# 1

# Introduction: Data and Its Properties, Analytical Methods and Jargon



**Points covered in this chapter**

- Types of data
- Sources of data
- The nature of data
- Scales of measurement
- Data distribution
- Population and sample properties
- Outliers
- Terminology

## PREAMBLE

This book is not a textbook although it does aim to teach the reader how to do things and explain how or why they work. It can be thought of as a handbook of data analysis; a sort of workshop manual for the mathematical and statistical procedures which scientists may use in order to extract information from their experimental data. It is written for scientists who want to analyse their data 'properly' but who don't have the time or inclination to complete a degree course in statistics in order

*A Practical Guide to Scientific Data Analysis* David Livingstone

to do this. I have tried to keep the mathematical and statistical theory to a minimum, sufficient to explain the basis of the methods but not too much to obscure the point of applying the procedures in the first case.

I am a chemist by training and a 'drug designer' by profession so it is inevitable that many examples will be chemical and also from the field of molecular design. One term that may often appear is QSAR. This stands for Quantitative Structure Activity Relationships, a term which covers methods by which the biological activity of chemicals is related to their chemical structure. I have tried to include applications from other branches of science but I hope that the structure of the book and the way that the methods are described will allow scientists from all disciplines to see how these sometimes obscure-seeming methods can be applied to their own problems.

For those readers who work within my own profession I trust that the more 'generic' approach to the explanation and description of the techniques will still allow an understanding of how they may be applied to their own problems. There are, of course, some particular topics which only apply to molecular design and these have been included in Chapter 10 so for these readers I recommend the unusual approach of reading this book by starting at the end. The text also includes examples from the drug design field, in some cases very specific examples such as chemical library design, so I expect that this will be a useful handbook for the molecular designer.

## 1.1 INTRODUCTION

Most applications of data analysis involve attempts to fit a model, usually quantitative,[1] to a set of experimental measurements or observations. The reasons for fitting such models are varied. For example, the model may be purely empirical and be required in order to make predictions for new experiments. On the other hand, the model may be based on some theory or law, and an evaluation of the fit of the data to the model may be used to give insight into the processes underlying the observations made. In some cases the ability to fit a model to a set of data successfully may provide the inspiration to formulate some new hypothesis. The type of model which may be fitted to any set of data depends not only on the nature of the data (see Section 1.4) but also on the intended use of the model. In many applications a model is meant to be used predictively,

[1] According to the type of data involved, the model may be qualitative.

but the predictions need not necessarily be quantitative. Chapters 4 and 5 give examples of techniques which may be used to make qualitative predictions, as do the classification methods described in Chapter 7.

In some circumstances it may appear that data analysis is not fitting a model at all! The simple procedure of plotting the values of two variables against one another might not seem to be modelling, unless it is already known that the variables are related by some law (for example absorbance and concentration, related by Beer's law). The production of a bivariate plot may be thought of as fitting a model which is simply dictated by the variables. This may be an alien concept but it is a useful way of visualizing what is happening when multivariate techniques are used for the display of data (see Chapter 4). The resulting plots may be thought of as models which have been fitted by the data and as a result they give some insight into the information that the model, and hence the data, contains.

## 1.2 TYPES OF DATA

At this point it is necessary to introduce some jargon which will help to distinguish the two main types of data which are involved in data analysis. The observed or experimentally measured data which will be modelled is known as a *dependent variable* or variables if there are more than one. It is expected that this type of data will be determined by some features, properties or factors of the system under observation or experiment, and it will thus be dependent on (related by) some more or less complex function of these factors. It is often the aim of data analysis to predict values of one or more dependent variables from values of one or more *independent variables*. The independent variables are observed properties of the system under study which, although they may be dependent on other properties, are not dependent on the observed or experimental data of interest. I have tried to phrase this in the most general way to cover the largest number of applications but perhaps a few examples may serve to illustrate the point. Dependent variables are usually determined by experimental measurement or observation on some (hopefully) relevant test system. This may be a biological system such as a purified enzyme, cell culture, piece of tissue, or whole animal; alternatively it may be a panel of tasters, a measurement of viscosity, the brightness of a star, the size of a nanoparticle, the quantification of colour and so on. Independent variables may be determined experimentally, may be observed themselves, may be calculated or may be

| ID | Response | Ind 1 | Ind 2 | Ind 3 | Ind 4 | Ind 5 |
|---|---|---|---|---|---|---|
| Case 1 | 14 | 1.6 | 136 | 0.03 | -12.6 | 19542 |
| Case 2 | 24 | 2 | 197 | 0.07 | -8.2 | 15005 |
| Case 3 | -6 | 9.05 | 211 | 0.1 | -1 | 10098 |
| Case 4 | 19 | 6 | 55 | 0.005 | -0.99 | 17126 |
| Case 5 | 88.2 | 3.66 | 126 | 0.8 | 0 | 19183 |
| Case 6 | 43 | 12 | 83 | 0.79 | -1.3 | 12087 |
| ……. | ……. | ……. | ……. | ……. | ……. | ……. |
| ……. | ……. | ……. | ……. | ……. | ……. | ……. |
| Case *n* | 11 | 7.05 | 156 | 0.05 | -6.5 | 16345 |

**Figure 1.1** Example of a dataset laid out as a table.

controlled by the investigator. Examples of independent variables are temperature, atmospheric pressure, time, molecular volume, concentration, distance, etc.

One other piece of jargon concerns the way that the elements of a data set are 'labelled'. The data set shown in Figure 1.1 is laid out as a table in the 'natural' way that most scientists would use; each row corresponds to a sample or experimental observation and each column corresponds to some measurement or observation (or calculation) for that row.

