttercup, which reproduces more or less exclusively by sexual reproduction?

Using comparable sized plants growing under similar conditions in the same soil, the numbers of achenes (fruits) formed in 100 flowers of each species were counted and recorded. The results are given in Figure 21.4, and calculations of the SDs are shown in Figure 21.5. Note that you are not expected to know the formula for calculating SD. The purpose of presenting the steps to the calculation is to take away the mystery of a calculation normally carried out by a scientific calculator or programmed spreadsheet.

Alternative methods of calculating means and SDs

Rather than carry out all these steps manually – for example, using a dedicated worksheet as in Figure 21.5 – the value of the standard deviation may be obtained using a scientific or **statistics calculator**, or by means of a **spreadsheet** incorporating formulae, or by using Merlin (page 683).

Using a scientific calculator or spreadsheet, you can calculate the SDs of both the data of frequencies of worms on quadrats on soil treated with pesticide and on untreated soils, at this point, if you wish.

Once obtained, the value may be applied to the normal distribution curve, as shown in Figure 21.6. Note that 68% of the data occurs within ± 1 SD, and more than 95% of the data occurs within ± 2 SDs.

Figure 21.5 Calculating the means and SDs of the data in Figure 21.4

achene production in *Ranunculus acris*

Values obtained in ascending order x	Frequency f	fx	Deviation of x from the mean $(x - \bar{x})$ [= d]	d^2	fd^2
17	0				
18	1	18	–12	144	144
19	1	19	–11	121	121
20	1	20	–10	100	100
21	1	21	–9	81	81
22	1	22	–8	64	64
23	3	69	–7	49	147
24	4	96	–6	36	144
25	4	100	–5	25	100
26	5	130	–4	16	80
27	5	135	–3	9	45
28	6	168	–2	4	24
29	8	232	–1	1	8
30	14	420	0	0	0
31	12	372	1	1	12
32	10	320	2	4	40
33	7	231	3	9	63
34	3	102	4	16	48
35	2	70	5	25	50
36	3	108	6	36	108
37	2	74	7	49	98
38	3	114	8	64	192
39	2	78	9	81	162
40	2	80	10	100	200
41	0				
	$\Sigma f = 100$	$\Sigma fx = 2999$			$\Sigma fd^2 = 2031$

Mean of data $= \frac{\Sigma fx}{\Sigma f} = \frac{2999}{100} = 29.99$

SD $= \sqrt{\left(\frac{\Sigma fd^2}{\Sigma f - 1}\right)} = \sqrt{\frac{2031}{99}} = \sqrt{20.51} = 4.53$

Thus the mean of the sample *Ranunculus acris* = 29.99, and the SD = 4.53.

achene production in *Ranunculus repens*

Values obtained in ascending order x	Frequency f	fx	Deviation of x from the mean $(x - \bar{x})$ [= d]	d^2	fd^2
14	0				
15	1	16	–9	81	81
16	1	17	–8	64	64
17	1	36	–7	49	98
18	2	76	–6	36	144
19	4	80	–5	25	100
20	4	168	–4	16	128
21	8	154	–3	9	63
22	7	207	–2	4	36
23	9	240	–1	1	10
24	10	400	0	0	0
25	16	234	1	1	9
26	9	270	2	4	40
27	10	112	3	9	36
28	4	145	4	16	80
29	5	90	5	25	75
30	3	31	6	36	36
31	1	32	7	49	49
32	1	66	8	64	128
33	2	34	9	81	81
34	1	35	10	100	100
35	1	36	11	121	121
36	0				
	$\Sigma f = 100$	$\Sigma fx = 2479$			$\Sigma fd^2 = 1479$

Mean of data $= \frac{\Sigma fx}{\Sigma f} = \frac{2479}{100} = 24.79$

SD $= \sqrt{\left(\frac{\Sigma fd^2}{\Sigma f - 1}\right)} = \sqrt{\frac{1479}{99}} = \sqrt{14.93} = 3.86$

Thus the mean of the sample *Ranunculus repens* = 24.79, and the SD = 3.86.

The value of calculating standard deviation

We have noted that a standard deviation of low value indicates that the observation differs very little from the mean, and that high values of SD indicate a wider spread around the mean.

Thus the SDs can be used to help to decide whether the differences between the two related means are significant or not, such as those shown in Figure 21.2 (page 677).

If the SDs are much larger than the difference between the means, then the differences in the means are highly unlikely to be significant.

On the other hand, when SDs are much smaller than the differences between the means, then the differences between the means is almost certainly significant.

