Statistical Methods in Medical Research

P. Armitage
MA, PhD
Emeritus Professor of Applied Statistics
University of Oxford

G. Berry
MA, PhD
Professor in Epidemiology and Biostatistics
University of Sydney

J.N.S. Matthews
MA, PhD
Professor of Medical Statistics
University of Newcastle upon Tyne

FOURTH EDITION

Blackwell
Science

To J.O. Irwin

Mentor and friend

Statistical Methods in Medical Research

P. Armitage
MA, PhD
Emeritus Professor of Applied Statistics
University of Oxford

G. Berry
MA, PhD
Professor in Epidemiology and Biostatistics
University of Sydney

J.N.S. Matthews
MA, PhD
Professor of Medical Statistics
University of Newcastle upon Tyne

FOURTH EDITION

Blackwell
Science

a Blackwell Publishing company
Blackwell Science, Inc., 350 Main Street, Malden, Massachusetts 02148-5018, USA
Blackwell Science Ltd, Osney Mead, Oxford OX2 0EL, UK
Blackwell Science Asia Pty Ltd, 550 Swanston Street, Carlton, Victoria 3053, Australia
Blackwell Wissenschafts Verlag, Kurfürstendamm 57, 10707 Berlin, Germany

First published 1971
Reprinted 1973, 1974, 1977, 1980, 1983, 1985
Second edition 1987
Reprinted 1988 (twice), 1990, 1991, 1993
Thrid edition 1994
Reprinted 1995, 1996
Fourth edition 2002
Reprinted 2002

Library of Congress Cataloging-in-Publication Data

Armitage, P.
Statistical methods in medical research / P. Armitage,
G. Berry, J.N.S. Matthews.—4th ed.
p. cm.
Includes bibliographical references and indexes.
ISBN 0-632-05257-0
1. Medicine—Research—Statistical methods.
I. Berry, G (Geoffrey) II. Matthews, J.N.S. III. Title.
[DNLM: 1. Biometry. 2. Research—methods.
WA 950 A 733s 2001] R852 A75 2001 610′ .7′27—dc21

00–067992

ISBN 0-632-05257-0

A catalogue record for this title is available from the British Library

Set by Kolam Information Services Pvt. Ltd., Pondicherry, India
Printed and bound in the United Kingdom by MPG Books Ltd, Bodmin, Cornwall

Commissioning Editor: Alison Brown
Production Editor: Fiona Pattison
Production Controller: Kylie Ord

For further information on Blackwell Science, visit our website:
www.blackwell-science.com

Contents

Preface to the fourth edition

In the prefaces to the first three editions of this book, we set out our aims as follows: to gather together the majority of statistical techniques that are used at all frequently in medical research, and to describe them in terms accessible to the non-mathematician. We expressed a hope that the book would have two special assets, distinguishing it from other books on applied statistics: the use of examples selected almost entirely from medical research projects, and a choice of statistical topics reflecting the extent of their usage in medical research.

These aims are equally relevant for this new edition. The steady sales of the earlier editions suggest that there was a gap in the literature which this book has to some extent filled. Why then, the reader may ask, is a new edition needed? The answer is that medical statistics (or, synonymously, *biostatistics*) is an expanding subject, with a continually developing body of techniques, and a steadily growing number of practitioners, especially in medical research organizations and the pharmaceutical industry, playing an increasingly influential role in medical research. New methods, new applications and changing attitudes call for a fresh approach to the exposition of our subject.

The first three editions followed much the same infrastructure, with little change to the original sequence of chapters—essentially an evolutionary approach to the introduction of new topics. In planning this fourth edition we decided at an early stage that the structure previously adopted had already been stretched to its limits. Many topics previously added wherever they would most conveniently fit could be handled better by a more radical rearrangement. The changing face of the subject demanded new chapters for topics now being treated at much greater length, and several areas of methodology still under active development needed to be described much more fully.

The principal changes from the third edition can be summarized as follows.

- Material on descriptive statistics is brought together in Chapter 2, following a very brief introductory Chapter 1.
- The basic results on sampling variation and inference for means, proportions and other simple measures are presented, in Chapters 4 and 5, in a more homogeneous way. For example, the important results for a mean are treated together in §4.2, rather than being split, as before, across two chapters.

- The important and influential approach to statistical inference using Bayesian methods is now dealt with much more fully—in Chapters 6 and 16, and in shorter references elsewhere in the book.
- Chapter 10 covers distribution-free methods and transformations, and also the new topics of permutation and Monte Carlo tests, the bootstrap and jackknife.
- Chapter 12 describes a wide range of special regression problems not covered in previous editions, including non-parametric and non-linear regression models, the construction of reference ranges for clinical test measurements, and multilevel models to take account of dependency between observations.
- In the treatment of categorical data primary emphasis is placed, in Chapter 14, on the use of logistic and related regression models. The older, and more empirical, methods based on χ^2 tests, are described in Chapter 15 and now related more closely to the model-based methods.
- Clinical trials, which now engage the attention of medical statisticians more intensively than ever, were allotted too small a corner in earlier editions. We now have a full treatment of the organizational and statistical aspects of trials in Chapter 18. This includes material on sequential methods, which find a natural home in §18.7.
- Chapter 19, on epidemiological statistics, includes topics previously treated separately, such as survey design and vital statistical rates.
- A new Chapter 20 on laboratory assays includes previous material on biological assay, and, in §§20.5 and 20.6, new topics such as dilution assays and tumour incidence studies.

The effect of this radical reorganization is, we hope, to improve the continuity and cohesion of the presentation, and to extend the scope to cover many new ideas now being introduced into the analysis of medical research data. We have tried to maintain the modest level of mathematical exposition which characterized earlier editions, essentially confining the mathematics to the statement of algebraic formulae rather than pursuing mathematical proofs. However, some of the newer methods involve formulae that cannot be expressed in simple algebraic terms, typically because they are most naturally explained by means of matrix algebra and/or calculus. We have attempted to ease the reader's route through these passages, but some difficulties will inevitably arise. When this happens the reader is strongly encouraged to skip the detail: continuity will not normally be lost, and the general points under discussion will usually emerge without recourse to advanced mathematics.

In the last two editions we included a final chapter on computing. Its omission from the present edition does not in any way indicate a downplaying of the role of computers in modern statistical analysis—rather the reverse. Few scientists, whether statisticians, clinicians or laboratory workers, would nowadays contemplate an analysis without recourse to a computer and a set of statistical programs, typically in the form of a standard statistics package.

