A Practical Guide to Scientific Data Analysis

David Livingstone
ChemQuest, Sandown, Isle of Wight, UK

WILEY
A John Wiley and Sons, Ltd., Publication

A Practical Guide to Scientific Data Analysis

A Practical Guide to Scientific Data Analysis

David Livingstone
ChemQuest, Sandown, Isle of Wight, UK

WILEY
A John Wiley and Sons, Ltd., Publication

This edition first published 2009

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Livingstone, D. (David)
A practical guide to scientific data analysis / David Livingstone.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-85153-1 (cloth : alk. paper)
1. QSAR (Biochemistry) – Statistical methods. 2. Biochemistry – Statistical methods.
I. Title.
QP517.S85L554 2009
615′.1900727–dc22

2009025910

A catalogue record for this book is available from the British Library.

ISBN 978-0470-851531

Typeset in 10.5/13pt Sabon by Aptara Inc., New Delhi, India.
Printed and bound in Great Britain by TJ International, Padstow, Corwall

This book is dedicated to the memory of my first wife, Cherry (18/5/52–1/8/05), who inspired me, encouraged me and helped me in everything I've done, and to the memory of Rifleman Jamie Gunn (4/8/87–25/2/09), whom we both loved very much and who was killed in action in Helmand province, Afghanistan.

Contents

Preface

The idea for this book came in part from teaching quantitative drug design to B.Sc. and M.Sc. students at the Universities of Sussex and Portsmouth. I have also needed to describe a number of mathematical and statistical methods to my friends and colleagues in medicinal (and physical) chemistry, biochemistry, and pharmacology departments at Wellcome Research and SmithKline Beecham Pharmaceuticals. I have looked for a textbook which I could recommend which gives *practical* guidance in the use and interpretation of the apparently esoteric methods of multivariate statistics, otherwise known as pattern recognition. I would have found such a book useful when I was learning the trade, and so this is intended to be that sort of guide.

There are, of course, many fine textbooks of statistics and these are referred to as appropriate for further reading. However, I feel that there isn't a book which gives a practical guide for scientists to the processes of data analysis. The emphasis here is on the application of the techniques and the interpretation of their results, although a certain amount of theory is required in order to explain the methods. This is not intended to be a statistical textbook, indeed an elementary knowledge of statistics is assumed of the reader, but is meant to be a statistical companion to the novice or casual user.

It is necessary here to consider the type of research which these methods may be used for. Historically, techniques for building models to relate biological properties to chemical structure have been developed in pharmaceutical and agrochemical research. Many of the examples used in this text are derived from these fields of work. There is no reason, however, why any sort of property which depends on chemical structure should not be modelled in this way. This might be termed quantitative structure–property relationships (QSPR) rather than QSAR where

A stands for activity. Such models are beginning to be reported; recent examples include applications in the design of dyestuffs, cosmetics, egg-white substitutes, artificial sweeteners, cheese-making, and prepared food products. I have tried to incorporate some of these applications to illustrate the methods, as well as the more traditional examples of QSAR.

There are also many other areas of science which can benefit from the application of statistical and mathematical methods to an examination of their data, particularly multivariate techniques. I hope that scientists from these other disciplines will be able to see how such approaches can be of use in their own work.

The chapters are ordered in a logical sequence, the sequence in which data analysis might be carried out – from planning an experiment through examining and displaying the data to constructing quantitative models. However, each chapter is intended to stand alone so that casual users can refer to the section that is most appropriate to their problem. The one exception to this is the Introduction which explains many of the terms which are used later in the book. Finally, I have included definitions and descriptions of some of the chemical properties and biological terms used in panels separated from the rest of the text. Thus, a reader who is already familiar with such concepts should be able to read the book without undue interruption.

David Livingstone
Sandown, Isle of Wight
May 2009

Abbreviations

π	hydrophobicity substituent constant
σ	electronic substituent constant
Λ_{alk}	hydrogen-bonding capability parameter
ΔH	enthalpy
AI	artificial intelligence
ANN	artificial neural networks
ANOVA	analysis of variance
BPN	back-propagation neural network
CA	cluster analysis
CAMEO	Computer Assisted Mechanistic Evaluation of Organic reactions
CASE	Computer Assisted Structure Evaluation
CCA	canonical correlation analysis
CoMFA	Comparative Molecular Field Analysis
CONCORD	CONnection table to CoORDinates
CR	continuum regression
CSA	cluster significance analysis
DEREK	Deductive Estimation of Risk from Existing Knowledge
ED_{50}	dose to give 50 % effect
ESDL10	electrophilic superdelocalizability
ESS	explained sum of squares
FA	factor analysis
FOSSIL	Frame Orientated System for Spectroscopic Inductive Learning
GABA	γ-aminobutyric acid
GC-MS	gas chromatography-mass spectrometry
HOMO	highest occupied molecular orbital
HPLC	high-performance liquid chromatography

