# 3 Choosing how to represent data

**Key words:** variable, unit, raw data, categorical, discrete, continuous, factor, frequency, frequency table, grouped data, two-way table, pie chart, bar chart, grouped bar chart, stacked bar chart, independent variable, dependent variable, data point, horizontal axis, vertical axis, line graph, gradient, time series, scatter graph.

Tables, charts and graphs are all ways of representing data, and they can be used for two broad purposes. The first is to support the collection, organisation and analysis of data as part of the process of a scientific study. The second is to help present the conclusions of a study to a wider audience. The choices of how to represent data are influenced by:

- the nature of the data
- the kinds of questions about the data that are of interest.

## 3.1 Using tables to collect and present data

When constructing a table of data, one consideration is what to put in the rows and what to put in the columns. A common form of data in science involves two related ***variables***, for example the temperature of a cooling object against time (Figure 3.1).

**Figure 3.1** Temperature of a cooling object

**(a)**

| Time (s) | Temperature (°C) |
|---|---|
| 0 | 59.5 |
| 30 | 54.3 |
| 60 | 51.2 |
| 90 | 48.4 |
| etc. | |

**(b)**

| Time (s) | 0 | 30 | 60 | 90 | etc. |
|---|---|---|---|---|---|
| Temperature (°C) | 59.5 | 54.3 | 51.2 | 48.4 | |

The reason that the table in Figure 3.1a works better for collecting data is because it can be easily extended downwards – it is not so easy to extend Figure 3.1b. However, it also has the advantage that, by aligning the values for each quantity vertically, it is easier for the eye to scan down and compare the sizes of values. This is harder to do when the eye has to scan across a horizontal arrangement of values. When making tables intended for *presenting* data, this is a particular consideration, and more complex tables may require careful thought.

The ***units*** of the values are included at the top of the column along with the variable name, so that the rest of the table just shows the *numbers*. Note that the units are enclosed in *brackets*,

for example 'Temperature (°C)'. This is the most usual convention in secondary school science and mathematics for the column headers in tables of data and for labelling axes on graphs. Another convention is to write 'Temperature in °C'. In scientific literature and in post-16 studies, a common convention is to use the '/' symbol (the *solidus*, or forward slash), for example 'Temperature/°C)'. This is intended to indicate that the values of temperature are divided by the unit °C to produce 'pure numbers'. This is a subtle idea, which is why brackets are more suitable for secondary science. In addition, pupils need to be familiar with the 'brackets' convention, since it is widely used on tables and charts intended for a general audience, including those produced by scientific organisations.

Since pupils may come across graphs using different conventions from a variety of sources (in books, on the internet and so on), teachers may wish their pupils to be familiar with all of them. In any case, it is important to check which convention they will meet in their examinations.

Care should be taken, however, if the pupils are not familiar with the use of negative indices in units, which is also usually not introduced until post-16 (see Section 2.5 *Index notation and powers* on page 19). Thus, while both 'velocity (m/s)' and 'velocity/m $s^{-1}$' are acceptable and considered to be correct, 'velocity/m/s' is ambiguous and confusing and it should thus be avoided.

## 3.2 Using tables to process data

Tables are also used to support the processing of ***raw data*** in various ways. One example is when further columns are added to a table to carry out calculations on existing columns. Figure 3.2 shows a table in which two of the columns (Mass and Volume) are used to collect measured values, while the final column (Density) contains calculated values.

**Figure 3.2** Calculating density from mass and volume

| Object | Mass | Volume | Density |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

Another example is when raw data from one table are *counted* to produce further tables showing the values of the counts. For example, in a survey of the pupils in a class, counting the raw data on eye colour (Figure 3.3a) produces a 'table of counts' (Figure 3.3b). Such a table is called a ***frequency table***. In mathematics, a ***frequency*** refers to the number obtained by counting objects or events. Thus, if there are 7 pupils with eye colour 'blue' then this category has a frequency of 7.

The column 'Eye colour' in Figure 3.3a contains ***categorical*** data. The raw data in this table also include 'shoe sizes', and these are ***discrete*** data. It would also be possible to count the number of pupils with each shoe size but there might be quite a large number of categories. Here, it may be more convenient to use fewer categories, by choosing some *groups of shoe sizes*, and to count the numbers in these. Figure 3.3c is also a frequency table but here is showing ***grouped data***.