However, descriptions of the characteristics of different packages quickly go out of date. Most potential users will have access to one or more packages, and probably to sources of advice about them. Detailed descriptions and instructions can, therefore, readily be obtained elsewhere. We have confined our descriptions to some general remarks in §2.2 and brief comments on specific programs at relevant points throughout the book.

As with earlier editions, we have had in mind a very broad class of readership. A major purpose of the book has always been to guide the medical research worker with no particular mathematical expertise but with the ability to follow algebraic formulae and, more particularly, the concepts behind them. Even the more advanced methods described in this edition are being extensively used in medical research and they find their way into the reports subsequently published in the medical press. It is important that the medical research worker should understand the gist of these methods, even though the technical details may remain something of a mystery.

Statisticians engaged in medical work or interested in medical applications will, we hope, find many points of interest in this new review of the subject. We hope especially that newly qualified medical statisticians, faced with the need to respond to the demands of unfamiliar applications, will find the book to be of value. Although the book developed from material used in courses for postgraduate students in the medical sciences, we have always regarded it primarily as a resource for research workers rather than as a course book. Nevertheless, much of the book would provide a useful framework for courses at various levels, either for students trained in medical or biological sciences or for those moving towards a career in medical statistics. The statistics teacher would have little difficulty in making appropriate selections for particular groups of students.

For much of the material included in the book, both illustrative and general, we owe our thanks to our present and former colleagues. We have attempted to give attributions for quoted data, but the origins of some are lost in the mists of time, and we must apologize to authors who find their data put to unsuspected purposes in these pages.

In preparing each of these editions for the press we have had much secretarial and other help from many people, to all of whom we express our thanks. We appreciate also the encouragement and support given by Stuart Taylor and his colleagues at Blackwell Science. Two of the authors (P.A. and G.B.) are grateful to the third (J.N.S.M.) for joining them in this enterprise, and all the authors thank their wives and families for their forbearance in the face of occasionally unsocial working practices.

P. Armitage
G. Berry
J.N.S. Matthews

1 The scope of statistics

In one sense medical statistics are merely numerical statements about medical matters: how many people die from a certain cause each year, how many hospital beds are available in a certain area, how much money is spent on a certain medical service. Such facts are clearly of administrative importance. To plan the maternity-bed service for a community we need to know how many women in that community give birth to a child in a given period, and how many of these should be cared for in hospitals or maternity homes. Numerical facts also supply the basis for a great deal of medical research; examples will be found throughout this book. It is no purpose of the book to list or even to summarize numerical information of this sort. Such facts may be found in official publications of national or international health departments, in the published reports of research investigations and in textbooks and monographs on medical subjects. This book is concerned with the general rather than the particular, with methodology rather than factual information, with the general principles of statistical investigations rather than the results of particular studies.

Statistics may be defined as the discipline concerned with the treatment of numerical data derived from groups of individuals. These individuals will often be people—for instance, those suffering from a certain disease or those living in a certain area. They may be animals or other organisms. They may be different administrative units, as when we measure the case-fatality rate in each of a number of hospitals. They may be merely different occasions on which a particular measurement has been made.

Why should we be interested in the numerical properties of groups of people or objects? Sometimes, for administrative reasons like those mentioned earlier, statistical facts are needed: these may be contained in official publications; they may be derivable from established systems of data collection such as cancer registries or systems for the notification of congenital malformations; they may, however, require specially designed statistical investigations.

This book is concerned particularly with the uses of statistics in medical research, and here—in contrast to its administrative uses—the case for statistics has not always been free from controversy. The argument occasionally used to be heard that statistical information contributes little or nothing to the progress of medicine, because the physician is concerned at any one time with the treatment of a single patient, and every patient differs in important respects from every

other patient. The clinical judgement exercised by a physician in the choice of treatment for an individual patient is based to an extent on theoretical considerations derived from an understanding of the nature of the illness. But it is based also on an appreciation of statistical information about diagnosis, treatment and prognosis acquired either through personal experience or through medical education. The important argument is whether such information should be stored in a rather informal way in the physician's mind, or whether it should be collected and reported in a systematic way. Very few doctors acquire, by personal experience, factual information over the whole range of medicine, and it is partly by the collection, analysis and reporting of statistical information that a common body of knowledge is built and solidified.

The phrase *evidence-based medicine* is often applied to describe the compilation of reliable and comprehensive information about medical care (Sackett *et al*., 1996). Its scope extends throughout the specialties of medicine, including, for instance, research into diagnostic tests, prognostic factors, therapeutic and prophylactic procedures, and covers public health and medical economics as well as clinical and epidemiological topics. A major role in the collection, critical evaluation and dissemination of such information is played by the Cochrane Collaboration, an international network of research centres (http://www.cochrane.org/).

In all this work, the statistical approach is essential. The variability of disease is an argument *for* statistical information, not *against* it. If the bedside physician finds that on one occasion a patient with migraine feels better after drinking plum juice, it does not follow, from this single observation, that plum juice is a useful therapy for migraine. The doctor needs statistical information showing, for example, whether in a group of patients improvement is reported more frequently after the administration of plum juice than after the use of some alternative treatment.

The difficulty of arguing from a single instance is equally apparent in studies of the aetiology of disease. The fact that a particular person was alive and well at the age of 95 and that he smoked 50 cigarettes a day and drank heavily would not convince one that such habits are conducive to good health and longevity. Individuals vary greatly in their susceptibility to disease. Many abstemious non-smokers die young. To study these questions one should look at the morbidity and mortality experience of groups of people with different habits: that is, one should do a statistical study.

The second chapter of this book is concerned mainly with some of the basic tools for collecting and presenting numerical data, a part of the subject usually called *descriptive statistics*. The statistician needs to go beyond this descriptive task, in two important respects. First, it may be possible to improve the quality of the information by careful planning of the data collection. For example, information on the efficacy of specific treatments is most reliably obtained from the experimental approach provided by a *clinical trial* (Chapter 18),

and questions about the aetiology of disease can be tackled by carefully designed *epidemiological surveys* (Chapter 19). Secondly, the methods of *statistical inference* provide a largely objective means of drawing conclusions from the data about the issues under research. Both these developments, of planning and inference, owe much to the work of R.A. (later Sir Ronald) Fisher (1890–1962), whose influence is apparent throughout modern statistical practice.

Almost all the techniques described in this book can be used in a wide variety of branches of medical research, and indeed frequently in the non-medical sciences also. To set the scene it may be useful to mention four quite different investigations in which statistical methods played an essential part.

