1 Collecting data

Key words: quantitative data, qualitative data, quantity, value, unit, resolution, scale, significant figures, range, variable, continuous, discrete, categorical, integer, experiment, survey, independent variable, dependent variable, control variable, factor, time series, raw data, primary data, secondary data.

Science is built, fundamentally, on observations of the world around us. In pure mathematics, numbers are often treated as abstract; however, in science, numbers are associated with values related to the real world. So, in thinking about the use of mathematics in science, a fitting place to start is to look at the nature of data collection.

1.1 Measuring and counting

A very obvious distinction about the types of data that can be collected is between ***quantitative data*** and ***qualitative data***. As an example, take the data that could be collected about the pupils in a class. Measuring the heights of the pupils would produce *quantitative* data, while their eye colours would be *qualitative* data.

As its name suggests, *quantitative* data relate to a ***quantity***. In this case, the quantity is 'height of pupil' and, for each pupil, there is a ***value*** for this quantity. The values are not just numbers. For example, if the height of a pupil is 119 cm then the value consists of a number (119) and a ***unit*** (cm). So, any calculation done on the *value* (119 cm) must be done on *the number and the unit*. Handling units in calculations is an important aspect of science. (See Section 2.1 *Calculations and units* on page 14 for further details.)

By contrast, the eye colours of pupils represent *qualitative* data – there are no numbers involved. If the numbers of pupils with each eye colour are *counted*, then each eye colour becomes a category with an associated numerical value. Thus, quantitative data can be generated from qualitative data. The 'eye colour of a pupil' is an *attribute of an individual* and is an example of qualitative data; the 'numbers of pupils with each eye colour' is a *variable* consisting of quantitative data.

So, there are two basic ways of collecting quantitative data – by *measuring* and by *counting*.

1.2 Measurement, resolution and significant figures

The values of measurements are limited by the ***resolution*** of the measuring instrument whether analogue (e.g. a ruler) or digital (e.g. an electronic balance). It is easy to do a calculation to get a number such as 2.3913043 displayed on a calculator; it is not possible, however, to use a school stopclock to record a value for a time such as 2.3913043 seconds. To do this would require an instrument with higher resolution.

A ruler generally has a ***scale*** that is divided into millimetres. When measuring with a ruler, it is usual to give the result to the nearest millimetre, since it is difficult by eye to judge fractions of a millimetre. The height of a sheet of A4 paper, as measured by a ruler, is 297 mm. This value has three ***significant figures*** – the number of digits that contribute information about the size of the value.

The number of significant figures is related to the *uncertainty* in the measurement. Reporting a measurement as '297 mm' does not imply that the 'true value' is *exactly* 297 mm, but that it is more likely to be nearer 297 mm than to 296 mm or 298 mm. (For more details about uncertainty, see Section 6.2 *Variability and measurement uncertainty* on page 51.)

Changing the units of measurement of a measured value does not change the number of significant figures. For example, the height of an A4 sheet of paper, 297 mm, could also be expressed as 29.7 cm or 0.297 m, or even as 0.000 297 km. All of these values have *three* significant figures, even though they have different numbers of digits after the decimal point. However, measuring a 2p coin with a ruler gives a result for the diameter of 26 mm. The resolution of the ruler is still the same, though in this case the value has only two significant figures.

In general, reading an analogue scale by eye will ideally produce a value with three significant figures, though sometimes it may have only two.

On a digital instrument, the values are read directly from the display. On kitchen scales reading to the nearest 1 g, the mass of a litre carton of orange juice might be displayed as 1082 g (four significant figures), while the mass of a 2p coin shows as 7 g (one significant figure). Again, the resolution of the instrument is the same, but the number of significant figures is different. For the coin, a more accurate value would be obtained using a balance with a greater resolution: such an instrument would tend to have a smaller ***range***. For example, kitchen scales that read to the nearest 1 g might measure up to 5000 g, while a 'pocket-sized' balance reading to the nearest 0.001 g might only read up to 20 g. On the latter balance, the mass of a 2p coin might then be displayed as 7.154 g (four significant figures).