It is also possible to create groups from ***continuous*** data; this is discussed in Section 6.4 *Displaying larger sets of values* on page 53.

**Figure 3.3** Survey of pupils in a class

**(a)** Table of raw data

| Pupil | Eye colour | Shoe size |
|---|---|---|
| 1 | brown | 1½ |
| 2 | blue | 5 |
| 3 | brown | 5½ |
| 4 | brown | 4 |
| etc. | | |

**(b)** Frequency table

| Eye colour | Number of pupils |
|---|---|
| Blue | |
| Brown | |
| Green | |

**(c)** Frequency table (grouped data)

| Shoe size | Number of pupils |
|---|---|
| 2½ or less | |
| 3 to 5 | |
| 5½ or more | |

**(d)** Two-way table

| Eye colour | Shoe size | | |
|---|---|---|---|
| | 2½ or less | 3 to 5 | 5½ or more |
| Blue | | | |
| Brown | | | |
| Green | | | |

The tables in Figure 3.3b and 3.3c each show the numbers of pupils categorised by one independent variable or ***factor*** (eye colour or shoe size).

The table in Figure 3.3d shows the numbers of pupils categorised by *both of these factors*. This is also a frequency table, and is called a ***two-way table***. Such tables are useful to see if two factors are related – for example, if there were a large number of pupils with green eyes and large shoe sizes, then this might suggest a relationship between the two factors (though perhaps unlikely in this example).

## 3.3 Presenting data visually

The most common types of charts and graphs for presenting data are ***pie charts***, ***bar charts***, ***line graphs*** and ***scatter graphs***. As with tables, visual displays of data can be useful both in the analysis of data and in the presentation of the results.

Displaying data visually can be particularly useful in *comparing the relative sizes of values* and in *looking for relationships between variables*. Visual displays are less useful in communicating actual values: people tend to focus on the patterns rather than the numbers. To emphasise actual values, a table is more effective.

Choosing what charts or graphs to draw is influenced by the nature of the data. The rest of this section will look at the different kinds of display that can be used to represent the following commonly found data structures:

- *A quantity categorised by one factor*
  (e.g. numbers of people in a sample categorised by eye colour)
- *A quantity categorised by two factors*
  (e.g. UK energy consumption categorised by type of fuel and year)
- *Two related quantities*
  (e.g. the extension of a spring related to the mass suspended from it).

Another type of data is simply a set of values for a single quantity (e.g. the heights of a sample of pupils). The analysis of this kind of data and the displays used (*histograms* and *boxplots*) are discussed later. (See Chapter 6 *Dealing with variability* on page 50 and Chapter 8 *Looking for relationships: batches and scatter graphs* on page 75.)

## 3.4 Charts showing a quantity categorised by one factor

Figure 3.4a shows a quantity (number of people in a sample) categorised by one ***factor*** (eye colour). Because it is meaningful to add these values to give a *total*, one possible display is a ***pie chart*** (Figure 3.4b).

An advantage of a pie chart is that it helps to show the size of each category relative to the whole (the category 'green eyes' represents nearly a quarter of the sample), but it is not always easy to compare the sizes of the sectors to each other (the sizes of 'blue' and 'brown' look very similar). Although pie charts are often found in everyday media reports, they are not much used in scientific publications. In mathematics, pupils construct their own pie charts and consider how they are used by others: constructing a pie chart draws on and develops a number of ideas, including data handling, working out percentages and doing calculations on angles.

Another possibility is a ***bar chart*** (Figure 3.4c). Here, it is much easier to compare the sizes of the three values ('blue' is a bit bigger than 'brown', and nearly twice as much as 'green'). However, now it is harder to judge the fractional size of each category compared with the whole.

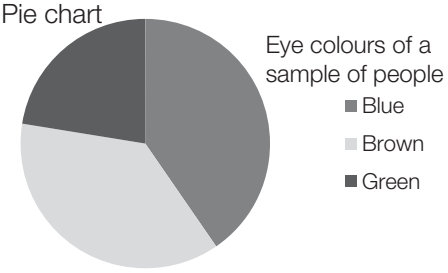
The choice of whether to use a *pie chart* or a *bar chart* depends on whether the focus is on the sizes of the categories relative to the whole or relative to each other.