1 MacKie *et al.* (1992) studied the trend in the incidence of primary cutaneous malignant melanoma in Scotland during the period 1979–89. In assessing trends of this sort it is important to take account of such factors as changes in standards of diagnosis and in definition of disease categories, changes in the pattern of referrals of patients in and out of the area under study, and changes in the age structure of the population. The study group was set up with these points in mind, and dealt with almost 4000 patients. The investigators found that the annual incidence rate increased during the period from 3·4 to 7·1 per 100 000 for men, and from 6·6 to 10·4 for women. These findings suggest that the disease, which is known to be affected by high levels of ultraviolet radiation, may be becoming more common even in areas where these levels are relatively low.

2 Women who have had a pregnancy with a neural tube defect (NTD) are known to be at higher than average risk of having a similar occurrence in a future pregnancy. During the early 1980s two studies were published suggesting that vitamin supplementation around the time of conception might reduce this risk. In one study, women who agreed to participate were given a mixture of vitamins including folic acid, and they showed a much lower incidence of NTD in their subsequent pregnancies than women who were already pregnant or who declined to participate. It was possible, however, that some systematic difference in the characteristics of those who participated and those who did not might explain the results. The second study attempted to overcome this ambiguity by allocating women randomly to receive folic acid supplementation or a placebo, but it was too small to give clear-cut results. The Medical Research Council (MRC) Vitamin Study Research Group (1991) reported a much larger randomized trial, in which the separate effects could be studied of both folic acid and other vitamins. The outcome was clear. Of 593 women receiving folic acid and becoming pregnant, six had NTD; of 602 not receiving folic acid, 21 had NTD. No effect of other vitamins was apparent. Statistical methods confirmed the immediate impression that the contrast between the folic acid and control

groups is very unlikely to be due to chance and can safely be ascribed to the treatment used.

3 The World Health Organization carried out a collaborative case–control study at 12 participating centres in 10 countries to investigate the possible association between breast cancer and the use of oral contraceptives (WHO Collaborative Study of Neoplasia and Steroid Contraceptives, 1990). In each hospital, women with breast cancer and meeting specific age and residential criteria were taken as cases. Controls were taken from women who were admitted to the same hospital, who satisfied the same age and residential criteria as the cases, and who were not suffering from a condition considered as possibly influencing contraceptive practices. The study included 2116 cases and 13 072 controls. The analysis of the association between breast cancer and use of oral contraceptives had to consider a number of other variables that are associated with breast cancer and which might differ between users and non-users of oral contraceptives. These variables included age, age at first live birth (2·7-fold effect between age 30 or older and less than 20 years), a socio-economic index (twofold effect), year of marriage and family history of breast cancer (threefold effect). After making allowance for these possible confounding variables as necessary, the risk of breast cancer for users of oral contraceptives was estimated as 1·15 times the risk for non-users, a weak association in comparison with the size of the associations with some of the other variables that had to be considered.

4 A further example of the use of statistical arguments is a study to quantify illness in babies under 6 months of age reported by Cole *et al.* (1991). It is important that parents and general practitioners have an appropriate method for identifying severe illness requiring referral to a specialist paediatrician. Whether this is possible can only be determined by the study of a large number of babies for whom possible signs and symptoms are recorded, and for whom the severity of illness is also determined. In this study the authors considered 28 symptoms and 47 physical signs. The analysis showed that it was sufficient to use seven of the symptoms and 12 of the signs, and each symptom or sign was assigned an integer score proportional to its importance. A baby's illness score was then derived by adding the scores for any signs or symptoms that were present. The score was then considered in three categories, 0–7, 8–12 and 13 or more, indicating well or mildly ill, moderate illness and serious illness, respectively. It was predicted that the use of this score would correctly classify 98% of the babies who were well or mildly ill and correctly identify 92% of the seriously ill.

These examples come from different fields of medicine. A review of research in any one branch of medicine is likely to reveal the pervasive influence of the statistical approach, in laboratory, clinical and epidemiological studies. Consider, for instance, research into the human immunodeficiency virus (HIV) and

the acquired immune deficiency syndrome (AIDS). Early studies extrapolated the trend in reported cases of AIDS to give estimates of the future incidence. However, changes in the incidence of clinical AIDS are largely determined by the trends in the incidence of earlier events, namely the original HIV infections. The timing of an HIV infection is usually unknown, but it is possible to use estimates of the incubation period to work backwards from the AIDS incidence to that of HIV infection, and then to project forwards to obtain estimates of future trends in AIDS. Estimation of duration of survival of AIDS patients is complicated by the fact that, at any one time, many are still alive, a standard situation in the analysis of survival data (Chapter 17). As possible methods of treatment became available, they were subjected to carefully controlled clinical trials, and reliable evidence was produced for the efficacy of various forms of combined therapy. The progression of disease in each patient may be assessed both by clinical symptoms and signs and by measurement of specific markers. Of these, the most important are the CD4 cell count, as a measure of the patient's immune status, and the viral load, as measured by an assay of viral RNA by the polymerase chain reaction (PCR) method or some alternative test. Statistical questions arising with markers include their ability to predict clinical progression (and hence perhaps act as surrogate measures in trials that would otherwise require long observation periods); their variability, both between patients and on repeated occasions on the same patient; and the stability of the assay methods used for the determinations.

Statistical work in this field, as in any other specialized branch of medicine, must take into account the special features of the disease under study, and must involve close collaboration between statisticians and medical experts. Nevertheless, most of the issues that arise are common to work in other branches of medicine, and can thus be discussed in fairly general terms. It is the purpose of this book to present these general methods, illustrating them by examples from different medical fields.

Statistical investigations

The statistical investigations described above have one feature in common: they involve observations of a similar type being made on each of a group of individuals. The individuals may be people (as in **1**–**4** above), animals, blood samples, or even inanimate objects such as birth certificates or parishes. The need to study groups rather than merely single individuals arises from the presence of random, unexplained variation. If all patients suffering from the common cold experienced well-defined symptoms for precisely 7 days, it might be possible to demonstrate the merits of a purported drug for the alleviation of symptoms by administering it to one patient only. If the symptoms lasted only 5 days, the reduction could safely be attributed to the new treatment. Similarly, if blood

pressure were an exact function of age, varying neither from person to person nor between occasions on the same person, the blood pressure at age 55 could be determined by one observation only. Such studies would not be statistical in nature and would not call for statistical analysis. Those situations, of course, do not hold. The duration of symptoms from the common cold varies from one attack to another; blood pressures vary both between individuals and between occasions. Comparisons of the effects of different medical treatments must therefore be made on groups of patients; studies of physiological norms require population surveys.