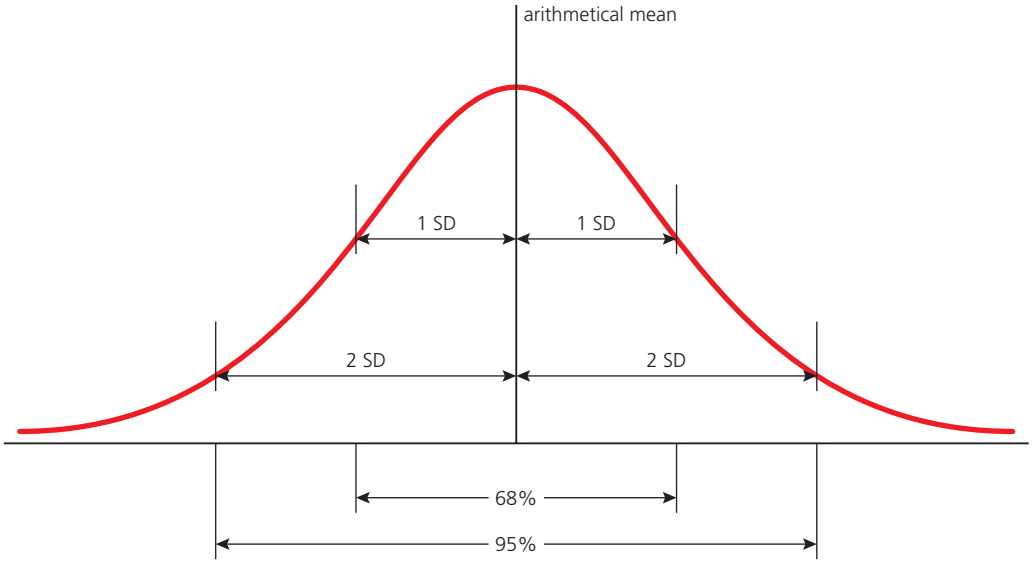


Figure 21.6 The normal distribution and its SD

Another statistical test – the *t*-test

1.1.5

Statistical tests typically compare large, randomly selected representative samples of normally distributed data. In practice, it is often the case that data can only be obtained from quite small samples. The ***t*-test** may be applied to sample sizes of more than 5 and less than 30 of normally distributed data. It provides a way of measuring the overlap between two sets of data – a large value of *t* indicates little overlap and makes it highly likely there is a significant difference between the two data sets. An example will illustrate the method. However, you should note that you are not expected to calculate values of *t*.

Applying the *t*-test

An ecologist was investigating woodland microhabitats, contrasting the communities in a shaded position with those in full sunlight. One of the plants was ivy (*Hedera helix*), but relatively few occurred at the locations under investigation. The issue arose: were the leaves in the shade actually larger than those in the sunlight?

Leaf widths were measured, but because the size of the leaves varied with the position on the plant, only the fourth leaf from each stem tip was measured. The results from the plants available are shown in Table 21.1.

Size-class/mm	Leaves in sunlight (a)	Leaves in shade (b)
20–24	24	
25–29	26, 26	26
30–34	30, 31, 31, 32, 32, 33	33, 34
35–39	37, 38	35, 35, 36, 36, 36, 37
40–44	43	41, 42
45–49		45

Table 21.1 Sizes of sun and shade leaves of *Hedera helix*

Steps to the *t*-test

1 The null hypothesis (negative hypothesis) assumes the difference under investigation has arisen by chance. In this example, the **null hypothesis** is:
'There is no difference in size between leaves in sunlight and leaves in shade.'
The role of the *t*-test is to determine whether to accept or reject the null hypothesis. If it is rejected here, we can have confidence that the difference in the leaf sizes of the two samples is statistically significant.

2 Next, check that the data are normally distributed. This is done by arranging the data for leaves in sunlight and leaves in shade as in Table 21.1 (and plot a histogram, if necessary).

3 *You are not expected to calculate values of t.* This statistic can be found by using a scientific or statistics calculator, or by means of a spreadsheet incorporating formulae.

Actually, a formula for the *t*-test for unmatched samples (data sets a and b) is:

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

where
$\bar{x}_a$ = mean of data set a
$\bar{x}_b$ = mean of data set b
s_a^2 = standard deviation for data set a, squared
s_b^2 = standard deviation for data set b, squared
n_a = number of data in set a
n_b = number of data in set b
$\sqrt{}$ = square root of

4 Once a value of *t* has been calculated (here $t = 2.10$), we determine the **degrees of freedom (df)** for the two samples, using the formula:
df = (total number of values in both samples) − 2
$= (n_a + n_b) - 2$
In this case:
df = (12 + 12) − 2 = 22.
Now we consult a table of critical values for the *t*-test.

5 A table of critical values for the *t*-test is given in Figure 21.7. Look down the column of significance levels (p) at the 0.05 level until you reach the line corresponding to df = 22. You will see that here, $p = 2.08$.

6 Since the calculated value of *t* (2.10) exceeds this critical value (2.08) at the 0.05 level of significance, it indicates that there is a lower than 0.05 probability (5%) that the difference between the two means is solely due to chance. Therefore, we can reject the null hypothesis, and conclude the difference between the two samples is significant.