**Figure 3.4** Eye colours of a sample of people

**(a)** Table of data

| Eye colour | Number |
|---|---|
| Blue | 63 |
| Brown | 58 |
| Green | 35 |
| **Total** | **156** |

**(b)** Pie chart




**(c)** Bar chart

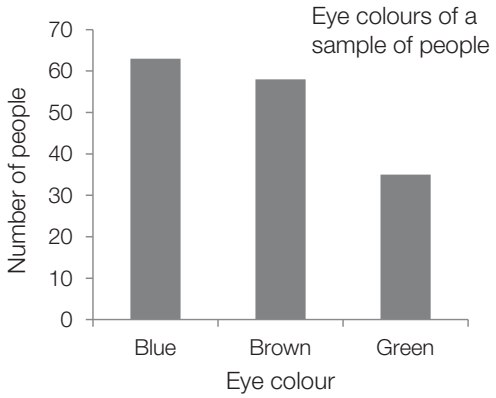
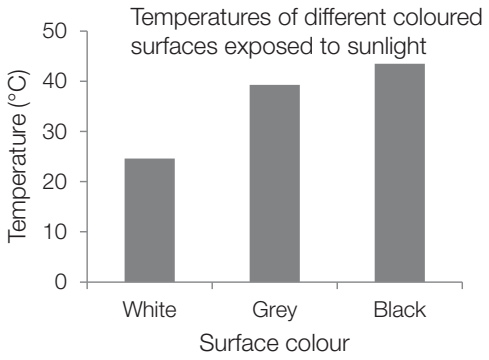



Figure 3.5 also shows a quantity (temperature) categorised by one factor (colour of surface). In this case, it is not meaningful to add these values to give a total. A pie chart would not make sense, so only a bar chart is possible.

**Figure 3.5** Temperatures of different coloured surfaces exposed to sunlight

| Surface colour | Temperature (°C) |
|---|---|
| White | 24.6 |
| Grey | 39.3 |
| Black | 43.5 |




In this example, it is not meaningful to compare the *ratios* of the values since the Celsius temperature scale has an arbitrary zero (i.e. it does not make sense to say that the value for 'black' is nearly twice as big as that for 'white'). However, comparing the heights of the bars does make it possible to compare the *differences* in temperature.

## 3.5 Charts showing a quantity categorised by two factors

Figure 3.6 shows a quantity (UK annual energy consumption) categorised by two ***factors*** (type of fuel and year). Here it is meaningful to add the values for the types of fuel together to give the total energy consumption for a year but it is *not* useful to add the values for each year together. There are a variety of different types of display that could be drawn for these data, as illustrated in Figure 3.7.

**Figure 3.6** UK annual energy consumption

Million tonnes of oil equivalent (mtoe)

| | 1975 | 2005 |
|---|---|---|
| Solid fuel | 73.7 | 39.9 |
| Petroleum | 85.0 | 78.2 |
| Gas | 35.1 | 94.3 |
| Bioenergy and waste | 0 | 4.2 |
| Primary electricity* | 8.5 | 19.8 |
| Total | **202.3** | **236.4** |

* from nuclear, hydro, wind, solar

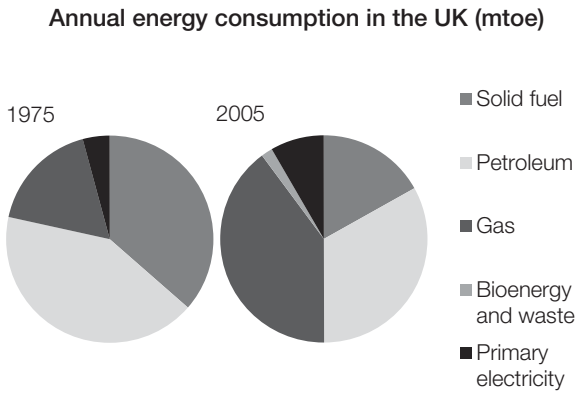
Figure 3.7a shows two ***pie charts***, one for each year, and drawn to the same size. As with a single pie chart, the focus is on comparisons of the parts with the whole for each year; however, it is not easy to compare the two to see how the proportions have changed from one year to the next. To represent the *actual* sizes of the values, the two pie charts could be drawn with *different* sizes, the area of each whole pie representing the value of the total. However, it can be very difficult to judge the relative sizes of segments from different sized pies and with different angles. In general, using multiple pie charts to make comparisons is often not very effective.