In the planning of a statistical study a number of administrative and technical problems are likely to arise. These will be characteristic of the particular field of research and cannot be discussed fully in the present context. Two aspects of the planning will almost invariably be present and are of particular concern to the statistician. The investigator will wish the inferences from the study to be sufficiently precise, and will also wish the results to be relevant to the questions being asked. Discussions of the statistical design of investigations are concerned especially with the general considerations that bear on these two objectives. Some of the questions that arise are: (i) how to select the individuals on which observations are to be made; (ii) how to decide on the numbers of observations falling into different groups; and (iii) how to allocate observations between different possible categories, such as groups of animals receiving different treatments or groups of people living in different areas.

It is useful to make a conceptual distinction between two different types of statistical investigation, the *experiment* and the *survey*. Experimentation involves a planned interference with the natural course of events so that its effect can be observed. In a survey, on the other hand, the investigator is a more passive observer, interfering as little as possible with the phenomena to be recorded. It is easy to think of extreme examples to illustrate this antithesis, but in practice the distinction is sometimes hard to draw. Consider, for instance, the following series of statistical studies:

1. A register of deaths occurring during a particular year, classified by the cause of death.
2. A survey of the types of motor vehicle passing a checkpoint during a certain period.
3. A public opinion poll.
4. A study of the respiratory function (as measured by various tests) of men working in a certain industry.
5. Observations of the survival times of mice of three different strains, after inoculation with the same dose of a toxic substance.
6. A clinical trial to compare the merits of surgery and conservative treatment for patients with a certain condition, the subjects being allotted randomly to the two treatments.

Studies **1** to **4** are clearly surveys, although they involve an increasing amount of interference with nature. Study **6** is equally clearly an experiment. Study **5** occupies an equivocal position. In its statistical aspects it is conceptually a survey, since the object is to observe and compare certain characteristics of three strains of mice. It happens, though, that the characteristic of interest requires the most extreme form of interference—the death of the animal—and the non-statistical techniques are more akin to those of a laboratory experiment than to those required in most survey work.

The general principles of experimental design will be discussed in §9.1, and those of survey design in §§19.2 and 19.4.

2 Describing data

2.1 **Diagrams**

One of the principal methods of displaying statistical information is the use of diagrams. Trends and contrasts are often more readily apprehended, and perhaps retained longer in the memory, by casual observation of a well-proportioned diagram than by scrutiny of the corresponding numerical data presented in tabular form. Diagrams must, however, be simple. If too much information is presented in one diagram it becomes too difficult to unravel and the reader is unlikely even to make the effort. Furthermore, details will usually be lost when data are shown in diagrammatic form. For any critical analysis of the data, therefore, reference must be made to the relevant numerical quantities.

Statistical diagrams serve two main purposes. The first is the presentation of statistical information in articles and other reports, when it may be felt that the reader will appreciate a simple, evocative display. Official statistics of trade, finance, and medical and demographic data are often illustrated by diagrams in newspaper articles and in annual reports of government departments. The powerful impact of diagrams makes them also a potential means of misrepresentation by the unscrupulous. The reader should pay little attention to a diagram unless the definition of the quantities represented and the scales on which they are shown are all clearly explained. In research papers it is inadvisable to present basic data solely in diagrams because of the loss of detail referred to above. The use of diagrams here should be restricted to the emphasis of important points, the detailed evidence being presented separately in tabular form.

The second main use is as a private aid to statistical analysis. The statistician will often have recourse to diagrams to gain insight into the structure of the data and to check assumptions which might be made in an analysis. This informal use of diagrams will often reveal new aspects of the data or suggest hypotheses which may be further investigated.

Various types of diagrams are discussed at appropriate points in this book. It will suffice here to mention a few of the main uses to which statistical diagrams are put, illustrating these from official publications.

1 *To compare two or more numbers*. The comparison is often by bars of different lengths (Fig. 2.1), but another common method (the *pictogram*) is

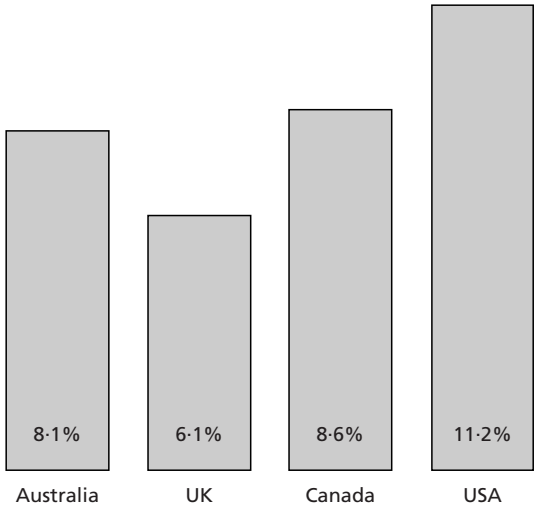


Fig. 2.1 A bar diagram showing the percentages of gross domestic product spent on health care in four countries in 1987 (reproduced with permission from Macklin, 1990).

to use rows of repeated symbols; for example, the populations of different countries may be depicted by rows of 'people', each 'person' representing 1 000 000 people. Care should be taken not to use symbols of the same shape but different sizes because of ambiguity in interpretation; for example, if exports of different countries are represented by money bags of different sizes the reader is uncertain whether the numerical quantities are represented by the linear or the areal dimensions of the bags.

2 *To express the distribution of individual objects or measurements into different categories*. The frequency distribution of different values of a numerical measurement is usually depicted by a histogram, a method discussed more fully in §2.3 (see Figs 2.6–2.8). The distribution of individuals into non-numerical categories can be shown as a *bar diagram* as in **1**, the length of each bar representing the number of observations (or *frequency*) in each category. If the frequencies are expressed as percentages, totalling 100%, a convenient device is the *pie chart* (Fig. 2.2).

3 *To express the change in some quantity over a period of time*. The natural method here is a graph in which points, representing the values of the quantity at successive times, are joined by a series of straight-line segments (Fig. 2.3). If the time intervals are very short the graph will become a smooth curve. If the variation in the measurement is over a small range centred some distance from zero it will be undesirable to start the scale (usually shown vertically) at zero for this will leave too much of the diagram completely blank. A non-zero origin should be indicated by a break in the axis at the

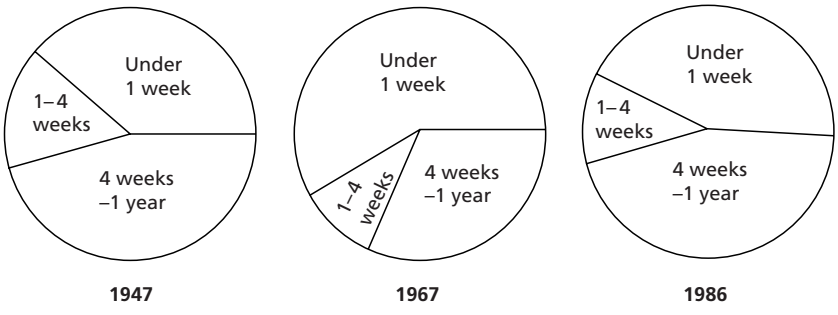


Fig. 2.2 A pie chart showing for three different years the proportions of infant deaths in England and Wales that occur in different parts of the first year of life. The amount for each category is proportional to the angle subtended at the centre of the circle and hence to the area of the sector.

