Chapter 1

Introduction to scientific data analysis

1.1 Introduction

'The principle of science, the definition almost, is the following: *The test of all knowledge is experiment*. Experiment is the *sole judge* of scientific "truth"'.

So wrote Richard Feynman, famous scientist and Nobel Prize winner, noted for his contributions to physics.[1]

It is possible that when Feynman wrote these words he had in mind elaborate experiments devised to reveal the 'secrets of the Universe', such as those involving the creation of new particles during high energy collisions in particle accelerators or others to determine the structure of DNA.[2] Experimentation encompasses an enormous range of more humble (but extremely important) activities such as testing the temperature of a baby's bath water by immersing an elbow into the water, or pressing on a bicycle tyre to establish whether it needs inflating. The absence of numerical measures of quantities distinguishes these experiments from those normally performed by scientists.

Many factors directly or indirectly influence the fidelity of data gathered during an experiment such as the quality of the experimental design, experimenter competence, instrument limitations and time available to perform the experiment. Identifying, appreciating and, where possible, accounting for, such factors are key tasks that must be carried out by an experimenter. After every care has been taken to acquire the best data possible, it is time to apply techniques of data analysis to extract the most from the data. The process of extraction requires qualitative as well as quantitative methods of analysis. The

[1] See Feynman, Leighton and Sands (1963).
[2] DNA stands for deoxyribonucleic acid.

first steps require consideration be given to how data may be summarised numerically and graphically.[3] This is the main focus of this chapter. Some of the ideas touched upon in this chapter, such as those relating to error and uncertainty, will be revisited in more detail in later chapters.

1.2 Scientific experimentation

To find out something about the world, we experiment. A child does this naturally, with no training or scientific apparatus. Through a potent combination of curiosity and trial and error, a child quickly creates a viable model of the 'way things work'. This allows the consequences of a particular action to be anticipated. Curiosity plays an equally important role in the professional life of a scientist who may wish to know the

- amount of contaminant in a pharmaceutical;
- concentration of CO_2 in the Earth's atmosphere;
- distribution of temperature across a leaf;
- stresses experienced by the wings of an aircraft;
- blood pressure of a person;
- frequency of electrical signals generated by the human brain.

Scientists look for relationships between quantities. For example, a scientist may wish to establish how the amount of energy radiated from a body each second depends on the temperature of that body. In formulating the problem, designing and executing the experiment and analysing the results, the intention may be to extend the domain of applicability of an established theory, or present convincing evidence of the limitations of that theory. Where results obtained conflict with accepted ideas or theories, a key goal is to provide a better explanation of the results. Before publishing a new and perhaps controversial explanation, the scientist needs to be confident in the data gathered and the methods used to analyse those data. This requires that experiments be well designed. In addition, good experimental design helps anticipate difficulties that may occur during the execution of the experiment and encourages the efficient use of resources.

Successful experimentation is often a combination of good ideas, good planning, perseverance and hard work. Though it is possible to discover something interesting and new 'by accident', it is usual for science to progress by small steps taken by many researchers. The insights gained by researchers (both experimentalists and theorists) combine to provide answers and explanations to some questions, and in the process create new questions that need to

[3] This is sometimes referred to as 'exploratory data analysis'.

be addressed. In fact, even if something new *is* found by chance, it is likely that the discovery will remain a curiosity until a serious scientific investigation is carried out to determine if the discovery or effect is real or illusory. While scientists are excited by new ideas, a healthy amount of scepticism remains until the ideas have been subjected to serious and sustained examination by others.

1.2.1 Aim of an experiment

An experiment needs a focus, more usually termed an 'aim', which is something the experimenter returns to during the design and analysis phases of the experiment. Essentially the aim embodies a question which can be expressed as 'what are we trying to find out by performing the experiment?'.

Expressing the aim clearly and concisely before the experiment begins is important, as it is reasonable to query as the experiment progresses whether the steps taken are succeeding in addressing the aim, or whether the experiment has deviated 'off track'. Deviating from the main aim is not necessarily a bad thing. After all, if you observe an interesting and unexpected effect during the course of an experiment, it would be quite natural to want to know more, as rigidly pursuing the original aim might cause an important discovery to be overlooked. Nevertheless, it is likely that if a new effect *has* been observed, this effect deserves its own separate and carefully planned experiment.

Implicit in the aim of the experiment is an idea or hypothesis that the experimenter wishes to promote or test, or an important question that requires clarification. Examples of questions that might form the basis of an experiment include the following.