Multiple pie charts are rarely used in scientific publications. However, in mathematics lessons, drawing pie charts of different sizes can be a helpful way for pupils to think about how the sizes of the values depend on both the size and the fraction of the total.

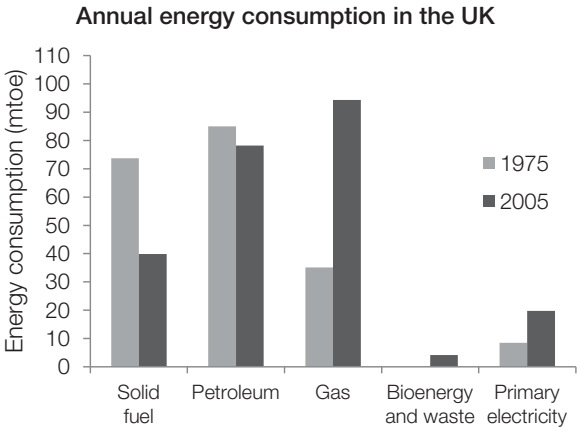
There are a number of choices of ***bar chart*** when the quantity is categorised by two factors. Figures 3.7b and 3.7c are both examples of a ***grouped bar chart*** (also known as a *clustered*

**Figure 3.7** A variety of charts showing UK annual energy consumption

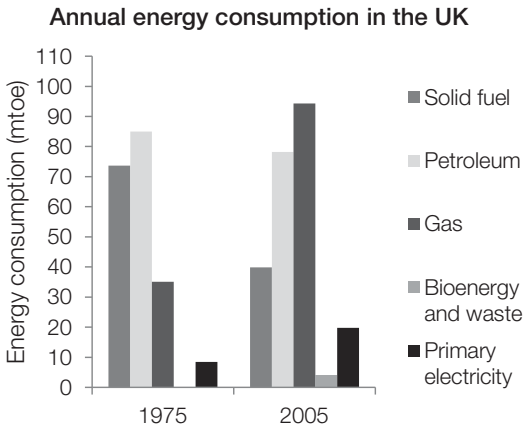
**(a)** Pie charts



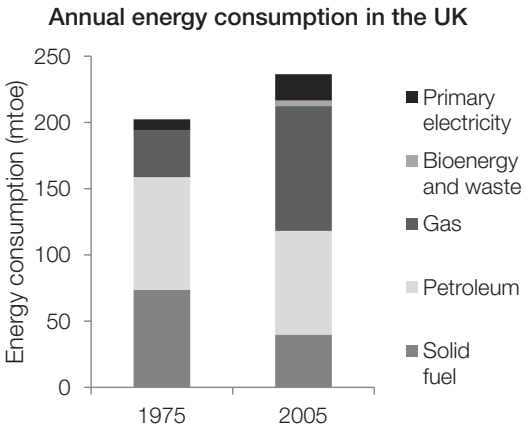

**(b)** Grouped bar chart 1



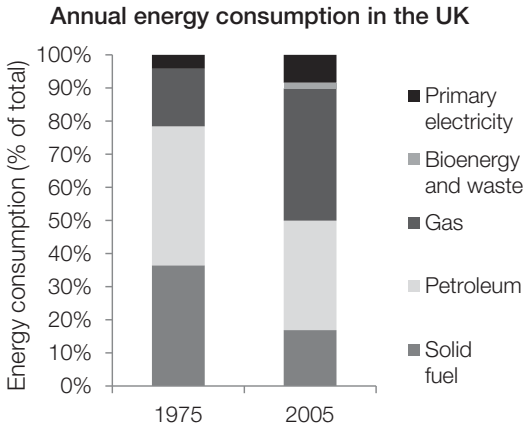

**(c)** Grouped bar chart 2



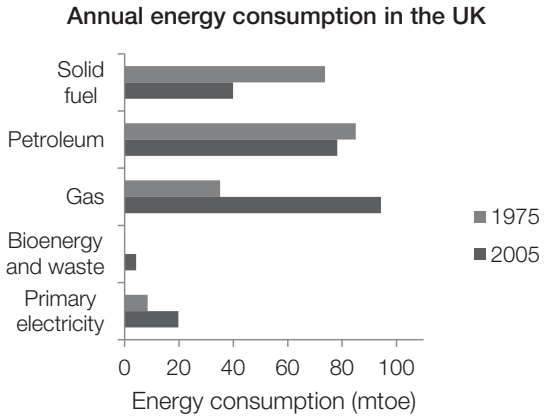

**(d)** Stacked bar chart 1




**(e)** Stacked bar chart 2




**(f)** Horizontal bar chart




*bar chart*). By showing a 'profile' for each year, Figure 3.7b makes it easier to compare the contributions of different fuels within each year, but harder to look at the change for each fuel over this period. By contrast, Figure 3.7c emphasises the change for each fuel over the period, but it is not as easy to see the contributions within a single year. The choice of which chart to draw depends on what comparison is of more interest.

Since the values for each type of fuel can be added to give a total, it is possible to draw a ***stacked bar chart*** (also known as *compound bar chart*). Two forms of the chart are shown:

one where the total height of each bar corresponds to the total quantity (Figure 3.7d) and the other where the values are expressed as percentages and the total height of each bar represents 100% (Figure 3.7e). Since the total energy consumption for each year is fairly similar, in this case the two charts are not very different from each other. In a sense, a stacked bar chart is a compromise between a pie chart and a grouped bar chart. It allows the sizes of parts to be compared with the whole (though not as easily as in a pie chart) and the parts to be compared with each other (though not as easily as in a grouped bar chart). This technique can also be applied to line graphs to show a number of different quantities (a *stacked line graph*), but these are often hard to interpret.