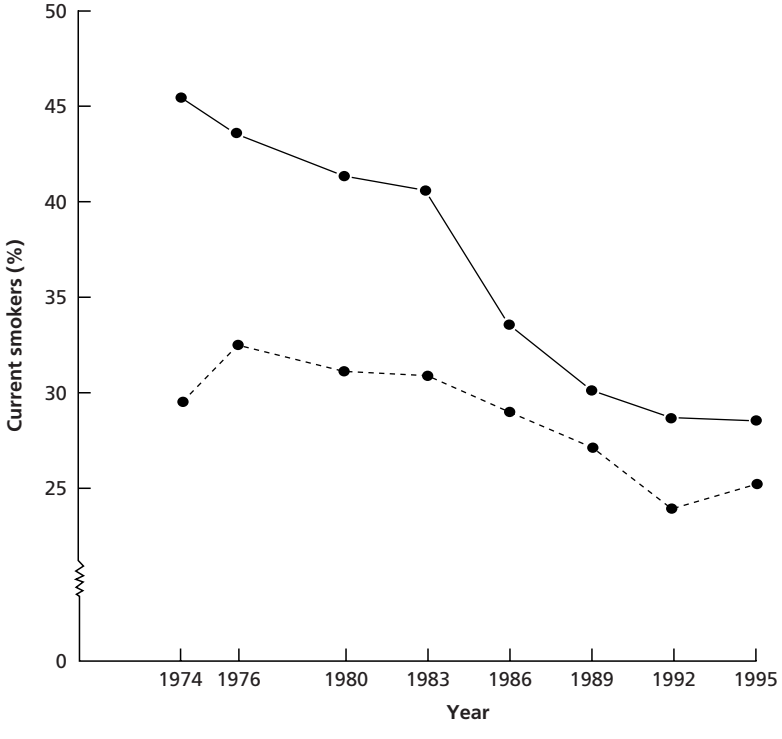


Fig. 2.3 A line diagram showing the changes between six surveys in the proportion of men (solid line) and women (dashed line) in Australia who were current smokers (adapted from Hill *et al.*, 1998).

lower end of the scale, to attract the readers' attention (Fig. 2.3). A slight trend can, of course, be made to appear much more dramatic than it really is by the judicious choice of a non-zero origin, and it is unfortunately only too easy for the unscrupulous to support a chosen interpretation of a time trend

by a careful choice of origin. A sudden change of scale over part of the range of variation is even more misleading and should almost always be avoided. Special scales based on logarithmic and other transformations are discussed in §§2.5 and 10.8.

4 *To express the relationship between two measurements, in a situation where they occur in pairs*. The usual device is the *scatter diagram* (see Fig. 7.1), which is described in detail in Chapter 7 and will not be discussed further here. Time trends, discussed in **3**, are of course a particular form of relationship, but they called for special comment because the data often consist of one measurement at each point of time (these times being often equally spaced). In general, data on relationships are not restricted in this way and the continuous graph is not generally appropriate.

Modern computing methods provide great flexibility in the construction of diagrams, by such features as interaction with the visual display, colour printing and dynamic displays of complex data. For extensive reviews of the art of graphical display, see Tufte (1983), Cleveland (1985, 1993) and Martin and Welsh (1998).

2.2 **Tabulation and data processing**

Tabulation

Another way of summarizing and presenting some of the important features of a set of data is in the form of a table. There are many variants, but the essential features are that the structure and meaning of a table are indicated by headings or labels and the statistical summary is provided by numbers in the body of the table. Frequently the table is two-dimensional, in that the headings for the horizontal rows and vertical columns define two different ways of categorizing the data. Each portion of the table defined by a combination of row and column is called a *cell*. The numerical information may be counts of numbers of individuals in different cells, mean values of some measurements (see §2.4) or more complex indices.

Some useful guidelines in the presentation of tables for publication are given by Ehrenberg (1975, 1977). Points to note are the avoidance of an unnecessarily large number of digits (since shorter, rounded-off numbers convey their message to the eye more effectively) and care that the layout allows the eye easily to compare numbers that need to be compared.

Table 2.1, taken from a report on assisted conception (AIH National Perinatal Statistics Unit, 1991), is an example of a table summarizing counts. It summarizes information on 5116 women who conceived following *in vitro* fertilization (IVF), and shows that the proportion of women whose pregnancy

Table 2.1 Outcome of pregnancies according to maternal age (adapted from AIH National Perinatal Statistics Unit, 1991).

Age		Live birth	Spontaneous abortion	Ectopic pregnancy	Stillbirth	Termination of pregnancy	Total
< 25	No.	94	21	10	2	0	127
	%	74·0	16·5	7·9	1·6	0·0	100·0
25–29	No.	962	272	96	36	2	1368
	%	70·3	19·9	7·0	2·6	0·1	99·9
30–34	No.	1615	430	143	58	8	2254
	%	71·7	19·1	6·3	2·6	0·4	100·1
35–39	No.	789	338	66	27	6	1226
	%	64·4	27·6	5·4	2·2	0·5	100·1
40 +	No.	69	60	6	1	5	141
	%	48·9	42·6	4·3	0·7	3·5	100·0
Total	No.	3529	1121	321	124	21	5116
	%	69·0	21·9	6·3	2·4	0·4	100·0

resulted in a live birth was related to age. How is such a table constructed? With a small quantity of data a table of this type could be formed by manual sorting and counting of the original records, but if there were many observations (as in Table 2.1) or if many tables had to be produced the labour would obviously be immense.

Data collection and preparation

We may distinguish first between the problems of preparing the data in a form suitable for tabulation, and the mechanical (or electronic) problems of getting the computations done. Some studies, particularly small laboratory experiments, give rise to relatively few observations, and the problems of data preparation are correspondingly simple. Indeed, tabulations of the type under discussion may not be required, and the statistician may be concerned solely with more complex forms of analysis.