The rows are called 'cases' and they may correspond to a sample or an observation, say, at a time point, a compound that has been tested for its pharmacological activity, a food that has been treated in some way, a particular blend of materials and so on. The first column is a label, or case identifier, and subsequent columns are variables which may also be called descriptors or properties or features. In the example shown in the figure there is one case label, one dependent variable and five independent variables for $n$ cases which may also be thought of as an $n$ by 6 matrix (ignoring the case label column). This may be more generally written as an $n$ by $p$ matrix where $p$ is the number of variables. There is nothing unsual in laying out a data set as a table. I expect most scientists did this for their first experiment, but the concept of thinking of a data set as a mathematical construct, a matrix, may not come so easily. Many of the techniques used for data analysis depend on matrix manipulations and although it isn't necessary to know the details of operations such as matrix multiplication in order to use them, thinking of a data set as a matrix does help to explain them.

Important features of data such as scales of measurement and distribution are described in later sections of this chapter but first we should consider the sources and nature of the data.

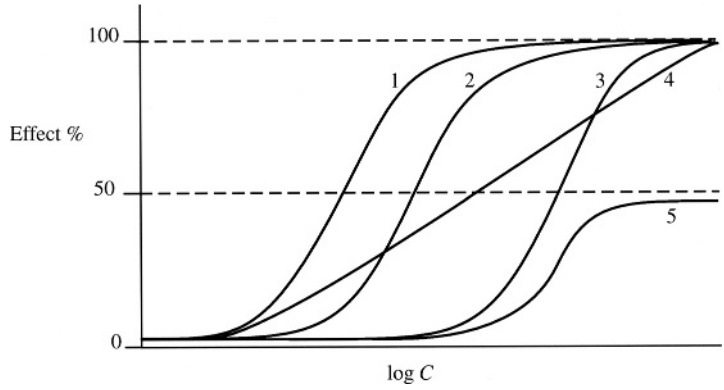


**Figure 1.2** Typical and not so typical dose–response curves for a set of five different compounds.

## 1.3 SOURCES OF DATA

### 1.3.1 Dependent Data

Important considerations for dependent data are that their measurement should be well defined experimentally, and that they should be consistent amongst the cases (objects, samples, observations) in a set. This may seem obvious, and of course it is good scientific practice to ensure that an experiment is well controlled, but it is not always obvious that data is consistent, particularly when analysed by someone who did not generate it. Consider the set of curves shown in Figure 1.2 where biological effect is plotted against concentration.

Compounds 1–3 can be seen to be 'well behaved' in that their dose–response curves are of very similar shape and are just shifted along the concentration axis depending on their potency. Curves of this sigmoidal shape are quite typical; common practice is to take 50 % as the measure of effect and read off the concentration to achieve this from the dose axis. The advantage of this is that the curve is linear in this region; thus if the $ED_{50}$ (the dose to give 50 % effect) has been bracketed by experimental measurements, it simply requires linear interpolation to obtain the $ED_{50}$. A further advantage of this procedure is that the effect is changing most rapidly with concentration in the 50 % part of the curve. Since small changes in concentration produce large changes in effect it is possible to get the most precise measure of the concentration

required to cause a standard effect. The curve for compound 4 illustrates a common problem in that it does not run parallel to the others; this compound produces small effects (<50 %) at very low doses but needs comparatively high concentrations to achieve effects in excess of 50 %. Compound 5 demonstrates yet another deviation from the norm in that it does not achieve 50 % effect. There may be a variety of reasons for these deviations from the usual behaviour, such as changes in mechanism, solubility problems, and so on, but the effect is to produce inconsistent results which may be difficult or impossible to analyse.

The situation shown here where full dose–response data is available is very good from the point of view of the analyst, since it is relatively easy to detect abnormal behaviour and the data will have good precision. However, it is often time-consuming, expensive, or both, to collect such a full set of data. There is also the question of what is required from the test in terms of the eventual application. There is little point, for example, in making precise measurements in the millimolar range when the target activity must be of the order of micromolar or nanomolar. Thus, it should be borne in mind that the data available for analysis may not always be as good as it appears at first sight. Any time spent in a preliminary examination of the data and discussion with those involved in the measurement will usually be amply repaid.

### 1.3.2 Independent Data

Independent variables also should be well defined experimentally, or in terms of an observation or calculation protocol, and should also be consistent amongst the cases in a set. It is important to know the precision of the independent variables since they may be used to make predictions of a dependent variable. Obviously the precision, or lack of it, of the independent variables will control the precision of the predictions. Some data analysis techniques assume that all the error is in the dependent variable, which is rarely ever the case.

There are many different types of independent variables. Some may be controlled by an investigator as part of the experimental procedure. The length of time that something is heated, for example, and the temperature that it is heated to may be independent variables. Others may be obtained by observation or measurement but might not be under the control of the investigator. Consider the case of the prediction of tropical storms where measurements may be made over a period of time of ocean temperature, air pressure, relative humidity, wind speed and so on. Any or all of these

parameters may be used as independent variables in attempts to model the development or duration of a tropical storm.

In the field of molecular design[2] the independent variables are most often physicochemical properties or molecular descriptors which characterize the molecules under study. There are a number of ways in which chemical structures can be characterized. Particular chemical features such as aromatic rings, carboxyl groups, chlorine atoms, double bonds and suchlike can be listed or counted. If they are listed, answering the question 'does the structure contain this feature?', then they will be binary descriptors taking the value of 1 for present and 0 for absent. If they are counts then the parameter will be a real valued number between 0 and some maximum value for the compounds in the set. Measured properties such as melting point, solubility, partition coefficient and so on are an obvious source of chemical descriptors. Other parameters, many of them, may be calculated from a knowledge of the 2-dimensional (2D) or 3-dimensional (3D) structure of the compounds [1, 2]. Actually, there are some descriptors, such as molecular weight, which don't even require a 2D structure.