For the experimenter, the significance of all this, is that there is a reason for the difference in the means, which can now be further investigated and fresh hypotheses proposed.

Extension: Merlin – statistical software available to biology students

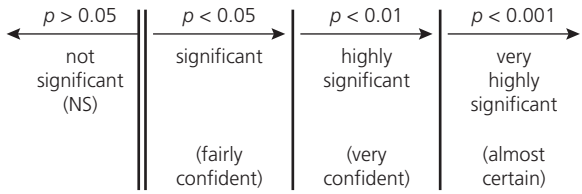
Merlin is a statistical package produced by Dr Neil Millar of Heckmondwike Grammar School, UK. This software is an add-in for Microsoft Excel and is easy to use. Merlin is available free of charge for educational and non-profit use. The package may be copied to laboratory and student computers from the school's website, currently at
www.heckgrammar.kirklees.sch.uk/index.php?p=310

URLs do change periodically. Merlin can also be located by means of a Google search under 'merlin+statistics'.

Once you have data from laboratory or field investigations, Merlin can be used to carry out a statistical test. There are no calculations required, and no look-up tables – no maths, no mistakes, in effect! Merlin also includes a basic introduction to statistics for biology students and a 'test chooser'. Here, in answer to a series of questions, Merlin selects the right test. Also with Merlin, Excel can display data in a range of graphs and charts, as appropriate.

Figure 21.7 Critical values for the *t*-test

	decreasing value of p →			
Degrees of freedom (df)	***p* values**			
	0.10	**0.05**	**0.01**	**0.001**
1	6.31	12.71	63.66	636.60
2	2.92	4.30	9.92	31.60
3	2.35	3.18	5.84	12.92
4	2.13	2.78	4.60	8.61
5	2.02	2.57	4.03	6.87
6	1.94	2.45	3.71	5.96
7	1.89	2.36	3.50	5.41
8	1.86	2.31	3.36	5.04
9	1.83	2.26	3.25	4.78
10	1.81	2.23	3.17	4.59
12	1.78	2.18	3.05	4.32
14	1.76	2.15	2.98	4.14
16	1.75	2.12	2.92	4.02
18	1.73	2.10	2.88	3.92
20	1.72	2.09	2.85	3.85
22	1.72	2.08	2.82	3.79
24	1.71	2.06	2.80	3.74
26	1.71	2.06	2.78	3.71
28	1.70	2.05	2.76	3.67
30	1.70	2.04	2.75	3.65
40	1.68	2.02	2.70	3.55
60	1.67	2.00	2.66	3.46
120	1.66	1.98	2.62	3.37
∞	1.64	1.96	2.58	3.29



Correlations do not establish causal relationships

1.1.6

By 'correlation' we mean 'a mutual relation between two (or more) things', or 'an interdependence of variable quantities'. The belief that because things have occurred together, one must be connected or related to the other in the sense that one is the cause of the other is an easily and commonly made mistake. Because two events (A and B) regularly occur together, it may appear to us that A causes B. This is not necessarily the case. In fact, there may be a common event that causes both, for example, or it may be an entirely spurious correlation. For example, some infants (very few) develop the symptoms of autism shortly after the normal time in childhood when the MMR inoculation is administered. Some parents of autistic children who had arranged for their child to be inoculated came to blame the vaccination for the child's condition. This confusion caught on for several years, many parents became anxious, and the practice of having the triple injection became unpopular. The numbers of vaccinated children fell to dangerously low levels. It was some time before detailed studies could convince the

majority of parents that the two events, MMR inoculation and the onset of autism, were not causally linked (page 360).

The fact that correlation does not prove cause was one of the reasons why Richard Doll's amassing of statistical evidence of a link between smoking and ill health was successfully resisted by the tobacco industry for an exceptionally long time (page 666). Now we know the various reasons why cigarette smoke triggers malfunctioning of body systems, ill heath and diseases of various sorts.

So, having applied statistical tests that indicate the possibility of a correlation, we cannot then assert that one event is the cause of the other. What we can do is have confidence that the events may well be linked, and so go on to investigate the mechanisms of the linkage – if there is one.

For example, a persistent condition of hypertension is directly linked to the raised incidence of coronary heart disease and vascular accidents of other sorts (page 665). In this case, the statistical relationship has been followed up, enabling us to understand why hypertension has these effects. Once the connections between events or conditions are understood, the relationship has been established. In other words, just because a correlation does *not* prove the cause does *not* mean there cannot be a causal relationship.

So, statistical confidence in the possibility of a causal link is a springboard to further investigation, not proof of a relationship.

TOK Link

Is the idea of 'cause and effect' generally uncritically accepted?

Does it permeate our culture?

Examine some current newspapers or journals that are frequently read in your country. Can you find examples of assumptions about cause and effect that are stated, but which are unproved and possibly dubious?