Data preparation is, in contrast, a problem of serious proportions in many large-scale investigations, whether with complex automated laboratory measurements or in clinical or other studies on a 'human' scale. In large-scale therapeutic and prophylactic trials, in prognostic investigations, in studies in epidemiology and social medicine and in many other fields, a large number of people may be included as subjects, and very many observations may be made on each subject. Furthermore, much of the information may be difficult to obtain in unambigu-

ous form and the precise definition of the variables may require careful thought. This subsection and the two following ones are concerned primarily with data from these large studies.

In most investigations of this type it will be necessary to collect the information on specially designed record forms or questionnaires. The design of forms and questionnaires is considered in some detail by Babbie (1989). The following points may be noted briefly here.

1 There is a temptation to attempt to collect more information than is clearly required, in case it turns out to be useful in either the present or some future study. While there is obviously a case for this course of action it carries serious disadvantages. The collection of data costs money and, although the cost of collecting extra information from an individual who is in any case providing some information may be relatively low, it must always be considered. The most serious disadvantage, though, is that the collection of marginally useful information may detract from the value of the essential data. The interviewer faced with 50 items for each subject may take appreciably less care than if only 20 items were required. If there is a serious risk of non-cooperation of the subject, as perhaps in postal surveys using questionnaires which are self-administered, the length of a questionnaire may be a strong disincentive and the list of items must be severely pruned. Similarly, if the data are collected by telephone interview, cooperation may be reduced if the respondent expects the call to take more than a few minutes.

2 Care should be taken over the wording of questions to ensure that their interpretation is unambiguous and in keeping with the purpose of the investigation. Whenever possible the various categories of response that are of interest should be enumerated on the form. This helps to prevent meaningless or ambiguous replies and saves time in the later classification of results. For example,

What is your working status? (circle number)

1 Domestic duties with no paid job outside home.
2 In part-time employment (less than 25 hours per week).
3 In full-time employment.
4 Unemployed seeking work.
5 Retired due to disability or illness (please specify cause).....
6 Retired for other reasons.
7 Other (please specify).....

If the answer to a question is a numerical quantity the units required should be specified. For example,

Your weight: . . . . . .kg.

In some cases more than one set of units may be in common use and both should be allowed for. For example,

Your height: . . . . . cm.
Or . . . . . feet . . . . . inches.

In other cases it may be sufficient to specify a number of categories. For example,

How many years have you lived in this town? (circle number)

1 Less than 5.
2 5–9.
3 10–19.
4 20–29.
5 30–39.
6 40 or more.

When the answer is qualitative but may nevertheless be regarded as a gradation of a single dimensional scale, a number of ordered choices may be given. For example,

How much stress or worry have you had in the last month with:

		None	A little	Some	Much	Very much
1	Your spouse?	1	2	3	4	5
2	Other members of your family?	1	2	3	4	5
3	Friends?	1	2	3	4	5
4	Money or finance?	1	2	3	4	5
5	Your job?	1	2	3	4	5
6	Your health?	1	2	3	4	5

Sometimes the data may be recorded directly into a computer. Biomedical data are often recorded on automatic analysers or other specialized equipment, and automatically transferred to a computer. In telephone interviews, it may be possible to dispense with the paper record, so that the interviewer reads a question on the computer screen and enters the response directly from the keyboard.

In many situations, though, the data will need to be transferred from data sheets to a computer, a process described in the next subsection.

Data transfer

The data are normally entered via the keyboard and screen on to disk, either the computer's own hard disk or a floppy disk (diskette) or both. Editing facilities allow amendments to be made directly on the stored data. As it is no longer necessary to keep a hard copy of the data in computer-readable form, it is essential to maintain back-up copies of data files to guard against computer malfunctions that may result in a particular file becoming unreadable.

There are two strategies for the entry of data. In the first the data are regarded as a row of characters, and no interpretation occurs until a data file

has been created. The second method is much more powerful and involves using the computer interactively as the data are entered. Questionnaires often contain items that are only applicable if a particular answer has been given to an earlier item. For example, if a detailed smoking history is required, the first question might be 'Have you smoked?' If the answer was 'yes', there would follow several questions on the number of years smoked, the amount smoked, the brands of cigarettes, etc. On the other hand, if the answer was 'no', these questions would not be applicable and should be skipped. With screen-based data entry the controlling program would automatically display the next applicable item on the screen.

There are various ways in which information from a form or questionnaire can be represented in a computer record. In the simplest method the reply to each question is given in one or more specific columns and each column contains a digit from 0 to 9. This means that non-numerical information must be 'coded'. For example, the coding of the first few questions might be as in Fig. 2.4. In some systems leading zeros must be entered, e.g. if three digits were allowed for a variable like diastolic blood pressure, a reading of 88 mmHg would be recorded as 088, whereas other systems allow blanks instead. For a subject with study number 122 who was a married woman aged 49, the first eight columns of the record given in Fig. 2.4 would be entered as the following codes:

Column	1	2	3	4	5	6	7	8
Code	0	1	2	2	2	4	9	2

Clearly the person entering the data must know which code to enter for any particular column. Two different approaches are possible. The information may

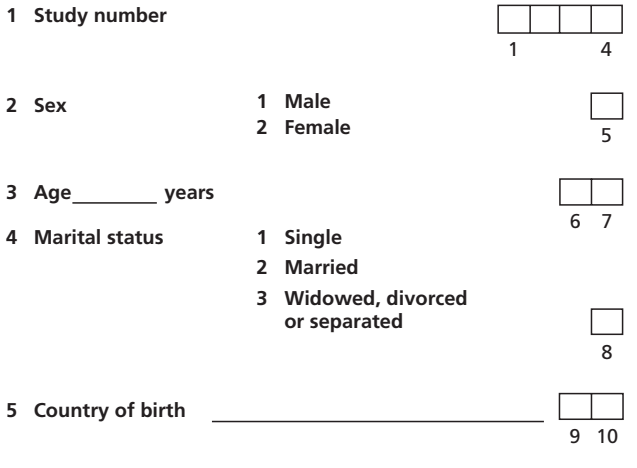


Fig. 2.4 An example of part of a questionnaire with coding indicated.

be transferred from the original record to a 'coding sheet' which will show for each column of each record precisely which code is to be entered. This may be a sheet of paper, ruled as a grid, in which the rows represent the different individuals and the vertical columns represent the columns of the record. Except for small jobs it will usually be preferable to design a special coding form showing clearly the different items; this will reduce the frequency of transcription errors. Alternatively, the coding may be included on the basic record form so that the transfer may be done direct from this form and the need for an intermediate coding sheet is removed. If sufficient care is given to the design of the record form, this second method is preferable, as it removes a potential source of copying errors. This is the approach shown in Fig. 2.4, where the boxes on the right are used for coding. For the first four items the codes are shown and an interviewer could fill in the coding boxes immediately. For item 5 there are so many possibilities that all the codes cannot be shown. Instead, the response would be recorded in open form, e.g. 'Greece', and the code looked up later in a detailed set of coding instructions.

