PART A

BASICS

Statistics is the science of collecting, summarizing, presenting and interpreting data, and of using them to estimate the magnitude of associations and test hypotheses. It has a central role in medical investigations. Not only does it provide a way of organizing information on a wider and more formal basis than relying on the exchange of anecdotes and personal experience, it takes into account the intrinsic variation inherent in most biological processes. For example, not only does blood pressure differ from person to person, but in the same person it also varies from day to day and from hour to hour. It is the interpretation of data in the presence of such variability that lies at the heart of statistics. Thus, in investigating morbidity associated with a particular stressful occupation, statistical methods would be needed to assess whether an observed average blood pressure above that of the general population could simply be due to chance variations or whether it represents a real indication of an occupational health risk.

Variability can also arise *unpredictably* (randomly) within a population. Individuals do not all react in the same way to a given stimulus. Thus, although smoking and heavy drinking are in general bad for the health, we may hear of a heavy smoker and drinker living to healthy old age, whereas a non-smoking teetotaller may die young. As another example, consider the evaluation of a new vaccine. Individuals vary both in their responsiveness to vaccines and in their susceptibility and exposure to disease. Not only will some people who are unvaccinated escape infection, but also a number of those who are vaccinated may contract the disease. What can be concluded if the proportion of people free from the disease is greater among the vaccinated group than among the unvaccinated? How effective is the vaccine? Could the apparent effect just be due to chance? Or, was there some bias in the way people were selected for vaccination, for example were they of different ages or social class, such that their baseline risk of contracting the disease was already lower than those selected into the non-vaccinated group? The methods of statistical analysis are used to address the first two of these questions, while the choice of an appropriate design should exclude the third. This example illustrates that the usefulness of statistics is not confined to the analysis of results. It also has a role to play in the design and conduct of a study.

In this first part of the book we cover the basics needed to understand data and commence formal statistical analysis. In Chapter 1 we describe how to use the book to locate the statistical methods needed in different situations, and to progress from basic techniques and concepts to more sophisticated analyses.

Before commencing an analysis it is essential to gain an understanding of the data. Therefore, in Chapter 2 we focus on defining the data, explaining the concepts of populations and samples, the structure of a dataset and the different types of variables that it may contain, while in Chapter 3 we outline techniques for displaying and tabulating data.

CHAPTER 1

Using this book

1.1 Introduction
1.2 Getting started (Part A)
1.3 Finding the right statistical method (Parts B–D)
1.4 Going further (Part E)
1.5 Understanding the links between study design, analysis and interpretation (Part F)
1.6 Trying out our examples
1.7 This book and evidence-based medicine

1.1 INTRODUCTION

People usually pick up a statistics book when they have data to analyse, or when they are doing a course. This has determined the structure of this book. The ordering of topics is based on a logical progression of both methods and practical concepts, rather than a formal mathematical development. Because different statistical methods are needed for different types of data, we start by describing how to define and explore a dataset (rest of Part A). The next three parts (B, C and D) then outline the standard statistical approaches for the three main types of outcome variables (see Section 1.3). Statistical ideas are introduced as needed, methods are described in the context of relevant examples drawn from real situations, and the data we have used are available for you to reproduce the examples and try further analyses (see Section 1.6). In Part E, we introduce a collection of more advanced topics, which build on common themes in Parts B to D. These are beyond the scope of most introductory texts. The final part of the book (Part F) is devoted to issues involved in the design and conduct of a study, and how to develop an analysis strategy to get the best out of the data collected.

This book is intended to appeal to a wide audience, and to meet several needs. It is a concise and straightforward introduction to the basic methods and ideas of medical statistics, and as such is suitable for self-instruction, or as a companion to lecture courses. It does not require a mathematical background. However, it is not just an introductory text. It extends well beyond this and aims to be a comprehensive reference text for anyone seriously involved in statistical analysis. Thus it covers the major topics a medical research worker, epidemiologist or medical statistician is likely to encounter when analysing data, or when reading a scientific paper. When dealing with the more advanced methods, the focus is on the principles involved, the context in which they are required and the interpretation of computer outputs and results, rather than on the statistical theory behind them.

1.2 GETTING STARTED (PART A)

The other chapters in Part A deal with the basics of getting to know your data. In Chapter 2 ('Defining the data') we explain the link between populations and samples, and describe the different types of variables, while in Chapter 3 we outline simple techniques for tabulating and displaying them.

In particular, we introduce the distinction between **exposure variables** or **risk factors** (that is variables which influence disease outcomes, including medical treatments) and **outcome variables** (the variables whose variation or occurrence we are seeking to understand). Assessing the size and strength of the influence of one or more exposure variables on the outcome variable of interest is the core issue that runs throughout this book, and is at the heart of the majority of statistical investigations.

1.3 FINDING THE RIGHT STATISTICAL METHOD (PARTS B–D)

The appropriate statistical methods to use depend on the nature of the outcome variable of interest. Types of outcome variables are described in detail in Chapter 2; they may be essentially one of three types:

1. Numerical outcomes, such as birthweight or cholesterol level.
2. Binary outcomes, summarized as proportions, risks or odds, such as the proportion of children diagnosed with asthma, the proportion of patients in each treatment group who are no longer hypertensive, or the risk of dying in the first year of life.
3. Rates of mortality, morbidity or survival measured longitudinally over time, such as the survival rates following different treatments for breast cancer, or the number of episodes of diarrhoea per person per year among AIDS patients.

Parts B, C and D comprehensively cover the full range of standard methods for these three types of outcome respectively, and will be sufficient for the majority of analysis requirements. The emphasis throughout is on how to choose the right method for the required analysis, how to execute the method and how to interpret the results from the computer output. A quick guide to the appropriate statistical methods for the analysis of the different types of outcome variable is included on the inside covers.

The key concepts underlying statistical methods are all introduced in Part B in the context of analysing numerical outcomes, but they apply equally to all the statistical methods in the book. Statistics is used to evaluate the association between an exposure variable and the outcome of interest. More specifically, it is used to measure this association in the data collected from the particular sample of individuals in our study and to make inferences about its likely size and strength in the population from which the sample was derived. In Chapter 6, we introduce the use of a **confidence interval**, to give a range of values within which the size of the association in the population is likely to lie, taking into account **sampling variation** and **standard error**, which reflect the inherent variation between individuals.

Hypothesis tests (also known as **significance tests**) and ***P*-values**, introduced in Chapter 7, are used to assess the strength of the evidence against the **null hypothesis** that there is no true association in the population from which the sample was drawn.

The methods in these three core parts of the book range from simple techniques such as *t*-tests or chi-squared tests for comparing two exposure groups, to the use of regression models for examining the effect of several exposure variables. Throughout we aim to show how these **regression models** arise as natural extensions to the simpler methods. These more sophisticated analyses are no longer the preserve of the trained statistician. They are widely available in statistical software packages and can be used by anyone with a desktop or notebook/laptop computer, and a moderate level of computer expertise. *The more advanced sections can be omitted at a first reading, as indicated at the relevant points in the text*. It is recommended, however, that the introductions of all chapters be read, as these put the different methods into context.

1.4 GOING FURTHER (PART E)

Parts B, C and D comprehensively cover the full range of standard methods for the three types of outcome variables. This range of methods will be sufficient for the majority of analysis requirements. Part E is for those who wish to go further, and to understand general issues in statistical modelling. It can be omitted until needed.

In Part E we explain the idea of **likelihood**, upon which most statistical methods are based, discuss generic issues in regression modelling, so that skills learned in applying one type of regression model can be applied directly to the others, and describe methods that allow us to relax the assumptions made in standard statistical methods. We also include chapters for two specialised areas of analysis. The first is the analysis of **clustered data**, which arise, for example, in cluster-randomized trials where communities, rather than individuals, are randomized to receive the intervention or to act as control. The second is on **systematic reviews** and **meta-analyses**, which synthesize findings from several independent studies. Finally, we include a brief overview of the **Bayesian approach** to statistical inference.

In these more advanced chapters our emphasis is on a practical approach, focussing on what the reader needs to know to conduct such analyses, and what is needed to critically appraise their reporting in scientific papers. However, we recommend that only the introductions of the chapters be attempted at first reading. The detail can be omitted and used only when the necessity arises, and/or the reader has acquired experience of basic regression modelling.