## 1.4 THE NATURE OF DATA

One of the most frequently overlooked aspects of data analysis is consideration of the data that is going to be analysed. How accurate is it? How complete is it? How representative is it? These are some of the questions that should be asked about any set of data, preferably *before* starting to try and understand it, along with the general question 'what do the numbers, or symbols, or categories mean?'

So far, in this book the terms descriptor, parameter, and property have been used interchangeably. This can perhaps be justified in that it helps to avoid repetition, but they do actually mean different things and so it would be best to define them here. Descriptor refers to any means by which a sample (case, object) is described or characterized: for molecules the term aromatic, for example, is a descriptor, as are the quantities molecular weight and boiling point. Physicochemical property refers to a feature of a molecule which is determined by its physical or chemical properties, or a combination of both. Parameter is a term which is used

[2] Molecular design means the design of a biologically active substance such as a pharmaceutical or pesticide, or of a 'performance' chemical such as a fragrance, flavour, and so on or a formulation such as paint, adhesive, etc.

to refer to some numerical measure of a descriptor or physicochemical property. The two descriptors molecular weight and boiling point are also both parameters; the term aromatic is a descriptor but not a parameter, whereas the question 'How many aromatic rings?' gives rise to a parameter. All parameters are thus descriptors but not vice versa.

The next few sections discuss some of the more important aspects of the nature and properties of data. It is often the data itself that dictates which particular analytical method may be used to examine it and how successful the outcome of that examination will be.

### 1.4.1 Types of Data and Scales of Measurement

In the examples of descriptors and parameters given here it may have been noticed that there are differences in the 'nature' of the values used to express them. This is because variables, both dependent and independent, can be classified as *qualitative* or *quantitative*. Qualitative variables contain data that can be placed into distinct classes; 'dead' or 'alive', for example, 'hot' or 'cold', 'aromatic' or 'non-aromatic' are examples of binary or dichotomous qualitative variables. Quantitative variables contain data that is numerical and can be ranked or ordered. Examples of quantitative variables are length, temperature, age, weight, etc. Quantitative variables can be further divided into discrete or continuous. Discrete variables are usually counts such as 'how many objects in a group', 'number of hydroxyl groups', 'number of components in a mixture', and so on. Continuous variables, such as height, time, volume, etc. can assume any value within a given range.

In addition to the classification of variables as qualitative/quantitative and the further division into discrete/continuous, variables can also be classified according to how they are categorized, counted or measured. This is because of differences in the scales of measurement used for variables. It is necessary to consider four different scales of measurement: nominal, ordinal, interval, and ratio. It is important to be aware of the properties of these scales since the nature of the scales determines which analytical methods should be used to treat the data.

#### Nominal

This is the weakest level of measurement, i.e. has the lowest information content, and applies to the situation where a number or other symbol

is used to assign membership to a class. The terms male and female, young and old, aromatic and non-aromatic are all descriptors based on nominal scales. These are dichotomous descriptors, in that the objects (people or compounds) belong to one class or another, but this is not the only type of nominal descriptor. Colour, subdivided into as many classes as desired, is a nominal descriptor as is the question 'which of the four halogens does the compound contain?'

### Ordinal

Like the nominal scale, the ordinal scale of measurement places objects in different classes but here the classes bear some relation to one another, expressed by the term greater than (>). Thus, from the previous example, old > middle-aged > young. Two examples in the context of molecular design are toxic > slightly toxic > nontoxic, and fully saturated > partially saturated > unsaturated. The latter descriptor might also be represented by the number of double bonds present in the structures although this is not chemically equivalent since triple bonds are ignored. It is important to be aware of the situations in which a parameter might appear to be measured on an interval or ratio scale (see below), but because of the distribution of compounds in the set under study, these effectively become nominal or ordinal descriptors (see next section).

### Interval

An interval scale has the characteristics of a nominal scale, but in addition the distances between any two numbers on the scale are of known size. The zero point and the units of measurement of an interval scale are arbitrary: a good example of an interval scale parameter is boiling point. This could be measured on either the Fahrenheit or Celsius temperature scales but the information content of the boiling point values is the same.

### Ratio

A ratio scale is an interval scale which has a true zero point as its origin. Mass is an example of a parameter measured on a ratio scale, as are parameters which describe dimensions such as length, volume, etc. An additional property of the ratio scale, hinted at in the name, is that it

contains a true ratio between values. A measurement of 200 for one sample and 100 for another, for example, means a ratio of 2:1 between these two samples.

What is the significance of these different scales of measurement? As will be discussed later, many of the well-known statistical methods are parametric, that is, they rely on assumptions concerning the distribution of the data. The computation of parametric tests involves arithmetic manipulation such as addition, multiplication, and division, and this should only be carried out on data measured on interval or ratio scales. When these procedures are used on data measured on other scales they introduce distortions into the data and thus cast doubt on any conclusions which may be drawn from the tests. Nonparametric or 'distribution-free' methods, on the other hand, concentrate on an order or ranking of data and thus can be used with ordinal data. Some of the nonparametric techniques are also designed to operate with classified (nominal) data. Since interval and ratio scales of measurement have all the properties of ordinal scales it is possible to use nonparametric methods for data measured on these scales. Thus, the distribution-free techniques are the 'safest' to use since they can be applied to most types of data. If, however, the data does conform to the distributional assumptions of the parametric techniques, these methods may well extract more information from the data.

### 1.4.2 Data Distribution

Statistics is often concerned with the treatment of a small[3] number of samples which have been drawn from a much larger population. Each of these samples may be described by one or more variables which have been measured or calculated for that sample. For each variable there exists a population of samples. It is the properties of these populations of variables that allows the assignment of probabilities, for example, the likelihood that the value of a variable will fall into a particular range, and the assessment of significance (i.e. is one number significantly different from another). Probability theory and statistics are, in fact, separate subjects; each may be said to be the inverse of the other, but for the purposes of this discussion they may be regarded as doing the same job.