- Is a new spectroscopic technique better able to detect impurities in silicon than existing techniques?
- Does heating a glass substrate during vacuum deposition of a metal improve the quality of films deposited onto the substrate?
- To what extent does a reflective coating on windows reduce the heat transfer into a motor vehicle?
- In what way does the efficiency of a thermoelectric cooler depend on the size of the electrical current supplied to the cooler?
- How does the flow rate of fluid through a hollow tube depend on the internal diameter of that tube?

Such questions can be restated explicitly as aims of a scientific investigation. It is possible to express those aims in a number of different, but essentially equivalent, ways. For example:

(a) the aim of the experiment is to determine the change in heat transfer to a motor vehicle when a reflective coating is applied to the windows of that vehicle;
(b) the aim of the experiment is to test the hypothesis that a reflective coating applied to the windows of a motor vehicle reduces the amount of heat transferred into that vehicle.

Most physical scientists and engineers would recognise (a) as a familiar way in which an aim is expressed in their disciplines. By contrast, the explicit inclusion of a hypothesis to be tested, as stated in (b) is often found in studies in the biological, medical and behavioural sciences. The difference in the way the aim is expressed is largely due to the conventions adopted by each discipline, as all have a common goal of advancing understanding and knowledge through experimentation, observation and analysis.

1.2.2 Experimental design

Deciding the aim or purpose of an experiment at an early stage is important, as precious resources (including the time of the experimenter) are to be devoted to the experiment. Experimenting is such an absorbing activity that it is possible for the aims of an experiment to become too ambitious. For example, the aim of an experiment might be to determine the effect on the thermal properties of a ceramic when several types of atoms are substituted for (say) atoms of calcium in the ceramic. If a month is available for the study, careful consideration must be given to the number of samples of ceramic that can be prepared and tested and whether a more restricted aim, perhaps concentrating on the substitution of just one type of atom, would be more judicious.

Once the aim of an experiment is decided, a plan of how that aim might be achieved is devised. Matters that must be considered include the following.

- What quantities are to be measured during the experiment?
- Over what ranges should the controllable quantities be measured?
- What are likely to be the dominant sources of error, and how can the errors be minimised?
- What equipment is needed and what is its availability?
- In what ways are the data to be analysed?
- Does the experimenter need to become skilled in new techniques (say, how to operate an electron microscope) in order to complete the experiment?
- Does new equipment need to be designed/constructed/acquired or does existing equipment require modification?
- Is there merit in developing a computer controlled acquisition system to gather the data?

- How much time is available to carry out the experiment?
- Are the instruments to be used performing within their specifications?

A particularly important aspect of experimentation is the identification of influences that can affect any result obtained through experiment or observation. Such influences are regarded as sources of 'experimental error' and we will have cause to consider these in this text. In the physical sciences, many of the experimental variables that would affect a result are readily identifiable and some are under the control of the experimenter. Identifying sources that would adversely influence the outcomes of an experiment may lead to ways in which the influence might be minimised. For example, the quality of a metal film deposited onto a glass substrate may be dependent upon the temperature of the substrate during the deposition process. By improving the temperature control of the system, so that the variability of the temperature of the substrate is reduced to (say) less than 5 °C, the quality of the films may be enhanced.

Despite the existence of techniques that allow us to draw out much from experimental data, a good experimenter does not rely on data analysis to compensate for data of dubious quality. If large scatter is observed in data, a sensible option is to investigate whether improved experimental technique can reduce the scatter. For example, incorporating electromagnetic shielding as part of an experiment requiring the measurement of extremely small voltages can improve the quality of the data dramatically and is preferred to the application of sophisticated data analysis techniques which attempt to compensate for shortcomings in the data.

An essential feature of experiments in the physical sciences is that the measurement process yields numerical values for quantities such as temperature, pH, strain, pressure and voltage. These numerical values (often referred to as *experimental data*) may be algebraically manipulated, graphed, compared with theoretical predictions or related to values obtained by other experimenters who have performed similar experiments.

1.3 The vocabulary of measurement

Scientists draw on statistical methods as well as those deriving from the science of measurement (termed *metrology*) when analysing their data. A consequence is that sometimes there is inconsistency between the way terms, such as error and uncertainty, are used in texts on the treatment of data written by statisticians and by those written by metrologists. The diversity of terms can be a distraction. In this text we will tend to rely on the internationally recognised explanation of terms found in the 'International vocabulary of metrology' (usually abbreviated to the VIM).[4]

1.4 Units and standards

Whenever a value is recorded in a table or plotted on a graph, the unit of measurement must be stated, as numbers by themselves have little meaning. To encompass all quantities that we might measure during an experiment, we need units that are:

- comprehensive;
- clearly defined;
- internationally endorsed;
- easy to use.