Finally, although all of the bar charts shown so far have been drawn with vertical bars, any type can be drawn with horizontal bars, for example as in Figure 3.7f. The eye may sometimes find it easier to make comparisons of bars by looking down a chart (in the same way that it is easier to compare numbers when written in a column).

## 3.6 Line graphs and scatter graphs: two related quantities

The tables of data in Figures 3.8, 3.9 and 3.10 all have data about two related quantities. There are a number of *similarities* between these three sets of data and the way they can be displayed, but also some important *differences*.

All of the examples can be thought of as showing a ***dependent variable*** plotted against an ***independent variable***:

- *outside temperature* against *time*
  (time is normally considered as the independent variable)
- *extension of a spring* against *mass added*
  (the mass added is the independent variable, since this is what is being changed in the experiment)
- *mean lifespan for mammals* against *mean heart rate*
  (here, the hypothesis is being tested that heart rate affects lifespan, so heart rate is being treated as the independent variable).

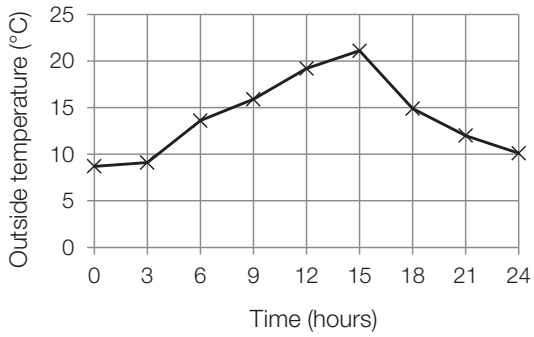
For each value of the independent variable, there is a corresponding value of the dependent variable. These values are used to plot a series of ***data points*** on a graph; the independent variable has been plotted along the ***horizontal axis*** and the dependent variable along the ***vertical axis***.

So far, all three sets of data have been treated the same – the differences arise when deciding whether to draw a line and how to draw it.

Figure 3.8 shows an example of a ***line graph***. It shows the change in outside temperature over a 24-hour period. Lines have been drawn *connecting each data point to the next one*. The assumption being made here is that each value for outside temperature is the *actual value for that particular time and place*, and so the line that is drawn passes through *all* the points. This example is a ***time series***, and the graph shows the variation of a quantity over time (a *trend*). The ***gradient*** of each line segment gives an indication of how quickly the quantity changes from one value to the next. For example, the temperature changed more slowly over the first 3-hour period than over the second.