[3] The term 'small' here may represent hundreds or even thousands of samples. This is a small number compared to a population which is often taken to be infinite.

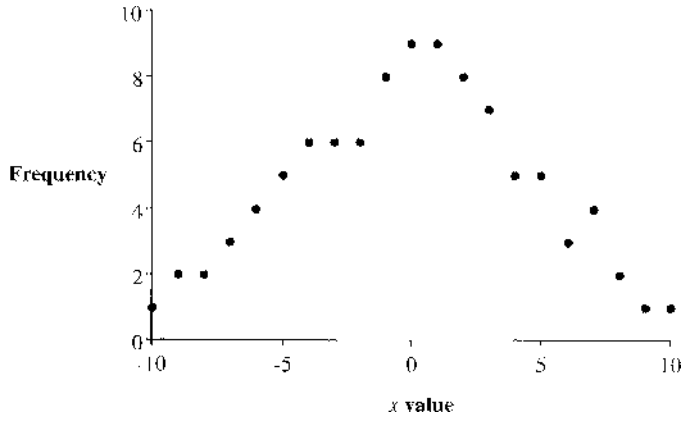


**Figure 1.3** Frequency distribution for the variable x over the range −10 to +10.

How are the properties of the population used? Perhaps one of the most familiar concepts in statistics is the frequency distribution. A plot of a frequency distribution is shown in Figure 1.3, where the ordinate ($y$-axis) represents the number of occurrences of a particular value of a variable given by the scales of the abscissa ($x$-axis).

If the data is discrete, usually but not necessarily measured on nominal or ordinal scales, then the variable values can only correspond to the points marked on the scale on the abscissa. If the data is continuous, a problem arises in the creation of a frequency distribution, since every value in the data set may be different and the resultant plot would be a very uninteresting straight line at $y = 1$. This may be overcome by taking ranges of the variable and counting the number of occurrences of values within each range. For the example shown in Figure 1.4 (where there are a total of 50 values in all), the ranges are 0–1, 1–2, 2–3, and so on up to 9–10.

It can be seen that these points fall on a roughly bell-shaped curve with the largest number of occurrences of the variable occurring around the peak of the curve, corresponding to the mean of the set. The mean of the sample is given the symbol $\overline{X}$ and is obtained by summing all the sample values together and dividing by the number of samples as shown in Equation (1.1).

$$\overline{X} = \frac{x_1 + x_2 + x_3 + \ldots\ldots x_n}{n} = \frac{\sum x}{n} \tag{1.1}$$

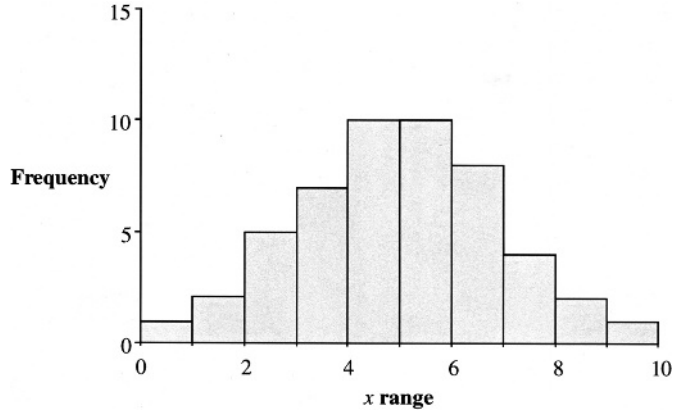


**Figure 1.4** Frequency histogram for the continuous variable x over the range 0 to +10.

The mean, since it is derived from a sample, is known as a *statistic*. The corresponding value for a population, the population mean, is given the symbol $\mu$ and this is known as a *parameter*, another use for the term. A convention in statistics is that Greek letters are used to denote parameters (measures or characteristics of the population) and Roman letters are used for statistics. The mean is known as a 'measure of central tendency' (others are the mode, median and midrange) which means that it gives some idea of the centre of the distribution of the values of the variable. In addition to knowing the centre of the distribution it is important to know how the data values are spread through the distribution. Are they clustered around the mean or do they spread evenly throughout the distribution? Measures of distribution are often known as 'measures of dispersion' and the most often used are variance and standard deviation. Variance is the average of the squares of the distance of each data value from the mean as shown in Equation (1.2):

$$s^2 = \frac{\sum (X - \overline{X})^2}{n - 1} \tag{1.2}$$

The symbol used for the sample variance is $s^2$ which at first sight might appear strange. Why use the square sign in a symbol for a quantity like this? The reason is that the standard deviation (s) of a sample is the square root of the variance. The standard deviation has the same units as the units of the original variable whereas the variance has units that are the square of the original units. Another odd thing might be noticed

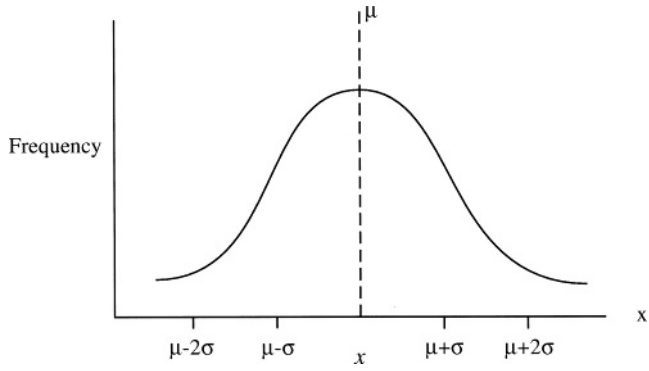


**Figure 1.5** Probability distribution for a very large number of values of the variable x; $\mu$ equals the mean of the set and $\sigma$ the standard deviation.

about Equation (1.2) and that is the use of $n - 1$ in the denominator. When calculating the mean the summation (Equation (1.1)) is divided by the number of data points, $n$, so why is $n - 1$ used here? The reason for this, apparently, is that the variance computed using $n$ usually underestimates the population variance and thus the summation is divided by $n - 1$ giving a slightly larger value. The corresponding symbols for the population parameters are $\sigma^2$ for the variance and $\sigma$ for the standard deviation. A graphical illustration of the meaning of $\mu$ and $\sigma$ is shown in Figure 1.5, which is a frequency distribution like Figures 1.3 and 1.4 but with more data values so that we obtain a smooth curve.