1.5 UNDERSTANDING THE LINKS BETWEEN STUDY DESIGN, ANALYSIS AND INTERPRETATION (PART F)

The results of a study are only as good as the data on which they are based. Part F addresses the links between study design, analysis and interpretation. It starts by explaining how to choose the right analysis for each of the main types of study

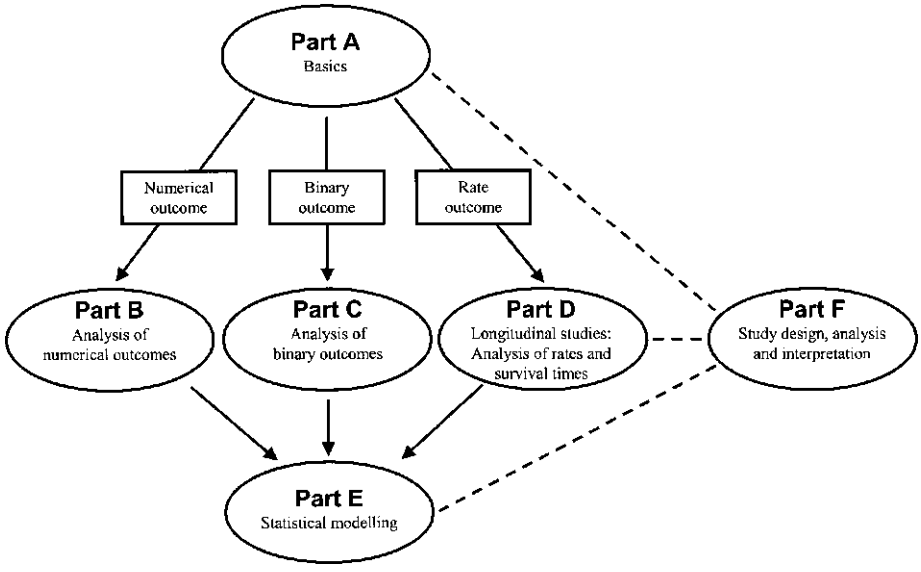


Fig. 1.1 Organization of this book.

design. It then describes how to choose an appropriate sample size, the effects of measurement error and misclassification, and the different ways in which associations can be measured and interpreted.

Finally, it is essential to plan and conduct statistical analyses in a way that maximizes the quality and interpretability of the findings. In a typical study, data are collected on a large number of variables, and it can be difficult to decide which methods to use and in what order. In Part F we aim to navigate you through this, by describing how to plan and conduct an analysis. Time invested here before you start pays off. Millions of trees must have been sacrificed to unplanned data analyses, where the data were looked at in every way imaginable. Equally often, gaps in analyses are discovered when the analyst tries to present the results. In fact it is not uncommon for people to find themselves going back to the drawing board at this stage. Careful planning of analyses should avoid these frustrations.

Of course, the issues discussed in Part F will affect all stages of the analysis of a study. This is illustrated in Figure 1.1, which shows how this book is organized.

1.6 TRYING OUT OUR EXAMPLES

Almost all statistical analyses are now done using computers, and all but very large datasets (those with measurements made on hundreds of thousands of individuals) can now be analysed using standard (desktop or laptop) office or home computers. Although simple analyses can be done with a hand-held calculator, even for these the use of a computer is recommended because results will be produced more quickly and be more accurate. For more complex analyses it is essential to use computers. Computers also allow production of high quality graphical displays.

For these reasons, we have conducted all analyses in this book using a computer. We have done these using the statistical package Stata (Stata Corporation, College Station, TX, USA; see www.stata.com). For simple analyses, we have included raw data where possible to enable readers to try out our examples for themselves. Most regression analyses presented in this book are based on datasets that are available for downloading from the book's web site, at www.blackwellpublishing.com/EssentialMedStats. Readers may wish to use these datasets either to check that they can reproduce the analyses presented in the book, or to practice further analyses.

In general, hand-held calculators do not provide facilities to perform a large enough range of statistical analyses for most purposes. In particular, they do not allow the storage of data or analysis commands that are needed to make sure that an analysis can be reproduced (see Chapter 38). However, calculators are useful for quick calculations and checking of results (both one's own and those in scientific papers). The minimum requirements are keys for scientific functions (such as square root and logarithm) and at least one memory. The new generation of handheld computers and personal organizers is blurring the distinction between calculators and computers, and it is likely that statistical software for such devices will become available in the future.

1.7 THIS BOOK AND EVIDENCE-BASED MEDICINE

As discussed above, statistics is the science of collecting, summarizing, presenting and interpreting data, and of using them to estimate the size and strengths of associations between variables. The core issue in medical statistics is how to assess the size and strength of the influence of one or more exposure variables (risk factors or treatments) on the outcome variable of interest (such as occurrence of disease or survival). In particular it aims to make inferences about this influence by studying a selected sample of individuals and using the results to make more general inferences about the wider population from which the sample was drawn.

The approach of evidence-based medicine is like a mirror to this. Inferences are made the other way around; by appraising the evidence based on the average effect of a treatment (or exposure) assessed on a large number of people, and judging its relevance to the management of a particular patient. More specifically, practitioners need to ask themselves what to consider before they can assume that the general finding will apply to a particular patient. For example, does the patient share the same characteristics as the group from which the evidence was gathered, such as age, sex, ethnic group, social class and the profile of related risk factors, such as smoking or obesity?

The evidence that the practitioner needs to appraise may come from a single study or, increasingly, from a systematic review of many. There has been an explosion in research evidence in recent decades: over two million articles are published annually in the biomedical literature and it is common for important issues to be addressed in several studies. Indeed, we might be reluctant to introduce a new treatment based on the result of one trial alone. A **systematic review**, or

overview, of the literature is a 'systematic assembly, critical appraisal and synthesis of all relevant studies on a specific topic'. The statistical methods for combining the results of a number of studies are known as **meta-analysis**. It should be emphasized that not all systematic reviews will contain a meta-analysis: this depends on the systematic review having located studies which are sufficiently similar that it is reasonable to consider combining their results. The increase in interest in meta-analysis is illustrated by the fact that while in 1987 there were 25 MEDLINE citations using the term 'meta-analysis'; this had increased to around 380 by 1991 and around 580 by 2001.

The majority of practitioners are concerned with using and appraising this evidence base, whereas the main focus of this book is on how to conduct the statistical analyses of studies that contribute to the evidence base. There are several excellent specialized evidence-based medicine books that lay out the issues in critically appraising a scientific paper or systematic review. We have therefore decided to refer the reader to these, rather than including a detailed discussion of critical appraisal in this book. We recommend Crombie (1996), Clarke and Croft (1998), Silagy and Haines (1998), Greenhalgh (2000) and Sackett *et al*. (2000).

The parts of this book that are particularly relevant to those practising evidence-based medicine are Chapters 32, 34 and 37. Thus in Chapter 32 on 'Systematic reviews and meta-analysis', we include a discussion of the sources of bias in meta-analysis and how these may be detected. In Chapter 34 we briefly review the most important aspects of the quality of randomized controlled trials. In Chapter 37 we describe the various different 'Measures of association and impact' and how to interpret them. These include numbers needed to treat or harm as well as risk ratios, odds ratios, attributable risks and absolute risk reductions. In addition, this book will be a useful companion for any practitioner who, as well as appraising the quality and relevance of the evidence base, wishes to understand more about the statistics behind the evidence generated.

CHAPTER 2

Defining the data

2.1 **Populations and samples**
2.2 **Types of variable**
Numerical variables
Binary and other categorical values
Rates
2.3 **Derived variables**
Calculated or categorized from recorded variables
Variables based on threshold values
Variables derived from reference curves, based on standard population values
Transformed variables
2.4 **Distinguishing between outcome and exposure variables**

2.1 POPULATIONS AND SAMPLES

Except when a full census is taken, we collect data on a **sample** from a much larger group called the **population**. The sample is of interest not in its own right, but for what it tells the investigator about the population. Statistics allows us to use the sample to make inferences about the population from which it was derived, as illustrated in Figure 2.1. Because of chance, different samples from the population will give different results and this must be taken into account when using a sample to make inferences about the population. This phenomenon, called **sampling variation**, lies at the heart of statistics. It is described in detail in Chapter 4.