**Figure 3.8** Outside temperature over a 24-hour period starting at midnight

| Time (hours) | Outside temperature (°C) |
|---|---|
| 0 | 8.7 |
| 3 | 9.1 |
| 6 | 13.6 |
| 9 | 15.9 |
| 12 | 19.2 |
| 15 | 21.1 |
| 18 | 14.9 |
| 21 | 12.0 |
| 24 | 10.1 |

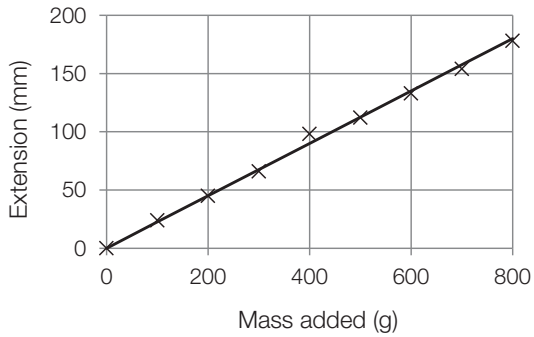



Care needs to be taken, however, in thinking about what happens between the measured values. Although the dependent variable (temperature) varies continuously throughout, and lines are used to connect the data points, these lines are *not* intended to indicate how the dependent variable changes between the data points.

The graph in Figure 3.9 is also an example of a ***line graph***. It shows how the extension of a spring depends on the mass suspended from it. Here, instead of connecting all the points together, a single straight line has been drawn that passes as close as possible to the points (though not necessarily through them), called a *line of best fit*. This type of line graph is very common in science. The assumption here is that there is a simple relationship between the two variables such that the *true values* all lie on the line: if all of the values could be measured with complete accuracy then every value of mass added would have a value of the extension of the spring that would lie on the line. In practice, not all the data points fit on this line because of *measurement uncertainties*.

**Figure 3.9** Effect of adding slotted masses to a spring

| Mass added (g) | Extension (mm) |
|---|---|
| 0 | 0 |
| 100 | 24 |
| 200 | 45 |
| 300 | 66 |
| 400 | 98 |
| 500 | 112 |
| 600 | 133 |
| 700 | 154 |
| 800 | 178 |




Unlike the previous type of line graph (in Figure 3.8), a fitted line *is* intended to indicate how the dependent variable changes between the data points. (For further details, see Section 7.5 *Interpolation and extrapolation on a line graph* on page 70).

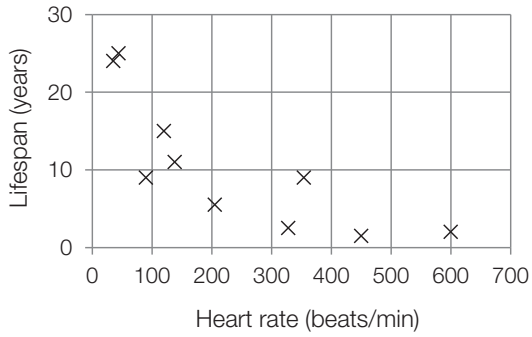
In both Figure 3.8 and Figure 3.9, the nature of the data suggests that for every value of the independent variable there will be a *single value* for the dependent variable. This is the justification for drawing a line to represent the relationship between the variables.

This is in contrast to the data in Figure 3.10, which shows the relationship between mean heart rate and mean lifespan for various types of mammal. This is a ***scatter graph*** and no line

has been drawn since the data are of a different type to the previous example involving the spring. The pattern of data points suggests that mammals with a higher heart rate (and higher metabolic rate) tend to have shorter lifespans. In this example, it is easy to imagine that there might be two types of mammal with the same mean heart rate but different lifespans (i.e. the same values for the independent variable but different values for the dependent variable). Unlike the example about the spring (Figure 3.9), there is not a unique value of the dependent variable for every value of the independent variable. A line thus cannot be drawn that passes *through or close to all the points*.

**Figure 3.10** Mean heart rates and lifespans for some selected types of mammal

| Mammal | Heart rate (beats/min) | Lifespan (years) |
|---|---|---|
| Badger | 138 | 11 |
| Cat | 120 | 15 |
| Elephant | 35 | 24 |
| Goat | 90 | 9 |
| Hamster | 450 | 1.5 |
| Horse | 44 | 25 |
| Mouse | 600 | 2 |
| Rabbit | 205 | 5.5 |
| Rat | 328 | 2.5 |
| Squirrel | 354 | 9 |




Since there does appear to be some kind of relationship between these two variables, it would be possible to draw a fitted line. The pattern of data points on the graph suggests that a curve would be a better fit than a straight line. This curve would have a different meaning to the line in Figure 3.9. Most of the data points would not be close to the fitted curve, and this would not be due to *measurement uncertainty* but to the *variability between different types of mammal*.