The figure shows that $\mu$ is located in the centre of the distribution, as expected, and that the values of the variable $x$ along the abscissa have been replaced by the mean +/− multiples of the standard deviation. This is because there is a theorem (Chebyshev's) which specifies the proportions of the spread of values in terms of the standard deviation, there is more on this later.

It is at this point that we can see a link between statistics and probability theory. If the height of the curve is standardized so that the area underneath it is unity, the graph is called a probability curve. The height of the curve at some point $x$ can be denoted by $f(x)$ which is called the probability density function (p.d.f.). This function is such that it satisfies the condition that the area under the curve is unity

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1 \tag{1.3}$$

This now allows us to find the probability that a value of $x$ will fall in any given range by finding the integral of the p.d.f. over that range:

$$probability\ (x_1 < x < x_2) = \int_{x_1}^{x_2} f(x)dx \tag{1.4}$$

This brief and rather incomplete description of frequency distributions and their relationship to probability distribution has been for the purpose of introducing the normal distribution curve. The normal or Gaussian distribution is the most important of the distributions that are considered in statistics. The height of a normal distribution curve is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{1.5}$$

This rather complicated function was chosen so that the total area under the curve is equal to 1 for all values of $\mu$ and $\sigma$. Equation (1.5) has been given so that the connection between probability and the two parameters $\mu$ and $\sigma$ of the distribution can be seen. The curve is shown in Figure 1.5 where the abscissa is marked in units of $\sigma$. It can be seen that the curve is symmetric about $\mu$, the mean, which is a measure of the *location* or 'central tendency' of the distribution. As mentioned earlier, there is a theorem that specifies the proportion of the spread of values in any distribution. In the special case of the normal distribution this means that approximately 68 % of the data values will fall within 1 standard deviation of the mean and 95 % within 2 standard deviations. Put another way, about one observation in three will lie more than one standard deviation ($\sigma$) from the mean and about one observation in 20 will lie more than two standard deviations from the mean. The standard deviation is a measure of the *spread* or 'dispersion'; it is these two properties, location and spread, of a distribution which allow us to make estimates of likelihood (or 'significance').

Some other features of the normal distribution can be seen by consideration of Figure 1.6. In part (a) of the figure, the distribution is no longer symmetrical; there are more values of the variable with a higher value.

This distribution is said to be skewed, it has a positive skewness; the distribution shown in part (b) is said to be negatively skewed. In part (c) three distributions are overlaid which have differing degrees of 'steepness' of the curve around the mean. The statistical term used

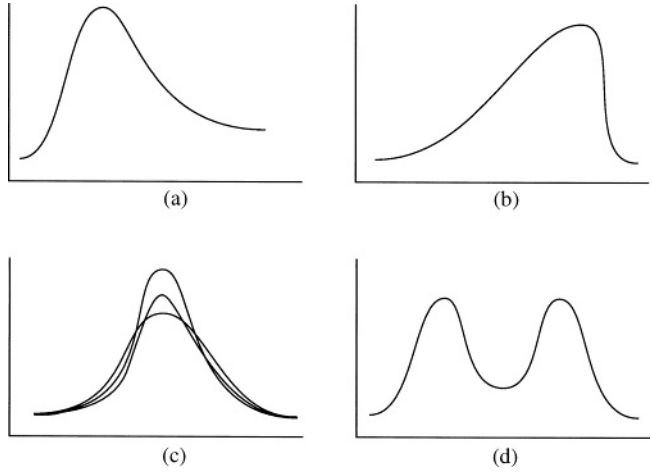


**Figure 1.6** Illustration of deviations of probability distributions from a normal distribution.

to describe the steepness, or degree of peakedness, of a distribution is *kurtosis*. Various measures may be used to express kurtosis; one known as the *moment ratio* gives a value of three for a normal distribution. Thus it is possible to judge how far a distribution deviates from normality by calculating values of skewness (= 0 for a normal distribution) and kurtosis. As will be seen later, these measures of how 'well behaved' a variable is may be used as an aid to variable selection. Finally, in part (d) of Figure 1.6 it can be seen that the distribution appears to have two means. This is known as a *bimodal* distribution, which has its own particular set of properties distinct to those of the normal distribution.

### 1.4.3 Deviations in Distribution

There are many situations in which a variable that might be expected to have a normal distribution does not. Take for example the molecular weight of a set of assorted painkillers. If the compounds in the set consisted of aspirin and morphine derivatives, then we might see a bimodal distribution with two peaks corresponding to values of around 180 (mol.wt. of aspirin) and 285 (mol.wt. of morphine). Skewed and kurtosed distributions may arise for a variety of reasons, and the effect they will have on an analysis depends on the assumptions employed in the analysis and the degree to which the distributions deviate from

normality, or whatever distribution is assumed. This, of course, is not a very satisfactory statement to someone who is asking the question, 'Is my data good enough (sufficiently well behaved) to apply such and such a method to it?' Unfortunately, there is not usually a simple answer to this sort of question. In general, the further the data deviates from the type of distribution that is assumed when a model is fitted, the less reliable will be the conclusions drawn from that model. It is worth pointing out here that real data is unlikely to conform perfectly to a normal distribution, or any other 'standard' distribution for that matter. Checking the distribution is necessary so that we know what type of method can be used to treat the data, as described later, and how reliable any estimates will be which are based on assumptions of distribution. A caution should be sounded here in that it is easy to become too critical and use a poor or less than 'perfect' distribution as an excuse not to use a particular technique, or to discount the results of an analysis.