The word 'population' is used in statistics in a wider sense than usual. It is not limited to a population of people but can refer to any collection of objects. For

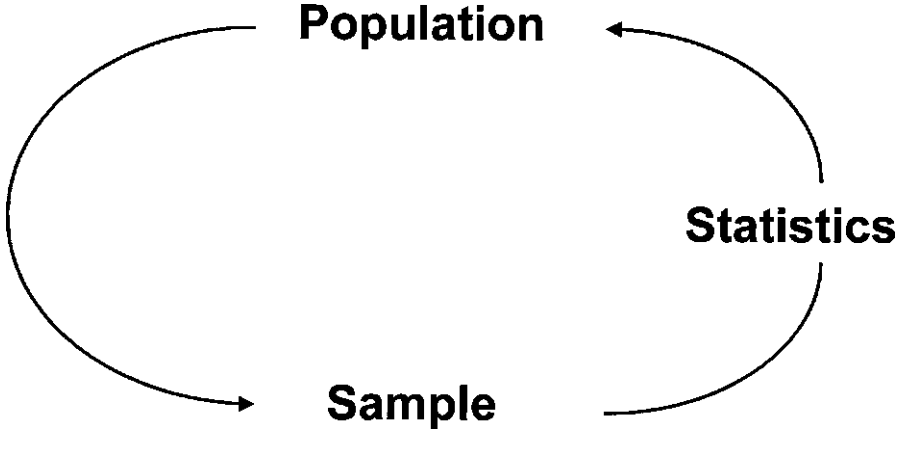


Fig. 2.1 Diagram to show the role of statistics in using information from a sample to make inferences about the population from which the sample was derived.

example, the data may relate to a sample of 20 hospitals from the population of all hospitals in the country. In such a case it is easy to imagine that the entire population can be listed and the sample selected directly from it. In many instances, however, the population and its boundaries are less precisely specified, and care must be taken to ensure that the sample truly represents the population about which information is required. This population is sometimes referred to as the **target population**. For example, consider a vaccine trial carried out using student volunteers. If it is reasonable to assume that in their response to the vaccine and exposure to disease students are typical of the community at large, the results will have general applicability. If, on the other hand, students differ in any respect which may materially affect their response to the vaccine or exposure to disease, the conclusions from the trial are restricted to the population of students and do not have general applicability. Deciding whether or not 'students are typical' is not a statistical issue, but depends on an informed judgement taking into account relevant biological and epidemiological knowledge.

Note that the target population often includes not only all persons living at present but also those that may be alive at some time in the future. This is the case in this last example evaluating the efficacy of the vaccine. It is obvious that the complete enumeration of such a population is not possible.

2.2 TYPES OF VARIABLE

The raw data of an investigation consist of **observations** made on individuals. In many situations the individuals are people, but they need not be. For instance, they might be red blood cells, urine specimens, rats, or hospitals. The number of individuals is called the **sample size**. Any aspect of an individual that is measured, like blood pressure, or recorded, like age or sex, is called a **variable**. There may be only one variable in a study or there may be many. For example, Table 2.1 shows the first six lines of data recorded in a study of outcome of treatment in tuberculosis patients treated in three hospitals. Each row of the table shows the data collected on a particular individual, while the columns of the table show the different variables which have been collected.

Table 2.1 First six lines of data from a study of outcome after diagnosis of tuberculosis.

Id	Hospital	Date of birth	Sex	Date of diagnosis	Weight (kg)	Smear result	Culture result	Skin test diameter (mm)	Alive after 6 months?
001	1	03/12/1929	M	23/08/1998	56.3	Positive	Negative	18	Y
002	1	13/04/1936	M	12/09/1998	73.5	Positive	Negative	15	Y
003	1	31/10/1931	F	17/06/1999	57.6	Positive	Positive	21	N
004	2	11/11/1922	F	05/07/1999	65.6	Uncertain	Positive	28	Y
005	2	01/05/1946	M	20/08/1999	81.1	Negative	Positive	6	Y
006	3	18/02/1954	M	17/09/1999	56.8	Positive	Negative	12	Y

A first step in choosing how best to display and analyse data is to classify the variables into their different types, as different methods pertain to each. The main division is between numerical (or quantitative) variables, categorical (or qualitative) variables and rates.

Numerical variables

A **numerical** variable is either **continuous** or **discrete**. A continuous variable, as the name implies, is a measurement on a continuous scale. In Table 2.1, weight is a continuous variable. In contrast, a discrete variable can only take a limited number of discrete values, which are usually whole numbers, such as the number of episodes of diarrhoea a child has had in a year.

Binary and other categorical variables

A **categorical** variable is non-numerical, for instance place of birth, ethnic group, or type of drug. A particularly common sort is a **binary variable** (also known as a **dichotomous** variable), which has only two possible values. For example, sex is male or female, or the patient may survive or die. We should also distinguish **ordered categorical** variables, whose categories, although non-numerical, can be considered to have a natural ordering. A common example of an ordered categorical variable is social class, which has a natural ordering from most deprived to most affluent. Table 2.2 shows the possible categories and sub-types of variable for each of the categorical variables in the data displayed in Table 2.1. Note that it could be debated whether smear result should be classified as ordered categorical or simply as categorical, depending on whether we can assume that ''uncertain'' is intermediate between 'negative' and 'positive'.

Rates

Rates of disease are measured in follow-up studies, and are the fundamental measure of the frequency of occurrence of disease over time. Their analysis forms the basis for Part D, and their exact definition can be found there. Examples include the survival rates following different treatments for breast cancer, or the number of episodes of diarrhoea/person/year among AIDS patients.

Table 2.2 Categorical (qualitative) variables recorded in the study of outcome after diagnosis of tuberculosis.

Variable	Categories	Type of variable
Hospital	1, 2, 3	Categorical
Sex	Male, female	Binary
Smear result	Negative, uncertain, positive	Ordered categorical
Culture result	Negative, positive	Binary
Alive at 6 months?	No, yes	Binary

2.3 DERIVED VARIABLES

Often, the variables included in a statistical analysis will be **derived** from those originally recorded. This may occur in a variety of different ways, and for a variety of reasons.

Calculated or categorized from recorded variables

We commonly derive a patient's *age* at diagnosis (in years) by calculating the number of days between their date of birth and date of diagnosis, and dividing this by 365.25 (the average number of days in a year, including leap years). We will often proceed to categorize age into *age groups*, for example we might define ten-year age groups as 30 to 39, 40 to 49, and so on. Age group is then an ordered categorical variable.

Another example is where the range of values observed for average monthly income is used to divide the sample into five equally-sized income groups (quintiles, see Section 3.3), and a new variable 'income group' created with '1' corresponding to the least affluent group in the population and '5' to the most affluent group.

Similarly, body mass index (BMI), which is calculated by dividing a person's weight (in kg) by the square of their height (in m), may be categorized into a 5-point scale going from $< 16\,\text{kg/m}^2$ being malnourished to $\geq 30\,\text{kg/m}^2$ defining obese. In contrast to the income group variable where the categorization is specific to the particular set of data, the categorization of the BMI scale has been carried out using conventionally agreed cut-off points to define the different groups. This type of variable, where the categorizing is based on pre-defined threshold values, is described in the next paragraph.

Variables based on threshold values

A particular group of derived variables are those based on **threshold values** of a measured variable. Two examples are given in Table 2.3. LBW is a binary variable for low birthweight ('yes' if the baby's birthweight was below 2500 g, and 'no' if

Table 2.3 Examples of derived variables based on threshold values.

Derived variable	Original variable
LBW (Low birthweight):	Birthweight:
Yes	< 2500 g
No	≥ 2500 g
Vitamin A status:	Serum retinol level:
Severe deficiency	< 0.35 μmol/l
Mild/moderate deficiency	0.35–0.69 μmol/l
Normal	≥ 0.70 μmol/l

the birthweight was 2500 g or above). Vitamin A status is an ordered categorical variable, derived from the serum retinol level.

Variables derived from reference curves, based on standard population values

A more refined comparison is based on comparing the value of a variable for the individual with **reference curves** based on the average and range of values for the whole population. For example, a child's growth can be monitored by plotting his/her weight (and height) against standard **growth curves**. This allows not only an assessment of where the child's weight (or height) lays compared to the average child at this age, but also allows growth faltering to be detected, if their growth curve appears to be dropping below what is usually expected for a child with their birthweight. How to calculate variables derived from a comparison with reference curves is postponed until Chapter 13 ('Transformations') at the end of Part B, since it requires an understanding of means, the normal distribution and *z*-scores, all of which are covered in Part B.

Transformed variables

In some cases it may be necessary to *transform* a numerical variable onto another scale in order to make it satisfy the assumptions needed for the relevant statistical methods. The **logarithmic transformation**, in which the value of the variable is replaced by its logarithm, is by far the most frequently applied. Its use is appropriate for a great variety of variables including incubation periods, parasite counts, titres, dose levels, concentrations of substances, and ratios. The reasons why a variable should be transformed, the different types of transformation, and how to choose between them are covered in detail in Chapter 13 at the end of part B.

2.4 DISTINGUISHING BETWEEN OUTCOME AND EXPOSURE VARIABLES

In order to choose appropriate data displays and statistical methods, it is very important to distinguish between *outcome* and *exposure* variables, in addition to identifying the types of each of the variables in the data set. The **outcome** variable is the variable that is the focus of our attention, whose variation or occurrence we are seeking to understand. In particular we are interested in identifying factors, or **exposures**, that may influence the size or the occurrence of the outcome variable. Some examples are given in Table 2.4. The purpose of a statistical analysis is to quantify the magnitude of the association between one or more exposure variables and the outcome variable.

A number of different terms are used to describe exposure and outcome variables, depending on the context. These are listed in Table 2.5. In particular, in a

Table 2.4 Examples of outcome and exposure variables.