Note that, in mathematics, when pupils encounter 'line graph' it is usually of the type shown in Figure 3.8, and they would talk of connecting each pair of data points with a 'line segment'. In science, 'line graphs' of the type shown in Figure 3.9 are more common. When pupils draw a line of best fit in mathematics, it is more likely to be for the type of data shown in Figure 3.10 (a 'scatter graph') rather than for that shown in Figure 3.9, and the fitted line would be straight. In science lessons, pupils are expected to judge whether a line of fit should be straight or curved. (See Section 8.8 *Drawing a line of best fit on a scatter graph* on page 85.)

Care in using terminology also needs to be taken when drawing a line graph or a scatter graph with a computer spreadsheet, such as *Excel*. For example, in drawing a line graph such as that in Figure 3.9, the spreadsheet needs to have two columns of data, with values of mass and extension. Selecting these and choosing a 'line graph' option produces two lines – one for each variable plotted sequentially. Confusingly, whenever you want to plot one variable against another, a 'scatter graph' option needs to be selected, whether you want to draw a line on the graph or not.

Sometimes it can seem that different sciences have different ways of handling data. The important point made in this section is that *different types of data* are handled in different ways. Since biology, chemistry and physics are often concerned with different types of data,

each subject has different emphases on how to handle such data. Experiments involving a relationship between two continuous variables are found across the sciences but are a particular focus in physics. Such experiments lead to *'line graph' type data*. Surveys involving data collection from individuals in a population are more common in biology than other subjects and lead to *'scatter graph' type data*.

The distinction between these two fundamentally different kinds of data is very important. How such data are analysed is discussed in more detail in the following two chapters:

- Chapter 7 *Looking for relationships: line graphs* on page 64
- Chapter 8 *Looking for relationships: batches and scatter graphs* on page 75.

## 3.7 Bar charts and line graphs

Examples of ***bar charts*** and ***line graphs*** have been discussed earlier. The bar chart shown in Figure 3.5 has a horizontal axis that represents a ***categorical*** variable. The line graph shown in Figure 3.9 has a horizontal axis that represents a ***continuous*** variable. But what about data where the independent variable is a ***discrete*** variable? Is a bar chart or a line graph better for this kind of data? This is a question that can generate a good deal of disagreement. (See Section 1.4 *Naming different types of data* on page 12 for further details about the meanings of these terms, in particular the discussion about the similarities between continuous and discrete data.)

Figure 3.11 is an example of data that has a *discrete* independent variable. It shows the voltage measured across a number of batteries (or 'cells') connected in series. (The ones used were in fact standard D-size alkaline cells marked '1.5 V'.)

**Figure 3.11** Voltage across cells connected in series

| Number of cells | Voltage (V) |
|---|---|
| 1 | 1.55 |
| 2 | 3.11 |
| 3 | 4.66 |
| 4 | 6.22 |
| 5 | 7.78 |
| 6 | 9.33 |

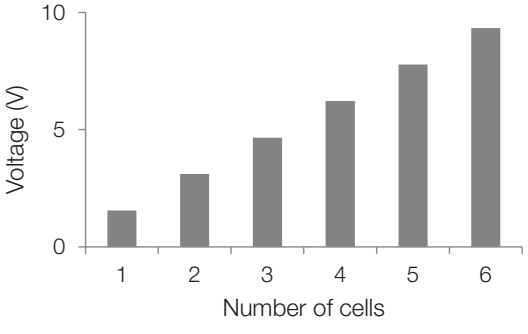
The independent variable here is 'number of cells' – it is a *discrete* variable. A discrete variable has similarities to both a *continuous* variable and a *categorical* variable:

- It is similar to a *continuous* variable in that they are both numerical (the numbers are related to the sizes of the values).
- It is similar to a *categorical* variable in that there are no 'in-between' values (e.g. '1½ cells' has no meaning).