Another problem which is frequently encountered in the distribution of data is the presence of outliers. Consider the data shown in Table 1.1 where calculated values of electrophilic superdelocalizability (ESDL10) are given for a set of analogues of antimycin $A_1$, compounds which kill human parasitic worms, *Dipetalonema vitae*.

The mean and standard deviation of this variable give no clues as to how well it is distributed and the skewness and kurtosis values of −3.15

**Table 1.1** Physicochemical properties and antifilarial activity of antimycin analogues (reproduced from ref. [3] with permission from American Chemical Society).

| Compound number | ESDL10 | Calculated log *P* | Melting point °C | Activity |
|---|---|---|---|---|
| 1 | −0.3896 | 7.239 | 81 | −0.845 |
| 2 | −0.4706 | 5.960 | 183 | −0.380 |
| 3 | −0.4688 | 6.994 | 207 | 1.398 |
| 4 | −0.4129 | 7.372 | 143 | 0.319 |
| 5 | −0.3762 | 5.730 | 165 | −0.875 |
| 6 | −0.3280 | 6.994 | 192 | 0.824 |
| 7 | −0.3649 | 6.755 | 256 | 1.839 |
| 8 | −0.5404 | 6.695 | 199 | 1.020 |
| 9 | −0.4499 | 7.372 | 151 | 0.420 |
| 10 | −0.3473 | 5.670 | 195 | 0.000 |
| 11 | −0.7942 | 4.888 | 212 | 0.097 |
| 12 | −0.4057 | 6.205 | 246 | 1.130 |
| 13 | −0.4094 | 6.113 | 208 | 0.920 |
| 14 | −1.4855 | 6.180 | 159 | 0.770 |
| 15 | −0.3427 | 5.681 | 178 | 0.301 |
| 16 | −0.4597 | 6.838 | 222 | 1.357 |

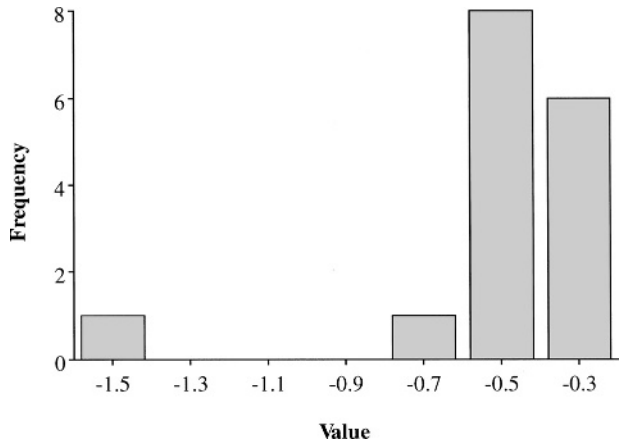


**Figure 1.7** Frequency distribution for the variable ESDL10 given in Table 1.1.

and 10.65 respectively might not suggest that it deviates too seriously from normal. A frequency distribution for this variable, however, reveals the presence of a single extreme value (compound 14) as shown in Figure 1.7.

This data was analysed by multiple linear regression (discussed further in Chapter 6), which is a method based on properties of the normal distribution. The presence of this outlier had quite profound effects on the analysis, which could have been avoided if the data distribution had been checked at the outset (particularly by the present author). Outliers can be very informative and should not simply be discarded as so frequently happens. If an outlier is found in one of the descriptor variables (physicochemical data), then it may show that a mistake has been made in the measurement or calculation of that variable for that compound. In the case of properties derived from computational chemistry calculations it may indicate that some basic assumption has been violated or that the particular method employed was not appropriate for that compound. An example of this can be found in semi-empirical molecular orbital methods which are only parameterized for a limited set of the elements. Outliers are not always due to mistakes, however. Consider the calculation of electrostatic potential around a molecule. It is easy to identify regions of high and low values, and these are often used to provide criteria for alignment or as a pictorial explanation of biological properties. The value of an electrostatic potential minimum or maximum, or the value of the potential at a given point, has been used as a parameter to describe sets of molecules. This is fine as long as each

molecule in the set has a maximum and/or minimum at approximately the same place. Problems arise if a small number of the structures do not have the corresponding values in which case they will form outliers. The effect of this is to cause the variable, apparently measured on an interval scale, to become a nominal descriptor. Take, for example, the case where 80 % of the members of the set have an electrostatic potential minimum of around −50 kcal/mole at a particular position. For the remaining members of the set, the electrostatic potential at this position is zero. This variable has now become an 'indicator' variable which has two distinct values (zero for 20 % of the molecules and −50 for the remainder) that identify two different subsets of the data. The problem may be overcome if the magnitude of a minimum or maximum is taken, irrespective of position, although problems may occur with molecules that have multiple minima or maxima. There is also the more difficult philosophical question of what do such values 'mean'.

When outliers occur in the biological or dependent data, they may also indicate mistakes: perhaps the wrong compound was tested, or it did not dissolve, a result was misrecorded, or the test did not work out as expected. However, in dependent data sets, outliers may be even more informative. They may indicate a change in biological mechanism, or perhaps they demonstrate that some important structural feature has been altered or a critical value of a physicochemical property exceeded. Once again, it is best not to simply discard such outliers, they may be very informative.

