1

Mathematics and Statistics in Science

Overview

Science students encounter mathematics and statistics in three main areas:

- Understanding and using theory.
- Carrying out experiments and analysing results.
- Presenting data in laboratory reports and essays.

Unfortunately, many students do not fully appreciate the need for understanding mathematics and/or statistics until it suddenly confronts them in a lecture or in the write-up of an experiment. There is indeed a 'chicken and egg' aspect to the problem:

> Some science students have little enthusiasm to study mathematics until it appears in a lecture or tutorial – by which time it is too late! Without the mathematics, they cannot *fully* understand the science that is being presented, and they drift into a habit of accepting a 'second-best' science *without* mathematics. The end result could easily be a drop of at least one grade in their final degree qualification.

All science is based on a *quantitative* understanding of the world around us – an understanding described ultimately by *measurable* values. Mathematics and statistics are merely the processes by which we handle these quantitative values in an effective and logical way.

Mathematics and statistics provide the network of links that tie together the details of our understanding, and create a sound basis for a fundamental appreciation of science as a whole. Without these quantifiable links, the ability of science to predict and move forward into new areas of understanding would be totally undermined.

In recent years, the data handling capability of information technology has made mathematical and statistical calculations far easier to perform, and has transformed the day-to-day work in many areas of science. In particular, a good spreadsheet program, like Excel, enables both scientists and students to carry out extensive calculations quickly, and present results and reports in a clear and accurate manner.

Essential Mathematics and Statistics for Science 2nd Edition Graham Currell and Antony Dowman

1.1 Data and Information

Real-world information is expressed in the mathematical world through **data**.

In science, some data values are believed to be fixed in nature. We refer to values that are fixed as **constants**, e.g. the constant c is often used to represent the speed of light in a vacuum, $c = 3.00 \times 10^8\ \mathrm{m\,s^{-1}}$.

However, most measured values are subject to change. We refer to these values as **variables**, e.g. T for temperature, pH for acidity.

The term **parameter** refers to a variable that can be used to describe a relevant characteristic of a scientific system, or a statistical population (see 7.2.2), e.g. the actual pH of a buffer solution, or the average (mean) age of the whole UK population. The term **statistic** refers to a variable that is used to describe a relevant characteristic of a *sampled* (see 7.2.2) set of data, e.g. five repeated measurements of the concentration of a solution, or the average (mean) age of 1000 members of the UK population.

Within this book we use the convention of printing letters and symbols that represent quantities (constants and variables) in italics, e.g. c, T and p.

The letters that represent units are presented in normal form, e.g. $\mathrm{m\,s^{-1}}$ gives the units of speed in metres per second.

There is an important relationship between data and information, which appears when analysing more complex data sets. It is a basic rule that:

> It is impossible to get more 'bits' of information from a calculation than the number of 'bits' of data that is put into the calculation.

For example, if a chemical mixture contains three separate compounds, then it is necessary to make at least three separate measurements on that mixture before it is possible to calculate the concentration of each separate compound.

In mathematics and statistics, the *number* of bits of information that are available in a data set is called the **degrees of freedom**, df, of that data set. This value appears in many statistical calculations, and it is usually easy to calculate the number of degrees of freedom appropriate to any given situation.

1.2 Experimental Variation and Uncertainty

The uncertainty inherent in scientific information is an important theme that appears throughout the book.

The **true value** of a variable is the value that we would measure if our measurement process were 'perfect'. However, because no process is perfect, the 'true value' is not normally known.

The **observed value** is the value that we produce as our *best estimate* of the true value.

The **error** in the measurement is the difference between the true value and the observed value:

$$\text{Error} = \text{Observed value} - \text{True value} \qquad [1.1]$$

As we do not normally know the 'true value', we cannot therefore know the actual error in any particular measurement. However, it is important that we have some idea of how large the error might be.

The **uncertainty** in the measurement is our *best estimate* of the magnitude of possible errors. The magnitude of the uncertainty must be derived on the basis of a proper understanding of the measurement process involved and the system being measured. The statistical interpretation of uncertainty is derived in 8.2.

The uncertainty in experimental measurements can be divided into two main categories:

Measurement uncertainty. Variations in the actual process of measurement will give some differences when the same measurement is repeated under exactly the same conditions. For example, repeating a measurement of alcohol level in the same blood sample may give results that differ by a few milligrams in each 100 millilitres of blood.

Subject uncertainty. A subject is a representative example of the system (9.1) being measured, but many of the systems in the real world have inherent variability in their responses. For example, in testing the effectiveness of a new drug, every person (subject) will have a slightly different reaction to that drug, and it would be necessary to carry out the test on a wide range of people before being confident about the 'average' response.

Whatever the source of uncertainty, it is important that any experiment must be designed both to counteract the effects of uncertainty and to quantify the magnitude of that uncertainty.

Within each of the two types of uncertainty, *measurement* and *subject*, it is possible to identify two further categories:

Random error. Each subsequent measurement has a random error, leading to *imprecision* in the result. A measurement with a low random error is said to be a *precise* measurement.

Systematic error. Each subsequent measurement has the same recurring error. A systematic error shows that the measurement is *biased*, e.g. when setting the liquid level in a burette, a particular student may always set the meniscus of the liquid a little too low.

The **precision** of a measurement is the best estimate for the purely *random error* in a measurement.

The **trueness** of a measurement is the best estimate for the *bias* in a measurement.

The **accuracy** of a measurement is the best estimate for the *overall error* in the final result, and includes both the effects of a lack of precision (due to random errors) and bias (due to systematic errors).

Example 1.1

Four groups of students each measure the pH (acidity) of a sample of soil, with each group preparing five replicate samples for testing. The results are given in Figure 1.1.