Reliable and accurate standards based on the definition of a unit must be available so that instruments designed to measure specific quantities may be compared against those standards. Without agreement between experimenters in, say, Australia, the United Kingdom and the United States, as to what constitutes a metre or a second, a comparison of values obtained by each experimenter would be impossible.

A variety of instruments may be employed to measure quantities in the physical sciences. These range from a simple hand-held stopwatch for timing a body in free-fall, to a state of the art HPLC[5] to determine the concentration of contaminant in a pharmaceutical. Whatever the particular details of a scientific investigation, we generally attach much importance to the 'numbers' that emerge from an experiment as they may provide support for a new theory of the origin of the Universe, assist in monitoring the concentration of CO_2 in the Earth's atmosphere, or help save a life. Referring to the outcome of a measurement as a 'number' is rather vague and misleading. Through experiment we obtain *values*. A value is the product of a number and the unit in which the measurement is made. The distinction in scientific contexts between number and value is important. Table 1.1 includes definitions of number, value, and other important terms as they are used in this text.

[4] ISO/IEC Guide 99:2007, International vocabulary of metrology – Basic and general concepts and associated terms (VIM). Available as a free download from http://www.bipm.org/en/publications/guides/vim.html [accessed 30/6/2011].

[5] HPLC stands for High Performance Liquid Chromatography.

Table 1.1. *Definitions of commonly used terms in data analysis.*

Term	Definition
Quantity	An attribute or property of a body, phenomenon or material. Examples of quantities are: the temperature, mass or electrical capacitance of a body, the time elapsed between two events such as starting and stopping a stopwatch, or the resistivity of a metal.
Unit	An amount of a quantity, suitably defined and agreed internationally, against which some other amount of the same quantity may be compared. As examples, the kelvin is a unit of temperature, the second is a unit of time and the ohm-metre is a unit of resistivity.
Value	The product of a number and a unit. As examples, 273 K is a value of temperature, 0.015 s is a value of time interval and 1.7×10^{-8} Ω·m is a value of resistivity.
Measurement	A process by which a value of a quantity is determined. For example, the measurement of water temperature using an alcohol-in-glass thermometer entails immersing a thermometer into the water followed by estimating the position of the top of a narrow column of alcohol against an adjacent scale.
Data	Values obtained through measurement or observation.

1.4.1 Units

The most widely used system of units in science is the SI system[6] and has been adopted officially by most countries around the world. Despite strongly favouring SI units in this text, we will also use some non-SI units such as the minute and the degree, as these are likely to remain in widespread use in science for the foreseeable future.

The origins of the SI system can be traced to pioneering work done on units in France in the late eighteenth century. In 1960 the name 'SI system' was adopted and at that time the system consisted of six fundamental or 'base' units. Since 1960 the system has been added to and refined and remains constantly under review. From time to time suggestions are made regarding how the definition of a unit may be improved. If this allows for easier or more accurate realisation of the

[6] SI stands for Système International. An authoritative document on the SI system prepared by the Bureau International des Poids et Mesures (custodians of the SI system) is freely available as a download from www.bipm.org/utils/common/pdf/si_brochure_8_en.pdf [accessed 9/11/2010].

Table 1.2. *SI base units, symbols and definitions.*

Quantity	Unit	Symbol	Definition
Mass	kilogram	kg	The kilogram is equal to the mass of the international prototype of the kilogram. (The prototype kilogram is made from an alloy of platinum and iridium and is kept under very carefully controlled environmental conditions by the Bureau International des Poids et Mesures (BIPM) in Sèvres near Paris, France.)
Length	metre	m	The metre is the length of the path travelled by light in a vacuum during a time interval of $\frac{1}{299792458}$ of a second.
Time	second	s	The second is the duration of 9192631770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom.
Thermodynamic temperature	kelvin	K	The kelvin is the fraction $\frac{1}{273.16}$ of the thermodynamic temperature of the triple point of water.
Electric current	ampere	A	The ampere is that current which, if maintained between two straight parallel conductors of infinite length, of negligible cross-section and placed one metre apart in a vacuum, would produce between these conductors a force of 2×10^{-7} newton per metre of length.
Luminous intensity	candela	cd	The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{14} hertz and that has a radiant intensity in that direction of $\frac{1}{683}$ watt per steradian.
Amount of substance	mole	mol	The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12.

unit as a standard (permitting, for example, improvements in instrument calibration), then appropriate modifications are made to the definition of the unit.