Outcome variable	Exposure variable
Baby born with low birth weight (yes, no)	Mother smoked during pregnancy (yes, no)
Anthropometric status at 1 year of age (weight-for-age *z*-score)	Duration of exclusive breastfeeding (weeks)
Number of diarrhoea episodes experienced in a year	Access to clean water supply (yes, no)
Child develops leukaemia (yes, no)	Proximity to nuclear power station (miles)
Survival time (months) following diagnosis of lung cancer	Socio-economic status (6 groups)

Table 2.5 Commonly used alternatives for describing exposure and outcome variables.

Outcome variable	Exposure variable
Response variable	Explanatory variable
Dependent variable	Independent variable
y-variable	*x*-variable
Case–control group	Risk factor
	Treatment group

clinical trial (see Chapter 34) the exposure is the **treatment** group, and in a **case–control study**, the outcome is the case–control status, and the exposure variables are often called **risk factors**.

The type of outcome variable is particularly important in determining the most appropriate statistical method. Part B of this book describes statistical methods for numerical outcome variables. Part C describes methods for binary outcome variables, with a brief description (Section 20.5) of methods for categorical outcomes with more than two types of response. Part D describes methods to be used for rates, arising in studies with binary outcomes in which individuals are followed over time.

CHAPTER 3

Displaying the data

3.1 **Introduction**
3.2 **Frequencies, frequency distributions and histograms**
Frequencies (categorical variables)
Frequency distributions (numerical variables)
Histograms
Frequency polygon
Frequency distribution of the population
Shapes of frequency distributions
3.3 **Cumulative frequency distributions, quantiles and percentiles**
Cumulative frequency distributions
Median and quartiles
Quantiles and percentiles
3.4 **Displaying the association between two variables**
Cross tabulations
Scatter plots
3.5 **Displaying time trends**

3.1 INTRODUCTION

With ready access to statistical software, there is a temptation to jump straight into complex analyses. This should be avoided. An essential first step of an analysis is to summarize and display the data. The familiarity with the data gained through doing this is invaluable in developing an appropriate analysis plan (see Chapter 38). These initial displays are also valuable in identifying **outliers** (unusual values of a variable) and revealing possible errors in the data, which should be checked and, if necessary, corrected.

This chapter describes simple tabular and graphical techniques for displaying the distribution of values taken by a single variable, and for displaying the association between the values of two variables. Diagrams and tables should always be clearly labelled and self-explanatory; it should not be necessary to refer to the text to understand them. At the same time they should not be cluttered with too much detail, and they must not be misleading.

3.2 FREQUENCIES, FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

Frequencies (categorical variables)

Summarizing categorical variables is straightforward, the main task being to count the number of observations in each category. These counts are called **frequencies**. They are often also presented as **relative frequencies**; that is as proportions or percentages of the total number of individuals. For example, Table 3.1 summarizes the method of delivery recorded for 600 births in a hospital. The

Table 3.1 Method of delivery of 600 babies born in a hospital.

Method of delivery	No. of births	Percentage
Normal	478	79.7
Forceps	65	10.8
Caesarean section	57	9.5
Total	600	100.0

variable of interest is the method of delivery, a categorical variable with three categories: normal delivery, forceps delivery, and caesarean section.

Frequencies and relative frequencies are commonly illustrated by a **bar chart** (also known as a **bar diagram**) or by a **pie chart**. In a bar chart the lengths of the bars are drawn proportional to the frequencies, as shown in Figure 3.1. Alternatively the bars may be drawn proportional to the percentages in each category; the shape is not changed, only the labelling of the scale. In either case, for ease of reading it is helpful to write the actual frequency and/or percentage to the right of the bar. In a pie chart (see Figure 3.2), the circle is divided so that the areas of the sectors are proportional to the frequencies, or equivalently to the percentages.

Frequency distributions (numerical variables)

If there are more than about 20 observations, a useful first step in summarizing a numerical (quantitative) variable is to form a **frequency distribution**. This is a table showing the number of observations at different values or within certain ranges. For a discrete variable the frequencies may be tabulated either for each value of the variable or for groups of values. With continuous variables, groups have to be formed. An example is given in Table 3.2, where haemoglobin has been measured

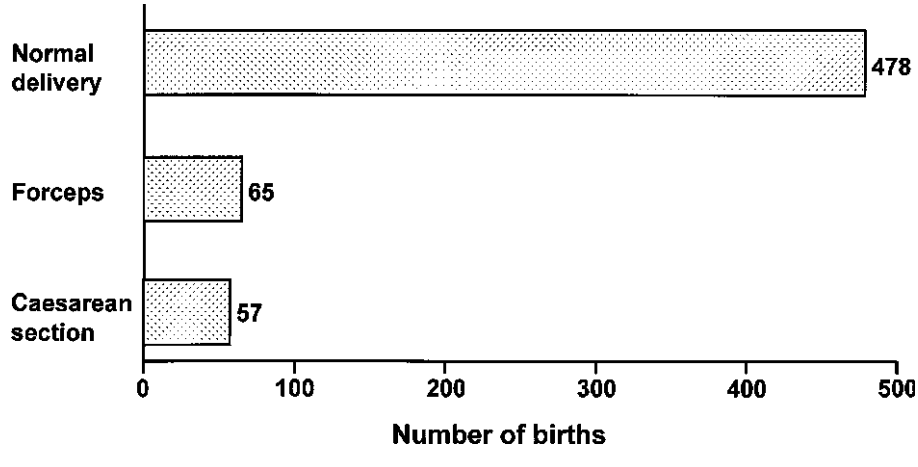


Fig. 3.1 Bar chart showing method of delivery of 600 babies born in a hospital.

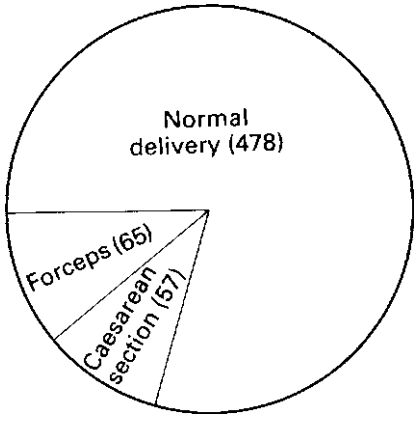


Fig. 3.2 Pie chart showing method of delivery of 600 babies born in a hospital.

to the nearest 0.1 g/100 ml and the group 11–, for example, contains all measurements between 11.0 and 11.9 g/100 ml inclusive.

When forming a frequency distribution, the first things to do are to count the number of observations and to identify the lowest and highest values. Then decide

Table 3.2 Haemoglobin levels in g/100 ml for 70 women.

(a) Raw data with the highest and lowest values underlined.

10.2	13.7	10.4	14.9	11.5	12.0	11.0
13.3	12.9	12.1	9.4	13.2	10.8	11.7
10.6	10.5	13.7	11.8	14.1	10.3	13.6
12.1	12.9	11.4	12.7	10.6	11.4	11.9
9.3	13.5	14.6	11.2	11.7	10.9	10.4
12.0	12.9	11.1	8.8	10.2	11.6	12.5
13.4	12.1	10.9	11.3	14.7	10.8	13.3
11.9	11.4	12.5	13.0	11.6	13.1	9.7
11.2	15.1	10.7	12.9	13.4	12.3	11.0
14.6	11.1	13.5	10.9	13.1	11.8	12.2

(b) Frequency distribution.

Haemoglobin (g/100 ml)	No. of women	Percentage
8–	1	1.4
9–	3	4.3
10–	14	20.0
11–	19	27.1
12–	14	20.0
13–	13	18.6
14–	5	7.1
15–15.9	1	1.4
Total	70	100.0

whether the data should be grouped and, if so, what grouping interval should be used. As a rough guide one should aim for 5–20 groups, depending on the number of observations. If the interval chosen for grouping the data is too wide, too much detail will be lost, while if it is too narrow the table will be unwieldy. The starting points of the groups should be round numbers and, whenever possible, all the intervals should be of the same width. There should be no gaps between groups. The table should be labelled so that it is clear what happens to observations that fall on the boundaries.

For example, in Table 3.2 there are 70 haemoglobin measurements. The lowest value is 8.8 and the highest 15.1 g/100 ml. Intervals of width 1 g/100 ml were chosen, leading to eight groups in the frequency distribution. Labelling the groups 8–, 9–, . . . is clear. An acceptable alternative would have been 8.0–8.9, 9.0–9.9 and so on. Note that labelling them 8–9, 9–10 and so on would have been confusing, since it would not then be clear to which group a measurement of 9.0 g/100 ml, for example, belonged.

Once the format of the table is decided, the numbers of observations in each group are counted. If this is done by hand, mistakes are most easily avoided by going through the data in order. For each value, a mark is put against the appropriate group. To facilitate the counting, these marks are arranged in groups of five by putting each fifth mark horizontally through the previous four (~~𝍸~~); these groups are called **five-bar gates**. The process is called **tallying**.