HTS	high throughput screening
I_{50}	concentration for 50 % inhibition
IC_{50}	concentration for 50 % inhibition
ID3	iterative dichotomizer three
IR	infrared
K_m	Michaelis–Menten constant
KNN	*k*-nearest neighbour technique
LC_{50}	concentration for 50 % lethal effect
LD_{50}	dose for 50 % death
LDA	linear discriminant analysis
LLM	linear learning machine
log*P*	logarithm of a partition coefficient
LOO	leave one out at a time
LV	latent variable
m.p.	melting point
MAO	monoamine oxidase
MIC	minimum inhibitory concentration
MLR	multiple linear regression
mol.wt.	molecular weight
MR	molar refractivity
MSD	mean squared distance
MSE	explained mean square
MSR	residual mean square
MTC	minimum threshold concentration
NLM	nonlinear mapping
NMR	nuclear magnetic resonance
NOA	natural orange aroma
NTP	National Toxicology Program
OLS	ordinary least square
PC	principal component
PCA	principal component analysis
PCR	principal component regression
p.d.f.	probability density function
pI_{50}	negative log of the concentration for 50 % inhibition
PLS	partial least squares
PRESS	predicted residual sum of squares
QDA	quantitative descriptive analysis
QSAR	quantitative structure-activity relationship
QSPR	quantitative structure-property relationship
R^2	multiple correlation coefficient
ReNDeR	Reversible Non-linear Dimension Reduction

RMSEP	root mean square error of prediction
RSS	residual or unexplained sum of squares
SE	standard error
SAR	structure-activity relationships
SIMCA	see footnote p. 195
SMA	spectral map analysis
SMILES	Simplified Molecular Input Line Entry System
SOM	self organising map
TD_{50}	dose for 50 % toxic effect
TOPKAT	Toxicity Prediction by Komputer Assisted Technology
TS	taboo search
TSD	total squared distance
TSS	total sum of squares
UFS	unsupervised forward selection
UHTS	ultra high throughput screening
UV	ultraviolet spectrophotometry
V_m	Van der Waals' volume

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.

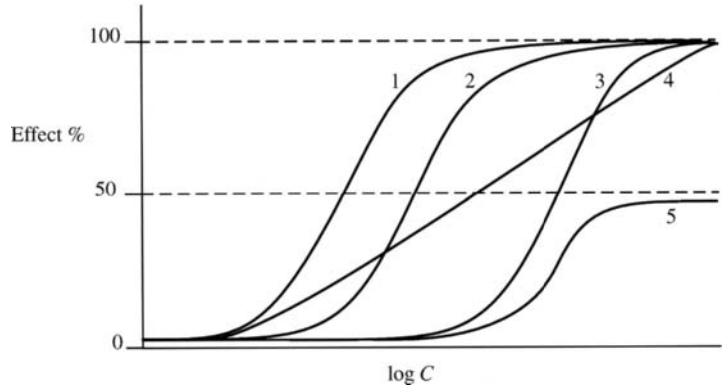

Figure 1.2 Typical and not so typical dose–response curves for a set of five different compounds.

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.

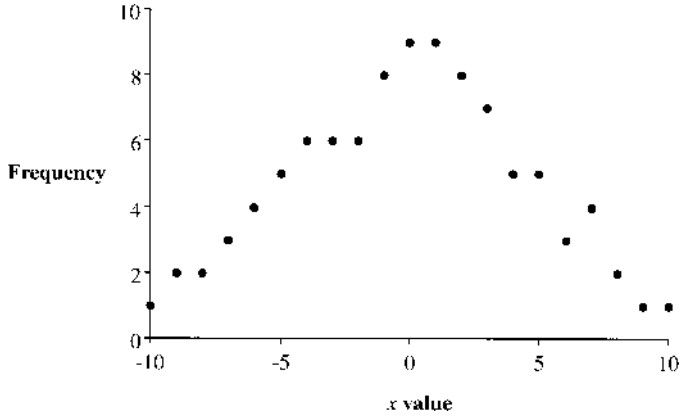

Figure 1.3 Frequency distribution for the variable x over the range −10 to +10.

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).

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

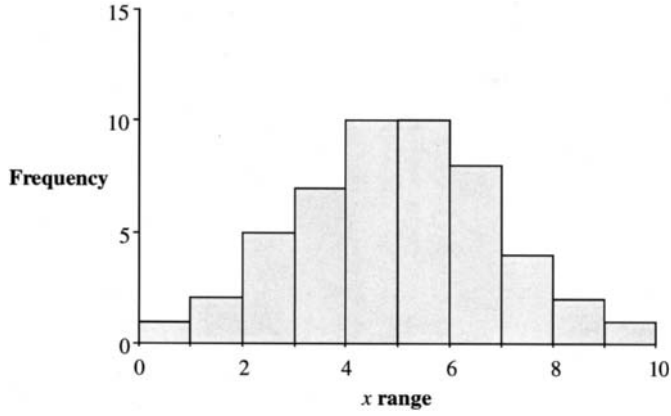

Figure 1.4 Frequency histogram for the continuous variable x over the range 0 to +10.

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 μ 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