Currently the SI system consists of seven base units as defined in table 1.2.

Other quantities may be expressed in terms of the base units. For example, energy can be expressed in units $kg{\cdot}m^2{\cdot}s^{-2}$ and electric potential difference in units $kg{\cdot}m^2{\cdot}s^{-3}{\cdot}A^{-1}$. The cumbersome nature of units expressed this way is such that other, so called *derived* units, are introduced which are formed from products of the base units. Some familiar quantities with their units expressed in derived and base units are shown in table 1.3.

Table 1.3. *Symbols and units of some common quantities.*

Quantity	Derived unit	Symbol	Unit of quantity expressed in base units
Energy, work	joule	J	$kg{\cdot}m^2{\cdot}s^{-2}$
Force	newton	N	$kg{\cdot}m{\cdot}s^{-2}$
Power	watt	W	$kg{\cdot}m^2{\cdot}s^{-3}$
Potential difference, electromotive force (emf)	volt	V	$kg{\cdot}m^2{\cdot}s^{-3}{\cdot}A^{-1}$
Electrical charge	coulomb	C	$s{\cdot}A$
Electrical resistance	ohm	Ω	$kg{\cdot}m^2{\cdot}s^{-3}{\cdot}A^{-2}$

Example 1

The farad is the SI derived unit of electrical capacitance. With the aid of table 1.3, express the unit of capacitance in terms of the base units, given that the capacitance, *C*, may be written

$$C = \frac{Q}{V}, \tag{1.1}$$

where *Q* represents electrical charge and *V* represents potential difference.

ANSWER

From table 1.3, the unit of charge expressed in base units is s·A and the unit of potential difference is $kg{\cdot}m^2{\cdot}s^{-3}{\cdot}A^{-1}$. It follows that the unit of capacitance can be expressed with the aid of equation 1.1 as,

$$\text{unit of capacitance} = \frac{s \cdot A}{kg \cdot m^2 \cdot s^{-3} \cdot A^{-1}} = kg^{-1} \cdot m^{-2} \cdot s^4 \cdot A^2.$$

Exercise A

The henry is the derived unit of electrical inductance in the SI system of units. With the aid of table 1.3, express the unit of inductance in terms of the base units, given the relationship

$$E = -L\frac{dI}{dt}, \tag{1.2}$$

where *E* represents emf, *L* represents inductance, *I* represents electric current, and *t* represents time.

1.4.2 Standards

How do the definitions of the SI units in table 1.2 relate to measurements made in a laboratory? For an instrument to measure a quantity in SI units, the definitions need to be made 'tangible' so that an example or *standard* of the unit is made available. Only when the definition is realised as a practical and maintainable standard, can values obtained by an instrument designed to measure the quantity be compared against that standard. Where a difference is established between standard and instrument, that difference is stated as a correction to the instrument. The process by which the comparison is made and the issuing of a statement of discrepancy is referred to as *calibration*.

Accurate standards based on the definitions of some of the units appearing in table 1.2 are realised in specialist laboratories. For example, a clock based on the properties of caesium atoms can reproduce the second to high accuracy.[7] By comparison, creating an accurate standard of the ampere based directly on the definition of the ampere appearing in table 1.2 is more challenging. In this case it is common for laboratories to maintain standards of related derived SI units such as the volt and the ohm, which can be implemented to high accuracy.

Most countries have a 'national standards laboratory', or equivalent, which maintains the most accurate standards achievable, referred to as *primary* standards. From time to time each national laboratory compares its standards with other primary standards held in other national laboratories around the world. In addition, a national laboratory creates and calibrates secondary standards by reference to the primary standard. Such secondary standards are found in government, industrial and university laboratories. Secondary standards in turn are used to calibrate and maintain working standards and eventually a working standard may be used to calibrate (for example) a hand-held voltmeter used in an experiment.

The result of a measurement is said to be *traceable* if, by a documented chain of comparisons involving secondary and working standards, the results can be compared with a primary standard. Traceability is very important in some situations, particularly when the 'correctness' of a value indicated by an instrument is in dispute.

[7] See appendix 2 of The International System of Units (English translation) 8th edition, 2006, published by BIPM. Appendix 2 is freely available as a download from http://www.bipm.org/utils/en/pdf/SIApp2_s_en.pdf [accessed 2/11/2010].