As well as the number of women, it is useful to show the percentage of women in each of the groups.

Histograms

Frequency distributions are usually illustrated by **histograms**, as shown in Figure 3.3 for the haemoglobin data. Either the frequencies or the percentages may be used; the shape of the histogram will be the same.

The construction of a histogram is straightforward when the grouping intervals of the frequency distribution are all equal, as is the case in Figure 3.3. If the intervals are of different widths, it is important to take this into account when drawing the histogram, otherwise a distorted picture will be obtained. For example, suppose the two highest haemoglobin groups had been combined in compiling Table 3.2(b). The frequency for this combined group (14.0–15.9 g/100 ml) would be six, but clearly it would be misleading to draw a rectangle of height six from 14 to 16 g/100 ml. Since this interval would be twice the width of all the others, the correct height of the line would be three, half the total frequency for this group. This is illustrated by the dotted line in Figure 3.3. The general rule for drawing a histogram when the intervals are not all the same width is to make the heights of the rectangles proportional to the frequencies divided by the widths, that is to make the areas of the histogram bars proportional to the frequencies.

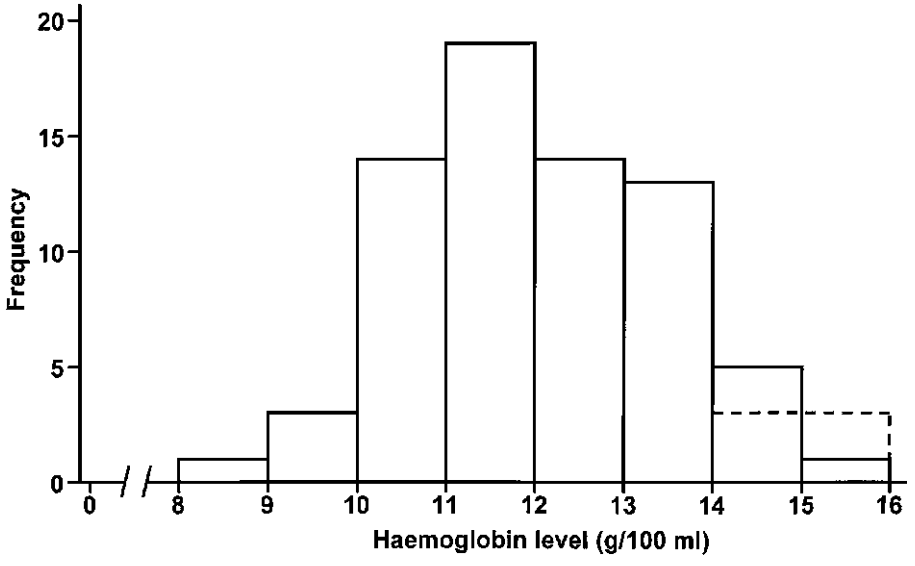


Fig. 3.3 Histogram of haemoglobin levels of 70 women.

Frequency polygon

An alternative but less common way of illustrating a frequency distribution is a **frequency polygon**, as shown in Figure 3.4. This is particularly useful when comparing two or more frequency distributions by drawing them on the same diagram. The polygon is drawn by imagining (or lightly pencilling) the histogram and joining

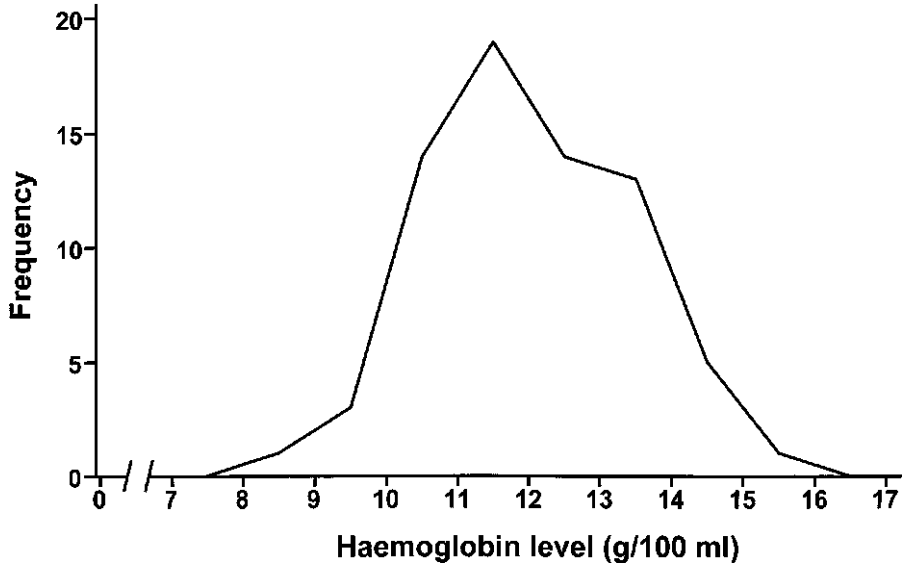


Fig. 3.4 Frequency polygon of haemoglobin levels of 70 women.

the midpoints of the tops of its rectangles. The endpoints of the resulting line are then joined to the horizontal axis at the midpoints of the groups immediately below and above the lowest and highest non-zero frequencies respectively. For the haemoglobin data, these are the groups 7.0–7.9 and 16.0–16.9 g/100 ml. The frequency polygon in Figure 3.4 is therefore joined to the axis at 7.5 and 16.5 g/100 ml.

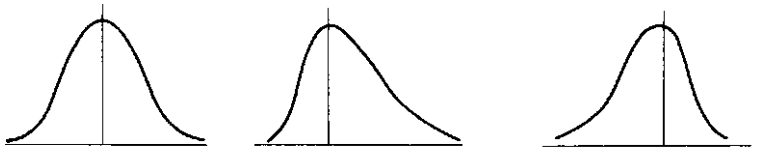
Frequency distribution of the population

Figures 3.3 and 3.4 illustrate the frequency distribution of the haemoglobin levels of a sample of 70 women. We use these data to give us information about the distribution of haemoglobin levels among women in general. For example, it seems uncommon for a woman to have a level below 9.0 g/100 ml or above 15.0 g/100 ml. Our confidence in drawing general conclusions from the data depends on how many individuals were measured. The larger the sample, the finer the grouping interval that can be chosen, so that the histogram (or frequency polygon) becomes smoother and more closely resembles the distribution of the total population. At the limit, if it were possible to ascertain the haemoglobin levels of the whole population of women, the resulting diagram would be a smooth curve.

Shapes of frequency distributions

Figure 3.5 shows three of the most common shapes of frequency distributions. They all have high frequencies in the centre of the distribution and low frequencies at the two extremes, which are called the **upper** and **lower tails** of the distribution. The distribution in Figure 3.5(a) is also **symmetrical** about the centre; this shape of curve is often described as 'bell-shaped'. The two other distributions are asymmetrical or **skewed**. The upper tail of the distribution in Figure 3.5(b) is longer than the lower tail; this is called **positively skewed** or skewed to the right. The distribution in Figure 3.5(c) is **negatively skewed** or skewed to the left.

All three distributions in Figure 3.5 are **unimodal**, that is they have just one peak. Figure 3.6(a) shows a **bimodal** frequency distribution, that is a distribution with two peaks. This is occasionally seen and usually indicates that the data are a mixture of



(a) Symmetrical and bell-shaped, e.g. height

(b) Positively skewed or skewed to the right, e.g. triceps skinfold measurement

(c) Negatively skewed or skewed to the left, e.g. period of gestation

Fig. 3.5 Three common shapes of frequency distributions with an example of each.

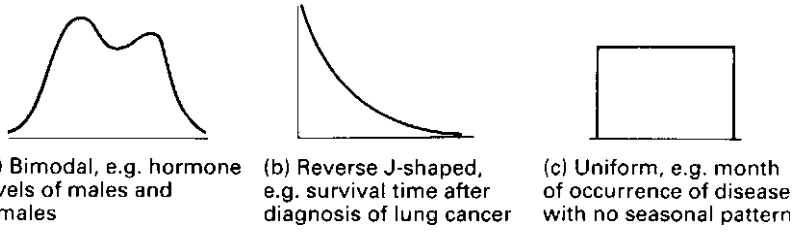


Fig. 3.6 Three less-common shapes of frequency distributions with an example of each.

two separate distributions. Also shown in Figure 3.6 are two other distributions that are sometimes found, the **reverse J-shaped** and the **uniform** distributions.