It was stated above that it is preferable to use the record form or questionnaire also for the coding. One reservation must, however, be made. The purpose of the questionnaire is to obtain accurate information, and anything that detracts from this should be removed. Particularly with self-administered questionnaires the presence of coding boxes, even though the respondent is not asked to use them, may reduce the cooperation a subject would otherwise give. This may be because of an abhorrence of what may be regarded as looking like an 'official' form, or it may be simply that the boxes have made the form appear cramped and less interesting. This should not be a problem where a few interviewers are being used but if there is any doubt, separate coding sheets should be used.

With screen-based entry the use of coding boxes is not necessary but care is still essential in the questionnaire design to ensure that the information required by the operator is easy to find.

The statistician or investigator wishing to tabulate the data in various ways using a computer must have access to a suitable program, and statistical packages are widely available for standard tasks such as tabulation.

It is essential that the data and the instructions for the particular analysis required be prepared in, or converted to, the form specified by the package. It may be better to edit the data in the way that leads to the fewest mistakes, and then to use a special editing program to get the data into the form needed for the package.

When any item of information is missing, it is inadvisable to leave a blank in the data file as that would be likely to cause confusion in later analyses. It is better to have a code such as '9' or '99' for 'missing'. However, when the missing information is numerical, care must be taken to ensure that the code cannot be

mistaken for a real observation. The coding scheme shown in Fig. 2.4 would be deficient in a survey of elderly people, since a code of '99' for an unknown age could be confused with a true age of 99 years, and indeed there is no provision for centenarians. A better plan would have been to use three digits for age, and to denote a missing reading by, say, '999'.

Data cleaning

Before data are subjected to further analyses, they should be carefully checked for errors. These may have arisen during data entry, and ideally data should be transferred by *double entry*, independently by two different operators, the two files being checked for consistency by a separate computer program. In practice, most data-processing organizations find this system too expensive, and rely on single entry by an experienced operator, with regular monitoring for errors, which should be maintained at a very low rate.

Other errors may occur because inaccurate information appeared on the initial record forms. Computer programs can be used to detect implausible values, which can be checked and corrected where necessary. Methods of data checking are discussed further in §2.7.

With direct entry of data, as in a telephone interview, logical errors or implausible values could be detected by the computer program and queried immediately with the respondent.

Statistical computation

Most of the methods of analysis described later in this book may be carried out using standard statistical packages or languages. Widely available packages include *BMDP* (BMDP, 1993), *SPSS* (SPSS, 1999), *Stata* (Stata, 2001) *MINITAB* (Minitab, 2000), *SAS* (SAS, 2000) and *SYSTAT* (SYSTAT, 2000). The scope of such packages develops too rapidly to justify any detailed descriptions here, but summaries can be found on the relevant websites, with fuller descriptions and operating instructions in the package manuals. Goldstein (1998) provides a useful summary. Many of these packages, such as *SAS*, offer facilities for the data management tasks described earlier in this section. *S-PLUS* (S-PLUS, 2000) provides an interactive data analysis system, together with a programming language, *S*. For very large data sets a database management system such as *Oracle* may be needed (Walker, 1998). *StatsDirect* (StatsDirect, 1999) is a more recent package covering many of the methods for medical applications that are described in this book.

Some statistical analyses may be performed on small data sets, or on compact tables summarizing larger data sets, and these may be read, item by item, directly into the computer. In larger studies, the analyses will refer to data

extracted from the full data file. In such cases it will be useful to form a derived file containing the subset of data needed for the analysis, in whatever form is required by the package program. As Altman (1991) remarks, the user is well advised as far as possible to use the same package for all his or her analyses, 'as it takes a considerable effort to become fully acquainted with even one package'.

In addition to the major statistical computing packages, which cover many of the standard methods described in this book, there are many other packages or programs suitable for some of the more specialized tasks. Occasional references to these are made throughout the book.

Although computers are increasingly used for analysis, with smaller sets of data it is often convenient to use a calculator, the most convenient form of which is the pocket calculator. These machines perform at high speed all the basic arithmetic operations, and have a range of mathematical functions such as the square, square root, exponential, logarithm, etc. An additional feature particularly useful in statistical work is the automatic calculation and accumulation of sums of squares of numbers. Some machines have a special range of extended facilities for statistical analyses. It is particularly common for the automatic calculation of the mean and standard deviation to be available. Programmable calculators are available and these facilitate repeated use of statistical formulae.

The user of a calculator often finds it difficult to know how much rounding off is permissible in the data and in the intermediate or final steps of the computations. Some guidance will be derived from the examples in this book, but the following general points may be noted.

1. Different values of any one measurement should normally be expressed to the same degree of precision. If a series of children's heights is generally given to the nearest centimetre, but a few are expressed to the nearest millimetre, this extra precision will be wasted in any calculations done on the series as a whole. All the measurements should therefore be rounded to the nearest centimetre for convenience of calculation.
2. A useful rule in rounding mid-point values (such as a height of 127·5 cm when rounding to whole numbers) is to round to the nearest even number. Thus 127·5 would be rounded to 128. This rule prevents a slight bias which would otherwise occur if the figures were always rounded up or always rounded down.
3. It may occasionally be justifiable to quote the results of calculations to a little more accuracy than the original data. For example, if a large series of heights is measured to the nearest centimetre the mean may sometimes be quoted to one decimal point. The reason for this is that, as we shall see, the effect of the rounding errors is reduced by the process of averaging.
4. If any quantity calculated during an intermediate stage of the calculations is quoted to, say, n significant digits, the result of any multiplication or division

of this quantity will be valid to, at the most, n digits. The significant digits are those from the first non-zero digit to the last meaningful digit, irrespective of the position of the decimal point. Thus, 1·002, 10·02, 100 200 (if this number is expressed to the nearest 100) all have four significant digits. Cumulative inaccuracy arises with successive operations of multiplication or division.

5 The result of an addition or subtraction is valid to, at most, the number of decimal digits of the least accurate figure. Thus, the result of adding 101 (accurate to the nearest integer) and 4·39 (accurate to two decimal points) is 105 (to the nearest integer). The last digit may be in error by one unit; for example, the exact figure corresponding to 101 may have been 101·42, in which case the result of the addition now should have been 105·81, or 106 to the nearest integer. These considerations are particularly important in subtraction. Very frequently in statistical calculations one number is subtracted from another of very similar size. The result of the subtraction may then be accurate to many fewer significant digits than either of the original numbers. For example, $3212{\cdot}78 - 3208{\cdot}44 = 4{\cdot}34$; three digits have been lost by the subtraction. For this reason it is essential in some early parts of a computation to keep more significant digits than will be required in the final result.