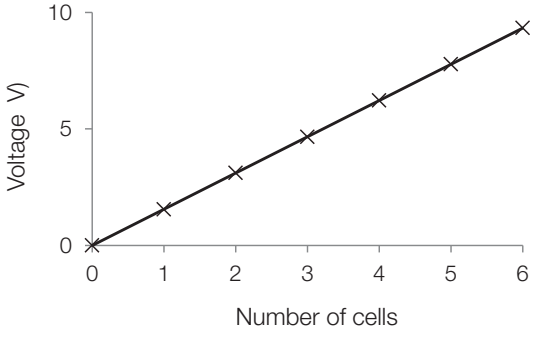
If 'number of cells' is treated as being more similar to a *categorical* variable then a *bar chart* would be plotted, as shown in Figure 3.12a; if it is treated as being more similar to a *continuous* variable then a *line graph* would be plotted, as shown in Figure 3.12b.

**Figure 3.12** Bar chart and line graph of the same data

**(a)** Bar chart




**(b)** Line graph



One of the arguments used for saying that a bar chart should be drawn for data like these is that it is not meaningful to draw a line when the values on the horizontal axis are discrete. This is because there are no 'in-between' values for the number of cells, and ***interpolation*** would not make sense (see Section 7.5 *Interpolation and extrapolation on a line graph* on page 70). For example, it is possible to read from the graph that to get a voltage of 5 V you would need about 3¼ cells – and you can't have 3¼ cells.

The problem with this argument is that, although there is nothing to stop anyone trying to interpolate on a line graph, it is not compulsory. By analogy, the relationship could be represented as an equation:

$$\text{voltage} = \text{number of cells} \times 1.55\,\text{V}$$

There is nothing here to indicate that 'number of cells' is a discrete variable and that only integer values could be substituted. In the same way, if we draw a line graph then we can use our judgement to decide on those aspects of the representation that may be useful (e.g. the gradient of the line) and those that may not (e.g. interpolation).

If there are only two values (for 1 and 2 cells) then a bar chart would certainly be better, since there are too few data points to draw a line. However, with a sufficient number of values, a line graph has many advantages.

- *Seeing patterns*: It is easier to see whether the points lie on a straight line or a curve, and to identify how close the measurements are to the fitted line.
- *Interpreting gradients*: It is possible to calculate the gradient of the line, and to obtain an equation for the relationship. The gradient is meaningful here because the numerical differences between values on the horizontal axis are meaningful (it is an *interval scale*).
- *Interpolating*: Although interpolation may not be meaningful if *all* of the discrete values have been measured, it does makes sense if there are 'missing' values. For example, if there are voltages for 1, 2, 5, 10, 15 and 20 cells then interpolation could be used to estimate voltages for other numbers of cells.
- *Extrapolating*: The line can be extrapolated to estimate values beyond the measured range.
- *Dealing with 'missing values'*: On a bar chart it may be difficult to represent discrete data consisting of a small number of values spread across a wide range. For example, suppose there are only five values corresponding to 5, 25, 50, 100 and 200 on the horizontal axis. One could either show these as equally spaced bars and lose the visual appearance of the relationship or show the whole scale from 1, 2, 3… 200, creating five narrow bars and a lot of spaces. Plotting a line graph would show the relationship more clearly. When the values of a discrete variable become very large (e.g. populations of countries), it certainly makes sense to treat these in the same way as continuous variables.
- *Meaningful non-integer values*: It is not possible to have, say, 2.5 rubber bands but, in an experiment involving forces related to rubber bands, interpolation may be meaningful. 'Number of rubber bands' becomes in effect a 'surrogate' unit of force, which is a continuous variable.

Line graphs of discrete data can also be useful when lines are used to join each data point to the next one. Such graphs are used to plot the properties of elements (e.g. their melting points) against atomic number (a discrete variable). Non-integer values of atomic number certainly have no meaning, but a bar chart of these data would be more difficult to interpret.

A line graph emphasises the peaks and the troughs in the data, and makes the periodic patterns stand out.

Sometimes, what appears to be a discrete variable on the horizontal axis actually reflects an underlying continuous variable. For example, 'Monday', 'Tuesday', 'Wednesday', and so on, look like discrete values. However, if the vertical axis represents a person's heart rate recorded at 8.00 am on each day then this really reflects a set of samples along a continuous scale. In principle, the heart rate could have been taken every hour or every minute. Similarly, if the vertical axis represents daily rainfall then this is just a conventional way of recording the total amount of this quantity. Again, in principle, the total could be recorded every hour or every minute. In both of these cases, a line graph could be justified, since the gradients are meaningful.