3.3 CUMULATIVE FREQUENCY DISTRIBUTIONS, QUANTILES AND PERCENTILES

Cumulative frequency distributions

Frequency distributions (and histograms) indicate the way data are distributed over a range of values, by showing the number or percentage of individuals within each group of values. **Cumulative distributions** start from the lowest value and show how the number and percentage of individuals accumulate as the values increase. For example, the cumulative frequency distribution for the first five observations of haemoglobin levels is shown in Table 3.3. There were 70 observations, so each represents $100/70 = 1.43\%$ of the total distribution. Rounding to one decimal place, the first observation (8.8 g/100 ml) corresponds to 1.4% of the distribution, the first and second observations to 2.9% of the distribution, and so on. Table 3.3 shows the values of these **cumulative percentages**, for different observations in the range of observed haemoglobin levels in the 70 women. A total of four women (5.7%) had levels below 10 g/100 ml. Similarly, 18 women (25.7%) had haemoglobin levels below 11 g/100 ml.

The **cumulative frequency distribution** is illustrated in Figure 3.7. This is drawn as a **step function**: the vertical jumps correspond to the increases in the cumulative percentages at each observed haemoglobin level. (Another example of plots that use step functions is **Kaplan–Meier** plots of cumulative survival probabilities over time; see Section 26.3.) Cumulative frequency curves are steep where there is a concentration of values, and shallow where values are sparse. In this example, where the majority of haemoglobin values are concentrated in the centre of the distribution, the curve is steep in the centre, and shallow at low and high values. If the haemoglobin levels were evenly distributed across the range, then the cumulative frequency curve would increase at a constant rate; all the steps would be the same width as well as the same height. An advantage of cumulative frequency distributions is that they display the shape of the distribution without the need for grouping, as required in plotting histograms (see Section 3.2). However the shape of a distribution is usually more clearly seen in a histogram.

Table 3.3 Cumulative percentages for different ranges of haemoglobin levels of 70 women.

Observation	Cumulative percentage	Haemoglobin level (g/100 ml)		Quartile
1	1.4	8.8	Minimum = 8.8	1
2	2.9	9.3		1
3	4.3	9.4		1
4	5.7	9.7		1
5	7.1	10.2		
⋮	⋮	⋮		
15	21.4	10.8		1
16	22.9	10.9		1
17	24.3	10.9	Lower quartile = 10.9	1
18	25.7	10.9		1
19	27.1	11.0		2
20	28.6	11.0		2
⋮	⋮	⋮		
33	47.1	11.7		2
34	48.6	11.8		2
35	50.0	11.8	Median = 11.85	2
36	51.4	11.9		3
37	52.9	11.9		3
38	54.3	12.0		3
⋮	⋮	⋮		
50	71.4	12.9		3
51	72.9	12.9		3
52	74.3	13.0		3
53	75.7	13.1	Upper quartile = 13.1	4
54	77.1	13.1		4
55	78.6	13.2		4
⋮	⋮	⋮		
66	94.3	14.6		4
67	95.7	14.6		4
68	97.1	14.7		4
69	98.6	14.9		4
70	100	15.1	Maximum = 15.1	4

Median and quartiles

Cumulative frequency distributions are useful in recoding a numerical variable into a categorical variable. The **median** is the midway value; half of the distribution lies below the median and half above it.

$$\text{Median} = \frac{(n+1)\text{th}}{2} \text{ value of the ordered observations}$$
$$(n = \text{number of observations})$$

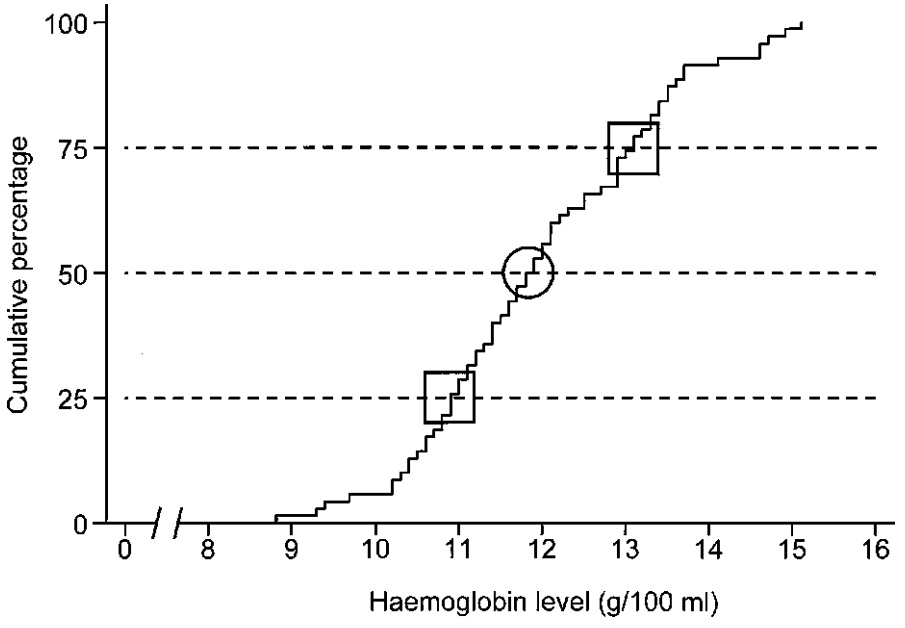


Fig. 3.7 Cumulative frequency distribution of haemoglobin levels of 70 women, with the median marked by a circle, and lower and upper quartiles marked by squares.

For the haemoglobin data, the median is the $71/2 = 35.5$th observation and so we take the average of the 35th and 36th observations. Thus the median is $(11.8+11.9)/2 = 11.85$, as shown in Table 3.3. Calculation of the median is also described in Section 4.2. When the sample size is reasonably large, the median can be estimated from the cumulative frequency distribution; it is the haemoglobin value corresponding to the point where the 50% line crosses the curve, as shown in Figure 3.7.

Also marked on Figure 3.7 are the two points where the 25% and 75% lines cross the curve. These are called the **lower** and **upper quartiles** of the distribution, respectively, and together with the median they divide the distribution into four equally-sized groups.

$$\text{Lower quartile} = \frac{(n+1)\text{th}}{4} \text{ value of the ordered observations}$$

$$\text{Upper quartile} = \frac{3 \times (n+1)\text{th}}{4} \text{ value of the ordered observations}$$

In the haemoglobin data, the lower quartile is the $71/4 = 17.75$th observation. This is calculated by taking three quarters of the difference between the 17th and 18th observations and adding it to the 17th observation. Since both the 17th and 18th observations equal 10.9 g/100 ml, so does the lower quartile, as shown

in Table 3.3. Similarly, $3 \times 71/4 = 53.25$, and since both the 53rd and 54th observations equal 13.1 g/100 ml, so does the upper quartile.

The **range** of the distribution is the difference between the minimum and maximum values. From Table 3.3, the minimum and maximum values for the haemoglobin data are 8.8 and 15.1 g/100 ml, so the range is $15.1 - 8.8 = 6.3$ g/100 ml. The difference between the lower and upper quartiles of the haemoglobin data is 2.2 g/100 ml. This is known as the **interquartile range**.

Range = highest value − lowest value

Interquartile range = upper quartile − lower quartile

A useful plot, based on these values, is a **box and whiskers plot**, as shown in Figure 3.8. The box is drawn from the lower quartile to the upper quartile; its length gives the interquartile range. The horizontal line in the middle of the box represents the median. Just as a cat's whiskers mark the full width of its body, the 'whiskers' in this plot mark the full extent of the data. They are drawn on either end of the box to the minimum and maximum values.

The right hand column of Table 3.3 shows how the median and lower and upper quartiles may be used to divide the data into equally sized groups called **quartiles**.

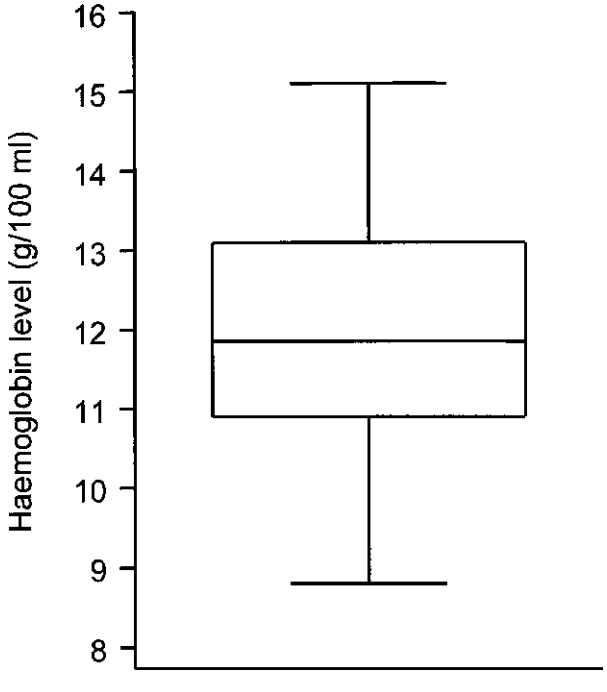


Fig. 3.8 Box and whiskers plot of the distribution of the haemoglobin levels of 70 women.