The 'zero' digit has an important role in expressing the number of significant figures for a value. If another coin is put on the same balance, and the display shows 7.200 g, then this should be recorded as '7.200 g' and not '7.2 g'. Writing '7.200 g' indicates that its mass was measured to the nearest '0.001 g', whereas writing '7.2 g' suggests that a balance with a lower resolution that only measured to the nearest 0.1 g was used.

On some digital instruments (e.g. multimeters), it is possible to change the range depending on the value being measured, in order to increase the resolution and obtain the maximum number of significant figures.

Note that the term 'range' is applied in a number of different ways in science – to measuring instruments, to the axes on a graph, and to the values of a set of data. (See the *Glossary for teachers* on page 119 for further details.)

The number of significant figures is an indication of the *precision* of a value. Thinking about the number of significant figures of a measured value is important when rounding the values obtained in calculations. The decision of how many significant figures should be given depends on the numbers of significant figures in the starting values. (See Section 2.3 *Rounding and significant figures* on page 16.)

For further information about the terminology related to values, units and measuring instruments, see the ASE/Nuffield publication *The Language of Measurement*.

1.3 Characteristics of different types of data

Statisticians have developed further distinctions beyond ‘quantitative’ and ‘qualitative’ to describe different types of data. For secondary school science, knowing the statistical terminology is not necessary, but it is worth being aware of some key characteristics of different types of data. The distinctions are important because different types of data can be analysed and represented in different ways.

Figures 1.1–1.4 show some examples of data. Each of the columns in these tables represents a ***variable***. The following questions identify key characteristics about the variables:

- *Can the data be put into a meaningful order?*
  Quantitative data can always be put into a meaningful order (from low values to high values), but qualitative data may or may not. For example, the days of the week (in Figure 1.2) form a sequence, while the pupils’ eye colours (brown, blue, etc.) have no particular order.
- *Can the data have any value or are there distinct categories?*
  There are always distinct categories for qualitative data. However, quantitative data can sometimes take on any value (e.g. length and temperature), but may have a small number of possible numerical values (often integer values, such as number of trees – there can be 3 trees or 4 trees but not 3.7 trees).

Figure 1.1 Pupils in a class

Pupil	Height (cm)	Eye colour
1	158	brown
2	148	blue
3	142	brown
4	168	brown
etc.		

Figure 1.2 Rainfall

Day	Rainfall (mm)
Sunday	0
Monday	8
Tuesday	13
Wednesday	0
etc.	

Figure 1.3 An object cooling

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

Figure 1.4 Tree survey

Quadrat	Number of trees
A	7
B	3
C	6
D	4

- *Are the numerical differences between values meaningful?*
  Numerical *differences* between measured quantities and counts are always meaningful. For example, in Figure 1.1, pupil 1 is 158 cm tall and pupil 2 is 148 cm tall: the difference, 10 cm, is meaningful as it represents how much taller one is than the other. Differences are also meaningful for the rainfall data and for temperature. However, although the column ‘Pupil’ contains the numbers 1, 2, 3 . . ., these are simply labels, and the differences between them have no meaning.
- *Are the ratios of values meaningful?*
  In Figure 1.4, it would make sense to talk in terms of the *ratios* of the numbers of trees; for example, ‘6 trees’ are twice as many trees as ‘3 trees’. Comparing the sizes of values in this way would also be meaningful for heights of pupils and for rainfall.

However, for temperature it would not make sense to talk about 20 °C being twice as hot as 10 °C. The choice of the zero on the Celsius scale is arbitrary, and so ratios of temperatures measured in degrees Celsius are not meaningful.

The technical terms that are used for such scales of measurement are interval scales (differences between values are meaningful) and ratio scales (ratios of values are meaningful). A ratio scale includes the properties of an interval scale (so differences between values on a ratio scale are also meaningful).