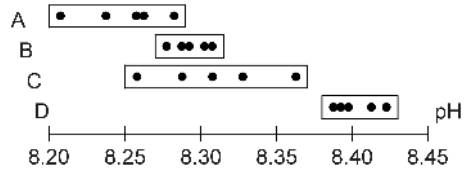

Figure 1.1 Precision and bias in experimental data.

What can be said about the *accuracy* of their results?

It is possible to say that the results from groups A and C show greater *random uncertainty* (less precision) than groups B and D. This could be due to such factors as a lack of care in preparing the five samples for testing, or some electronic instability in the pH meter being used.

Groups B and D show greater precision, but at least one of B or D must have some *bias* in their measurements, i.e. poor 'trueness'. The bias could be due to an error in setting the pH meter with a buffer solution, which would then make every one of the five measurements in the set wrong by the same amount.

With the information given, very little can be said about the overall *accuracy* of the measurements; the 'true' value is not known, and there is no information about possible bias in any of the results. For example *if the true value were* $pH = 8.40$, this would mean that groups A, B and C were all biased, with the most *accurate* measurement being group D.

The effect of random errors can be managed and quantified using suitable statistical methods (8.2, 8.3 and 15.1.2). The presentation of uncertainty as *error bars* on graphs is developed in an Excel tutorial on the Website.

Systematic errors are more difficult to manage in an experiment, but good experiment design (Chapter 15) aims to counteract their effect as much as possible.

1.3 Mathematical Models in Science

A fundamental building block of both science and mathematics is the *equation*.

Science uses the equation as a *mathematical model* to define the *relationship* between one or more factors in the real world (3.1.6). It may then be possible to use mathematics to investigate how that equation may lead to *new conclusions* about the world.

Perhaps the most famous equation, arising from the general theory of relativity, is:

$$E = mc^2$$

which relates the amount of energy, E (J), that would be released if a mass, m (kg), of matter was converted into energy (e.g. in a nuclear reactor). E and m are both variables and the constant $c(= 3.00 \times 10^8\ \mathrm{m\,s^{-1}})$ is the speed of light.

Example 1.2

Calculate the amount of matter, m, that must be converted *completely* into energy, if the amount of energy, E, is equivalent to that produced by a medium-sized power station in one year: $E = 1.8 \times 10^{13}$ J.

Rearranging the equation $E = mc^2$ gives:

$$m = \frac{E}{c^2}$$

Substituting values into the equation:

$$m = \frac{1.8 \times 10^{13}}{(3.00 \times 10^8)^2} \Rightarrow 0.000\,20 \text{ kg} \Rightarrow 0.20 \text{ g}$$

This equation tells us that if only 0.20 g of matter is converted into energy, it will produce an energy output equivalent to a power station operating for a year!

This is why the idea of nuclear power continues to be so very attractive.

Example 1.2 indicates some of the common mathematical processes used in handling equations in science: rearranging the equation, using scientific notation, changing of units, and 'solving' the equation to derive the value of an unknown variable.

Equations are used to represent many different types of scientific processes, and often employ a variety of *mathematical functions* to create suitable models.

In particular, many scientific systems behave in a manner that is best described using an exponential or logarithmic function, e.g. drug elimination in the human body, pH values. Example 1.3 shows how both the growth and decay of a bacteria population can be described, in part, by exponential functions.

Example 1.3

Figure 1.2 gives a plot of growth and decay in a bacteria batch colony, by plotting $\log(N)$ against time, t, where N is the number of cells per millilitre.

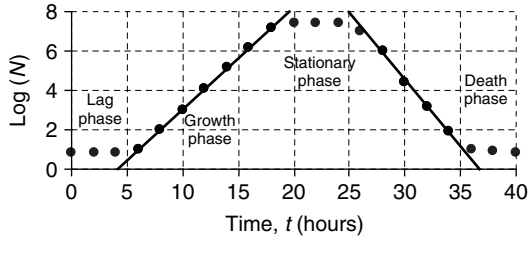

Figure 1.2 Lifecycle of a bacterial population.

The 'straight line' sections of the graph in the 'growth' and 'death' phases are two sections of the lifecycle that can be described by *exponential* functions (5.2).

Another aspect of real systems is that they often have significant inherent *variability*, e.g. similar members of a plant crop grow at different rates, or repeated measurements of the refractive index of glass may give different results. In these situations, we need to develop *statistical models* that we can use to describe the underlying behaviour of the system as a whole.

The particular statistical model that best fits the observed data is often a good guide to the scientific processes that govern the system being measured. Example 1.4 shows the Poisson distribution that could be expected if plants were distributed *randomly* with an average of 3.13 plants per unit area.

Example 1.4

Figure 1.3 shows the numbers (frequencies) of specific plants measured in 100 quadrats of unit area. In sampling the *random* distribution of plants it was found that 22 quadrats had 3 plants, 11 quadrats had 5 plants, etc.

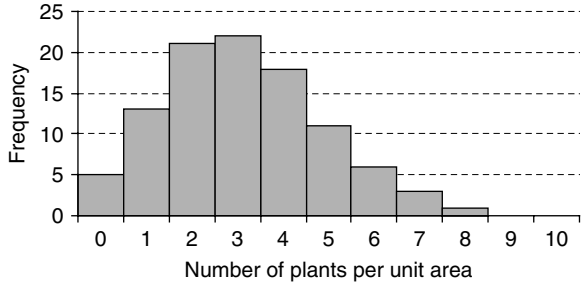

Figure 1.3 Poisson distribution of random plant abundance.

If the distribution of plants were affected by clumping or by competition for survival, then we would expect the *shape of the distribution* to be different.

Excel spreadsheets have become particularly useful for implementing mathematical models of very complex scientific systems. Throughout this book we continue to develop mathematics and statistics in conjunction with their practical applications through Excel.