Is there anything that can be done to improve a poorly distributed variable? The answer is yes, but it is a qualified yes since the use of too many 'tricks' to improve distribution may introduce other distortions which will obscure useful patterns in the data. The first step in improving distribution is to identify outliers and then, if possible, identify the cause(s) of such outliers. If an outlier cannot be 'fixed' it may need to be removed from the data set. The second step involves the consideration of the rest of the values in the set. If a variable has a high value of kurtosis or skewness, is there some good reason for this? Does the variable really measure what we think it does? Are the calculations/measurements sound for all of the members of the set, particularly at the extremes of the range for skewed distributions or around the mean where kurtosis is a problem. Finally, would a transformation help? Taking the logarithm of a variable will often make it behave more like a normally distributed variable, but this is not a justification for always taking logs!

A final point on the matter of data distribution concerns the non-parametric methods. Although these techniques are not based on

distributional assumptions, they may still suffer from the effects of 'strange' distributions in the data. The presence of outliers or the effective conversion of interval to ordinal data, as in the electrostatic potential example, may lead to misleading results.

## 1.5 ANALYTICAL METHODS

This whole book is concerned with analytical methods, as the following chapters will show, so the purpose of this section is to introduce and explain some of the terms which are used to describe the techniques. These terms, like most jargon, also often serve to obscure the methodology to the casual or novice user so it is hoped that this section will help to unveil the techniques.

First, we should consider some of the expressions which are used to describe the methods in general. *Biometrics* is a term which has been used since the early 20th century to describe the development of mathematical and statistical methods to data analysis problems in the biological sciences. *Chemometrics* is used to describe 'any mathematical or statistical procedure which is used to analyse chemical data' [4]. Thus, the simple act of plotting a calibration curve is chemometrics, as is the process of fitting a line to that plot by the method of least squares, as is the analysis by principal components of the spectrum of a solution containing several species. Any chemist who carries out quantitative experiments is also a chemometrician! *Univariate statistics* is (perhaps unsurprisingly) the term given to describe the statistical analysis of a single variable. This is the type of statistics which is normally taught on an introductory course; it involves the analysis of variance of a single variable to give quantities such as the mean and standard deviation, and some measures of the distribution of the data. *Multivariate statistics* describes the application of statistical methods to more than one variable at a time, and is perhaps more useful than univariate methods since most problems in real life are multivariate. We might more correctly use the term *multivariate analysis* since not all multivariate methods are statistical. Chemometrics and multivariate analysis refer to more or less the same things, chemometrics being the broader term since it includes univariate techniques.[4]

*Pattern recognition* is the name given to any method which helps to reveal the patterns within a data set. A definition of pattern recognition is that it 'seeks similarities and regularities present in the data'. Some

[4] But, of course, it is restricted to chemical problems.

**Table 1.2** Anaesthetic activity and hydrophobicity of a series of alcohols (reproduced from ref. [5] with permission from American Society for Pharmacology and Experimental Therapeutics (ASPET)).

| Alcohol | $\Sigma\pi$ | Anaesthetic activity $(\log 1/C)$ |
|---|---|---|
| $C_2H_5OH$ | 1.0 | 0.481 |
| *n*–$C_3H_7OH$ | 1.5 | 0.959 |
| *n*–$C_4H_9OH$ | 2.0 | 1.523 |
| *n*–$C_5H_{11}OH$ | 2.5 | 2.152 |
| *n*–$C_7H_{15}OH$ | 3.5 | 3.420 |
| *n*–$C_8H_{17}OH$ | 4.0 | 3.886 |
| *n*–$C_9H_{19}OH$ | 4.5 | 4.602 |
| *n*–$C_{10}H_{21}OH$ | 5.0 | 5.00 |
| *n*–$C_{11}H_{23}OH$ | 5.5 | 5.301 |
| *n*–$C_{12}H_{25}OH$ | 6.0 | 5.124 |

of the display techniques described in Chapter 4 are quite obvious examples of pattern recognition since they result in a visual display of the patterns in data. However, consider the data shown in Table 1.2 where the anaesthetic activity of a series of alcohols is given as the logarithm of the reciprocal of the concentration needed to induce a particular level of anaesthesia.

The other column in this table ($\Sigma\pi$) is a measure of the hydrophobicity of each of the alcohols. Hydrophobicity, which means literally 'water hating', reflects the tendency of molecules to partition into membranes in a biological system (see Chapter 10 for more detail) and is a physicochemical descriptor of the alcohols. Inspection of the table reveals a fairly obvious relationship between log 1/C and $\Sigma\pi$ but this is most easily seen by a plot as shown in Figure 1.8.

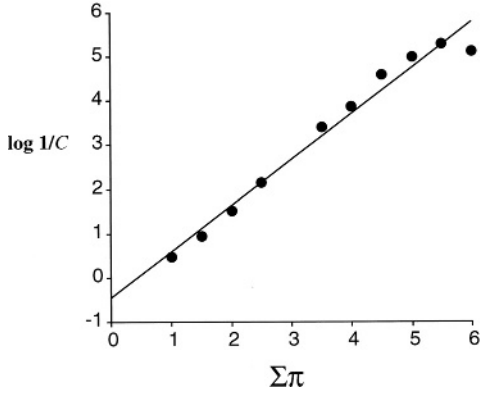


**Figure 1.8** Plot of biological response (log 1/C) against $\Sigma\pi$ (from Table 1.2).

This relationship can be expressed in a very concise form as shown in Equation (1.6):

$$\log {}^{1}/_{C} = 1.039 \sum \pi - 0.442 \tag{1.6}$$

This is an example of a simple linear regression equation. Regression equations and the statistics which may be used to describe their 'goodness of fit', to a linear or other model, are explained in detail in Chapter 6. For the purposes of demonstrating this relationship it is sufficient to say that the values of the logarithm of a reciprocal concentration (log $1/C$) in Equation (1.6) are obtained by multiplication of the $\Sigma\pi$ values by a coefficient (1.039) and the addition of a constant term ($-0.442$). The equation is shown in graphical form (Figure 1.8); the slope of the fitted line is equal to the regression coefficient (1.039) and the intercept of the line with the zero point of the $x$-axis is equal to the constant ($-0.442$). Thus, the pattern obvious in the data table may be shown by the simple bivariate plot and expressed numerically in Equation (1.6). These are examples of pattern recognition although regression models would not normally be classed as pattern recognition methods.