1.4 Naming different types of data

There are several terms used in statistics to describe the characteristics of different types of data, but three terms are commonly used in secondary school mathematics: ***continuous***, ***discrete*** and ***categorical***. These terms may also be encountered in secondary school science but they are used less frequently than in mathematics.

- *Continuous data*: These are numerical data for which the values can take on any value within a certain range. *Measurement* produces continuous data; for example, the heights of pupils or the temperatures of an object.
- *Discrete data*: These are also numerical data but they can only take on certain values. *Counting* produces discrete data. Counts have whole number or ***integer*** values; for example, number of trees in a survey area.
- *Categorical data*: These are not numerical values so they cannot be ordered but they can be sorted into categories; for example, the eye colour of pupils.

The differences between continuous and discrete data may be less marked than their definitions suggest. Although in principle it is possible for a measurement of length or temperature (*continuous data*) to have any value, in reality a measurement will have a limited number of significant figures. For example, in Figure 1.1, the height of the first pupil is measured as 158 cm. It is quite likely that, in a sufficiently large group, there will be other pupils whose heights will also be measured as 158 cm. Similarly, in a survey counting a small sample of trees (*discrete data*), there may only be a few possible values of counts (1, 2, 3, etc.), but there could be many possible values when counting large populations. Thus, in practice, it may be that continuous and discrete data are treated rather similarly, for example in deciding whether to draw a bar chart or a line graph (see Section 3.7 *Bar charts and line graphs* on page 32).

Note that the term 'discontinuous data' is sometimes used in school science; the intention is usually to mean categorical data but the term is ambiguous and can be confusing, since there are two ways that data can be 'not continuous' (discrete and categorical).

1.5 Where do data come from?

Many scientific studies are concerned with gathering evidence about the relationships between ***variables***, and can be broadly characterised as ***experiments*** or ***surveys***.

Pupils need to be aware that both of these terms have particular meanings in science, which may be different from the way they are used in everyday life. An 'experiment' may often be seen as an unstructured and random exploration, unlike the organised and purposeful 'experiment' of science. In everyday language, the term 'survey' often suggests a study involving a questionnaire to find out about people's opinions – a meaning that is also common in the mathematics classroom. In science, the term has a broader meaning.

In the simplest kind of experiment, the experimenter changes just one variable (the ***independent variable***) and observes the effect on another variable (the ***dependent variable***). All other variables (***control variables***) are kept constant by the experimenter.

Surveys are used in more complex situations where it is harder to manipulate the variables, so data are collected by observing the outcomes in various conditions. There may be a number of independent variables, as well as other unknown variables that affect the outcomes. Sometimes, there is no clear distinction between independent and dependent variables, and a survey just explores whether relationships exist without thinking in terms of causation.

An independent variable is often referred to as a ***factor***, particularly when it is a categorical variable. (Note that the term *factor* is also used in mathematics with an entirely different meaning. See the *Glossary for teachers* on page 119.)

In experiments or surveys where changes are observed over time, the data are called a ***time series*** and time is treated as the independent variable.

So, a variable may be described as '*continuous, discrete or categorical*', and as '*independent, dependent or control*'. The first set of terms relate to the *nature of the data* of a variable, while the second set refers to the role of the variable *in the context of an investigation*.

Note that, in mathematics, the term ***variable*** refers to a quantity that can take on a range of values and is often represented by a letter (e.g. x, y) in an algebraic equation. Much of science is concerned with algebraic modelling and uses the term 'variable' in the same way. However, in science, the term 'variable' is also used in situations where an algebraic relationship is not known. (See Section 9.2 *Variables, constants and coefficients* on page 88.)

Data collected directly from experiments or surveys, before calculations are performed, are called ***raw data***. If the data are collected directly by the user, these are called ***primary data***. If the data are obtained indirectly from other sources reporting raw or processed data (such as books, articles or web pages), these are ***secondary data***.