Values between 8.8 and 10.9 g/100 ml are in the first quartile, those between 11 and 11.8 g/100 ml are in the second quartile and so on. Note that equal values should always be placed in the same group, even if the groups are then of slightly different sizes.

Quantiles and percentiles

Equal-sized divisions of a distribution are called **quantiles**. For example, we may define **tertiles**, which divide the data into three equally-sized groups, and **quintiles**, which divide them into five. An example was described in Section 2.3, where the range of values observed for average monthly income was used to divide the sample into five equally-sized income groups, and a new variable 'income group' created with '1' corresponding to the least affluent group in the population and '5' to the most affluent group. Quintiles are estimated from the intersections with the cumulative frequency curve of lines at 20%, 40%, 60% and 80%. Divisions into ten equally sized groups are called **deciles**.

More generally, the kth **percentile** (or **centile** as it is also called) is the point below which k% of the values of the distribution lie. For a distribution with n observations, it is defined as:

$$k\text{th percentile} = \frac{k \times (n+1)\text{th}}{100} \text{ value of ordered observations}$$

It can also be estimated from the cumulative frequency curve; it is the x value corresponding to the point where a line drawn at k% intersects the curve. For example, the 5% point of the haemoglobin values is estimated to be 9.6 g/100 ml.

3.4 DISPLAYING THE ASSOCIATION BETWEEN TWO VARIABLES

Having examined the distribution of a single variable, we will often wish to display the way in which the distribution of one variable relates to the distribution of another. Appropriate methods to do this will depend on the type of the two variables.

Cross tabulations

When both variables are categorical, we can examine their relationship informally by cross-tabulating them in a **contingency table**. A useful convention is for the rows of the table to correspond to the exposure values and the columns to the outcomes. For example, Table 3.4 shows the results from a survey to compare the principal water sources in 150 households in three villages in West Africa. In this example, it would be natural to ask whether the household's village affects their likely water source, so that water source is the *outcome* and village is the *exposure*.

Table 3.4 Comparison of principal sources of water used by household in three villages in West Africa.

	Water source		
Village	River	Pond	Spring
A	20	18	12
B	32	20	8
C	18	12	10

The interpretability of contingency tables can be improved by including **marginal totals** and **percentages**:

- The marginal row totals show the total number of households in each village, and the marginal columns show the total numbers using each water source.
- Percentages (or proportions) can be calculated with respect to the row variable, the column variable, or the total number of individuals. A useful guide is that the percentages should correspond to the *exposure* variable. If the exposure is the row variable, as here, then row percentages should be presented, whereas if it is the column variable then column percentages should be presented.

In Table 3.4, the exposure variable, village, is the row variable, and Table 3.5 therefore shows row percentages together with marginal (row and column) totals. We can now see that, for example, the proportion of households mainly using a river was highest in Village B, while village A had the highest proportion of households mainly using a pond. By examining the column totals we can see that overall, rivers were the principal water source for 70 (47%) of the 150 households.

Table 3.5 Comparison of principal sources of water used by households in three villages in West Africa, including marginal totals and row percentages.

	Water source			
Village	River	Pond	Spring	Total
A	20 (40%)	18 (36%)	12 (24%)	50 (100%)
B	32 (53%)	20 (33%)	8 (13%)	60 (100%)
C	18 (45%)	12 (30%)	10 (25%)	40 (100%)
Total	70 (47%)	50 (33%)	30 (20%)	150 (100%)

Scatter plots

When we wish to examine the relationship between two numerical variables, we should start by drawing a scatter plot. This is a simple graph where each pair of values is represented by a symbol whose horizontal position is determined by the value of the first variable and vertical position is determined by the value of the second variable. By convention, the outcome variable determines vertical position and the exposure variable determines horizontal position.

For example, Figure 3.9 shows data from a study of lung function among 636 children aged 7 to 10 years living in a deprived suburb of Lima, Peru. The maximum volume of air which the children could breath out in 1 second (Forced Expiratory Volume in 1 second, denoted as FEV_1) was measured using a spirometer. We are interested in how FEV_1 changes with age, so that age is the exposure variable (horizontal axis) and FEV_1 is the outcome variable (vertical axis). The plot gives the clear impression that FEV_1 increases in an approximately linear manner with age.

Scatter plots may also be used to display the relationship between a categorical variable and a continuous variable. For example, in the study of lung function we are also interested in the relationship between FEV_1 and respiratory symptoms experienced by the child over the previous 12 months. Figure 3.10 shows a scatter plot that displays this relationship.

This figure is difficult to interpret, because many of the points overlap, particularly in the group of children who did not report respiratory symptoms. One solution to this is to scatter the points randomly along the horizontal axis, a process known as '**jittering**'. This produces a clearer picture, as shown in Figure 3.11. We can now see that FEV_1 tended to be higher in children who did not report respiratory symptoms in the previous 12 months than in those who did.

An alternative way to display the relationship between a numerical variable and a discrete variable is to draw **box and whiskers plots**, as described in Section 3.3. Table 3.6 shows the data needed to do this for the two groups of children: those who did and those who did not report respiratory symptoms. All the statistics displayed are

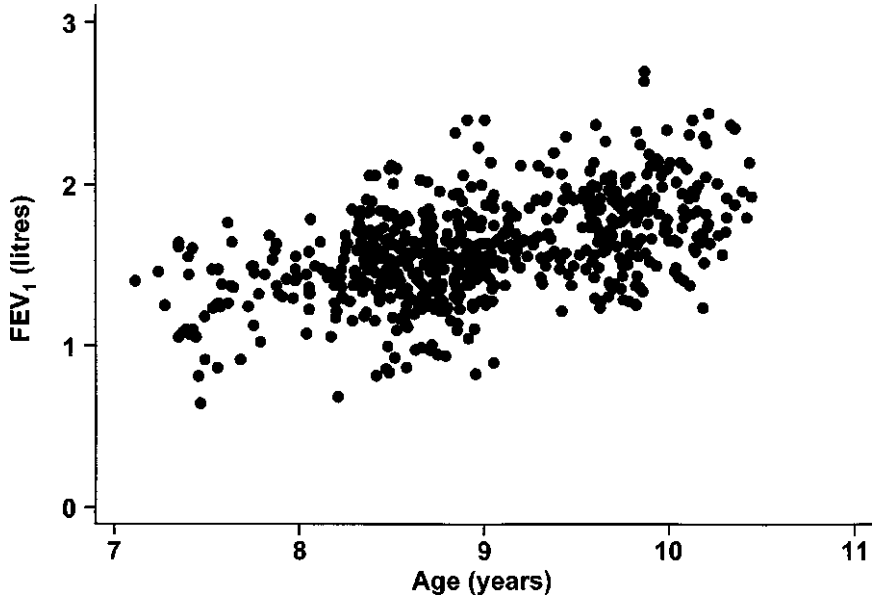


Fig. 3.9 Scatter plot showing the relationship between FEV_1 and age in 636 children living in a deprived suburb of Lima, Peru.

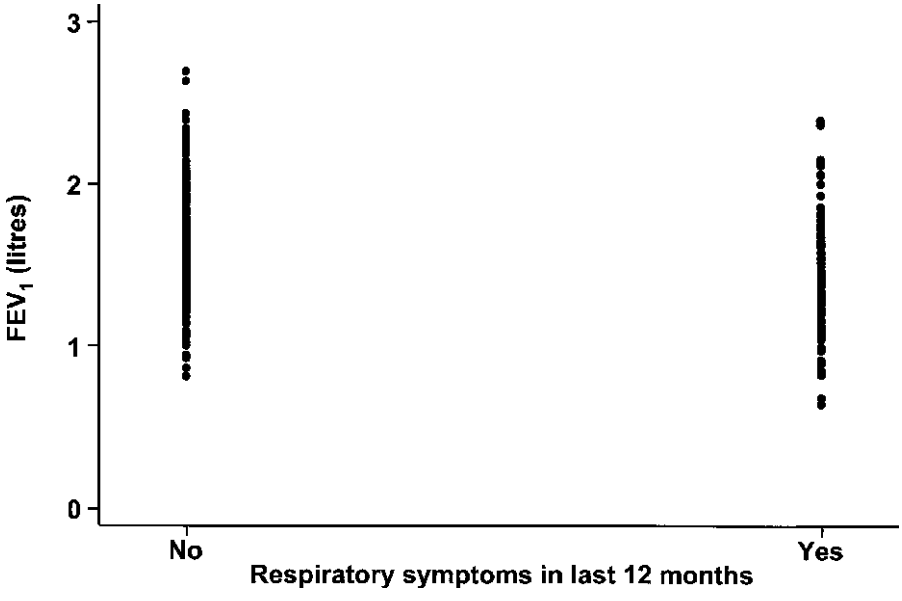


Fig. 3.10 Scatter plot showing the relationship between FEV_1 and respiratory symptoms in 636 children living in a deprived suburb of Lima, Peru.