Pattern recognition and chemometrics are more or less synonymous. Some of the pattern recognition techniques are derived from research into artificial intelligence. We can 'borrow' some useful jargon from this field which is related to the concept of 'training' an algorithm or device to carry out a particular task. Suppose that we have a set of data which describes a collection of compounds which can be classified as active or inactive in some biological test. The descriptor data, or independent variables, may be whole molecule parameters such as melting point, or may be substituent constants, or may be calculated quantities such as molecular orbital energies. One simple way in which this data may be analysed is to compare the values of the variables for the active compounds with those of the inactives (see discriminant analysis in Chapter 7). This may enable one to establish a rule or rules which will distinguish the two classes. For example, all the actives may have melting points above 250 °C and/or may have highest occupied molecular orbital (HOMO) energy values below $-10.5$. The production of these rules, by inspection of the data or by use of an algorithm, is called *supervised learning* since knowledge of class membership was used to generate them. The dependent variable, in this case membership of the active or inactive class, is used in the learning or training process. *Unsupervised learning*, on the other hand, does not make use of a dependent variable.

An example of unsupervised learning for this data set might be to plot the values of two of the descriptor variables against one another. Class membership for the compounds could then be marked on the plot and a pattern may be seen to emerge from the data. If we chose melting point and HOMO as the two variables to plot, we may see a grouping of the active compounds where HOMO $< -10.5$ and melting point $>250\ ^{\circ}C$.

The distinction between supervised and unsupervised learning may seem unimportant but there is a significant philosophical difference between the two. When we seek a rule to classify data, there is a possibility that any apparent rule may happen by chance. It may, for example, be a coincidence that all the active compounds have high melting points; in such a case the rule will not be predictive. This may be misleading, embarrassing, expensive, or all three! Chance effects may also occur with unsupervised learning but are much less likely since unsupervised learning does not seek to generate rules. Chance effects are discussed in more detail in Chapters 6 and 7. The concept of learning may also be used to define some data sets. A set of compounds which have already been tested in some biological system, or which are about to be tested, is known as a *learning* or *training set*. In the case of a supervised learning method this data will be used to train the technique but this term applies equally well to the unsupervised case. Judicious choice of the training set will have profound effects on the success of the application of any analytical method, supervised or unsupervised, since the information contained in this set dictates the information that can be extracted (see Chapter 2). A set of untested or yet to be synthesized compounds is called a *test set*, the objective of data analysis usually being to make predictions for the test set (also sometimes called a *prediction set*). A further type of data set, known as an *evaluation set*, may also be used. This consists of a set of compounds for which test results are available but which is not used in the construction of the model. Examination of the prediction results for an evaluation set can give some insight into the validity and accuracy of the model.

Finally we should define the terms *parametric* and *nonparametric*. A measure of the distribution of a variable (see Section 1.4.2) is a measure of one of the parameters of that variable. If we had measurements for all possible values of a variable (an infinite number of measurements), then we would be able to compute a value for the population distribution. Statistics is concerned with a much smaller set of measurements which forms a sample of that population and for which we can calculate a sample distribution. A well-known example of this is the Gaussian or normal distribution. One of the assumptions made in statistics is that a sample

distribution, which we can measure, will behave like a population distribution which we cannot. Although population distributions cannot be measured, some of their properties can be predicted by theory. Many statistical methods are based on the properties of population distributions, particularly the normal distribution. These are called *parametric techniques* since they make use of the distribution parameter. Before using a parametric method, the distribution of the variables involved should be calculated. This is very often ignored, although fortunately many of the techniques based on assumptions about the normal distribution are quite robust to departures from normality. There are also techniques which do not rely on the properties of a distribution, and these are known as *nonparametric* or '*distribution free*' methods.

## 1.6 SUMMARY

In this chapter the following points were covered:

1. dependent and independent variables and how data tables are laid out;
2. where data comes from and some of its properties;
3. descriptors, parameters and properties;
4. nominal, ordinal, interval and ratio scales;
5. frequency distributions, the normal distribution, definition and explanation of mean, variance and standard deviation. skewness and kurtosis;
6. the difference between sample and population properties;
7. factors causing deviations in distribution;
8. terminology – univariate and multivariate statistics, chemometrics and biometrics, pattern recognition, supervised and unsupervised learning. Training, test and evaluation sets, parametric and nonparametric or 'distribution free' techniques.

## REFERENCES

[1] Livingstone, D.J. (2000). 'The Characterization of Chemical Structures Using Molecular Properties. A Survey', *Journal of Chemical Information and Computer Science*, **40**, 195–209.

[2] Livingstone, D.J. (2003). 'Theoretical Property Predictions', *Current Topics in Medicinal Chemistry*, **3**, 1171–92.

[3] Selwood, D.L, Livingstone, D.J., Comley, J.C.W. *et al.* (1990). *Journal of Medicinal Chemistry*, **33**, 136–42.
[4] Kowalski, B., Brown, S., and Van de Ginste, B. (1987). *Journal of Chemometrics*, **1**, 1–2.
[5] Hansch, C., Steward, A. R., Iwasa, J., and Deutsch, E.W. (1965). *Molecular Pharmacology*, **1**, 205–13.