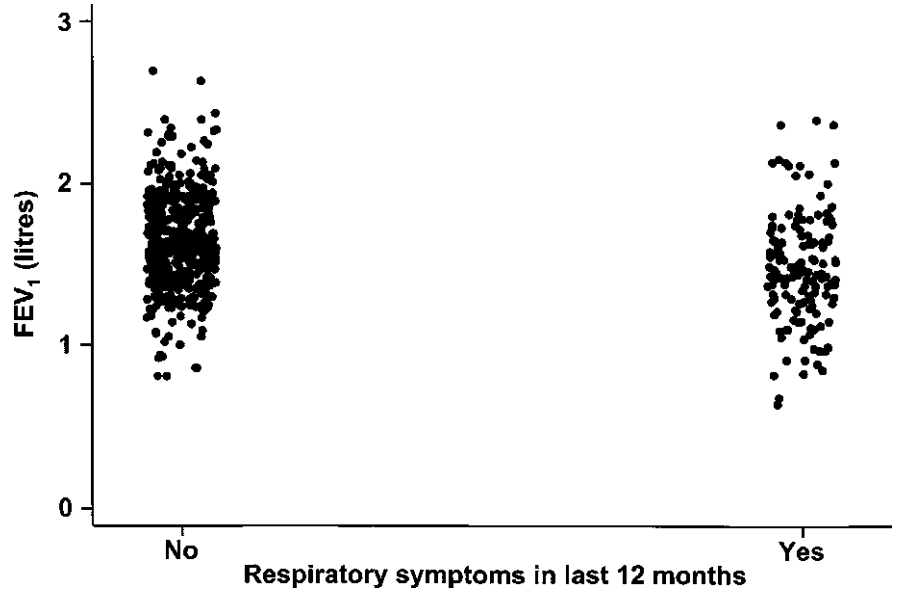


Fig. 3.11 Scatter plot showing the relationship between FEV_1 and respiratory symptoms in 636 children living in a deprived suburb of Lima, Peru. The position of the points on the horizontal axis was moved randomly ('jittered') in order to separate them.

Table 3.6 Median, interquartile range, and range of FEV_1 measurements on 636 children living in a deprived suburb of Lima, Peru, according to whether the child reported respiratory symptoms in the previous 12 months.

Respiratory symptoms in the previous 12 months	*n*	Lowest FEV_1 value	Lower quartile (25th centile)	Median	Upper quartile (75th centile)	Highest FEV_1 value
No	491	0.81	1.44	1.61	1.82	2.69
Yes	145	0.64	1.28	1.46	1.65	2.39
Totals	636	0.64	1.40	1.58	1.79	2.69

lower in children who reported symptoms. This is reflected in Figure 3.12, where all the points in the box and whiskers plot of FEV_1 values for children who reported respiratory symptoms are lower than the corresponding points in the box and whiskers plot for children who did not report symptoms.

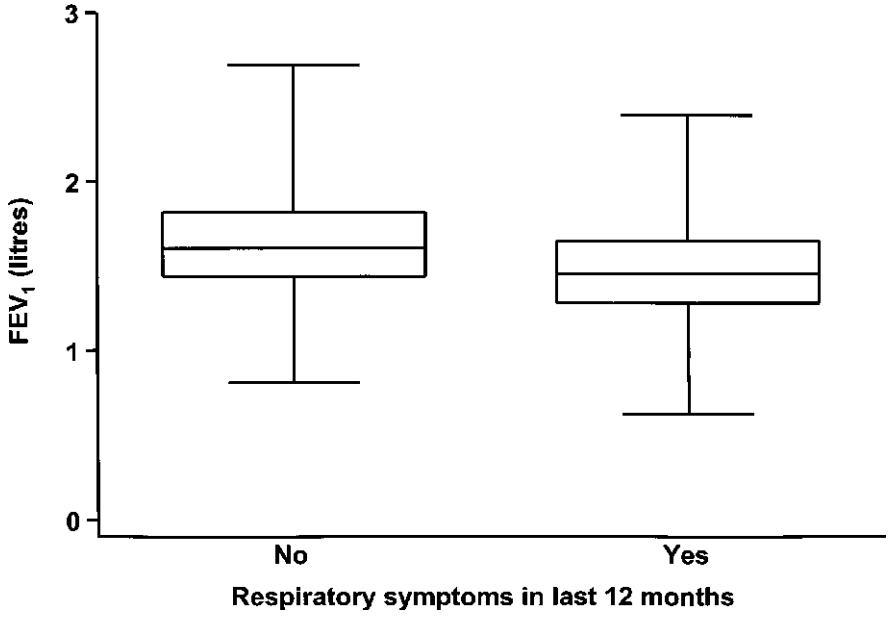


Fig. 3.12 Box and whiskers plots of the distribution of FEV_1 in 636 children living in a deprived suburb of Lima, Peru, according to whether they reported respiratory symptoms in the previous 12 months.

3.5 DISPLAYING TIME TRENDS

Graphs are also useful for displaying trends over time, such as the declines in child mortality rates that have taken place in all regions of the world in the latter half of the twentieth century, as shown in Figure 3.13. The graph also indicates the enormous differentials between regions that still remain. Note that the graph shows absolute changes in mortality rates over time. An alternative would be to

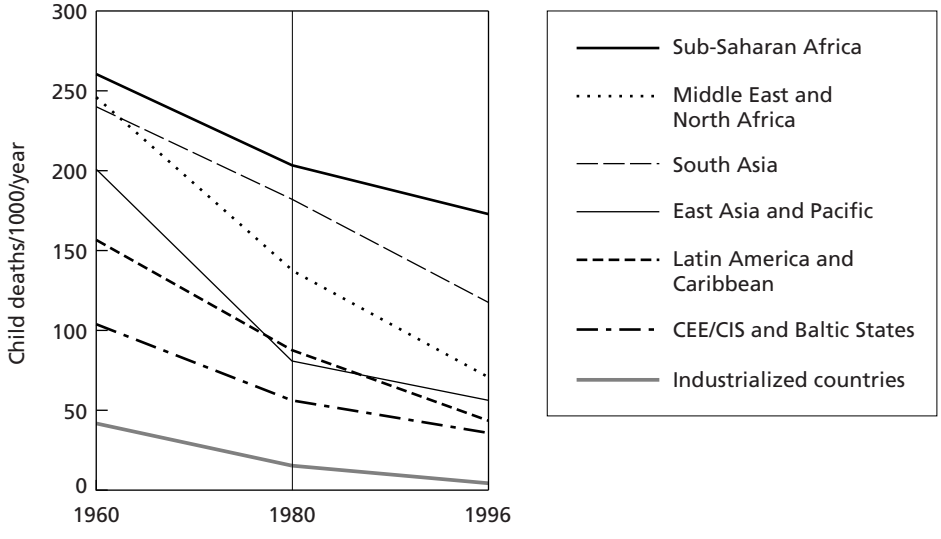


Fig. 3.13 Trends in under-five mortality rates by region of the world.

plot the logarithms of the death rates (see Chapter 13). The slopes of the lines would then show proportional declines, enabling rates of progress between regions to be readily compared.

Breaks and discontinuities in the scale(s) should be clearly marked, and avoided whenever possible. Figure 3.14(a) shows a common form of misrepresentation due to an inappropriate use of scale. The decline in infant mortality rate (IMR) has been made to look dramatic by expanding the vertical scale, while in reality the decrease over the 10 years displayed is only slight (from 22.7 to 22.1 deaths/1000 live births/year). A more realistic representation is shown in Figure 3.14(b), with the vertical scale starting at zero.

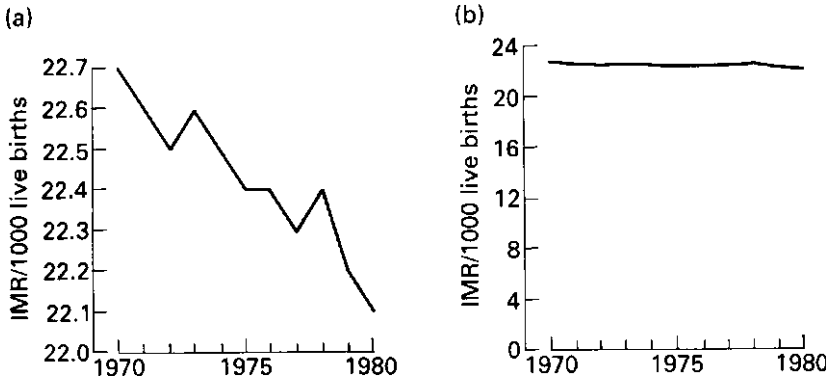


Fig. 3.14 Decline in infant mortality rate (IMR) between 1970 and 1980. (a) Inappropriate choice of scale has misleadingly exaggerated the decline. (b) Correct use of scale.