A final general point about computation is that the writing down of intermediate steps offers countless opportunities for error. It is therefore important to keep a tidy layout on paper, with adequate labelling and vertical and horizontal alignment of digits, and without undue crowding.

2.3 **Summarizing numerical data**

The raw material of all statistical investigations consists of individual observations, and these almost always have to be summarized in some way before any use can be made of them. We have discussed in the last two sections the use of diagrams and tables to present some of the main features of a set of data. We must now examine some particular forms of table, and the associated diagrams, in more detail. As we have seen, the aim of statistical methods goes beyond the mere presentation of data to include the drawing of inferences from them. These two aspects—description and inference—cannot be entirely separated. We cannot discuss the descriptive tools without some consideration of the purpose for which they are needed. In the next few sections, we shall occasionally have to anticipate questions of inference which will be discussed in more detail later in the book.

Any class of measurement or classification on which individual observations are made is called a *variable* or *variate*. For instance, in one problem the variable might be a particular measure of respiratory function in schoolboys, in another it might be the number of bacteria found in samples of water. In most problems

many variables are involved. In a study of the natural history of a certain disease, for example, observations are likely to be made, for each patient, on a number of variables measuring the clinical state of the patient at various times throughout the illness, and also on certain variables, such as age, not directly relating to the patient's health.

It is useful first to distinguish between two types of variable, *qualitative* (or *categorical*) and *quantitative*. Qualitative observations are those that are not characterized by a numerical quantity, but whose possible values consist of a number of categories, with any individual recorded as belonging to just one of these categories. Typical examples are sex, hair colour, death or survival in a certain period of time, and occupation. Qualitative variables may be subdivided into *nominal* and *ordinal* observations. An ordinal variable is one where the categories have an unambiguous natural order. For example, the stage of a cancer at a certain site may be categorized as state A, B, C or D, where previous observations have indicated that there is a progression through these stages in sequence from A to D. Sometimes the fact that the stages are ordered may be indicated by referring to them in terms of a number, stage 1, 2, 3 or 4, but the use of a number here is as a label and does not indicate that the variable is quantitative. A nominal variable is one for which there is no natural order of the categories. For example, certified cause of death might be classified as infectious disease, cancer, heart disease, etc. Again, the fact that cause of death is often referred to as a number (the International Classification of Diseases, or ICD, code) does not obscure the fact that the variable is nominal, with the codes serving only as shorthand labels.

The problem of summarizing qualitative nominal data is relatively simple. The main task is to count the number of observations in various categories, and perhaps to express them as proportions or percentages of appropriate totals. These counts are often called *frequencies* or *relative frequencies*. Examples are shown in Tables 2.1 and 2.2. If relative frequencies in certain subgroups are shown, it is useful to add them to give 1·00, or 100%, so that the reader can easily see which total frequencies have been subdivided. (Slight discrepancies in these totals, due to rounding the relative frequencies, as in Tables 2.1 and 2.3, may be ignored.)

Ordinal variables may be summarized in the same way as nominal variables. One difference is that the order of the categories in any table or figure is predetermined, whereas it is arbitrary for a nominal variable. The order also allows the calculation of *cumulative relative frequencies*, which are the sums of all relative frequencies below and including each category.

A particularly important type of qualitative observation is that in which a certain characteristic is either present or absent, so that the observations fall into one of two categories. Examples are sex, and survival or death. Such variables are variously called *binary, dichotomous* or *quantal*.

Table 2.2 Result of sputum examination 3 months after operation in group of patients treated with streptomycin and control group treated without streptomycin.

	Streptomycin		Control	
	Frequency	%	Frequency	%
Smear negative, culture negative	141	45·0	117	41·8
Smear negative, not cultured	90	28·8	67	23·9
Smear or culture positive	82	26·2	96	34·3
Total with known sputum result	313	100·0	280	100·0
Results not known	12		17	
Total	325		297	

Table 2.3 Frequency distribution of number of lesions caused by smallpox virus in egg membranes.

Number of lesions	Frequency (number of membranes)	Relative frequency (%)
0–	1	1
10–	6	8
20–	14	18
30–	14	18
40–	17	21
50–	8	10
60–	9	11
70–	3	4
80–	6	8
90–	1	1
100–	0	0
110–119	1	1
Total	80	101

Quantitative variables are those for which the individual observations are numerical quantities, usually either measurements or counts. It is useful to subdivide quantitative observations into *discrete* and *continuous* variables. Discrete measurements are those for which the possible values are quite distinct and separated. Often they are counts, such as the number of times an individual has been admitted to hospital in the last 5 years.

Continuous variables are those which can assume a continuous uninterrupted range of values. Examples are height, weight, age and blood pressure. Continuous measurements usually have an upper and a lower limit. For instance, height

cannot be less than zero, and there is presumably some lower limit above zero and some upper limit, but it would be difficult to say exactly what these limits are. The distinction between discrete and continuous variables is not always clear, because all continuous measurements are in practice rounded off; for instance, a series of heights might be recorded to the nearest centimetre and so appear discrete. Any ambiguity rarely matters, since the same statistical methods can often be safely applied to both continuous and discrete variables, particularly if the scale used for the latter is fairly finely subdivided. On the other hand, there are some special methods applicable to counts, which as we have seen must be positive whole numbers. The problems of summarizing quantitative data are much more complex than those for qualitative data, and the remainder of this chapter will be devoted almost entirely to them.

Sometimes a continuous or a discrete quantitative variable may be summarized by dividing the range of values into a number of categories, or *grouping intervals*, and producing a table of frequencies. For example, for age a number of age groups could be created and each individual put into one of the groups. The variable, age, has then been transformed into a new variable, age group, which has all the characteristics of an ordered categorical variable. Such a variable may be called an *interval* variable.

A useful first step in summarizing a fairly large collection of quantitative data is the formation of a *frequency distribution*. This is a table showing the number of observations, or frequency, at different values or within certain ranges of values of the variable. For a discrete variable with a few categories the frequency may be tabulated at each value, but, if there is a wide range of possible values, it will be convenient to subdivide the range into categories. An example is shown in Table 2.3. (In this example the reader should note the distinction between two types of count—the variable, which is the number of lesions on an individual chorioallantoic membrane, and the frequency, which is the number of membranes on which the variable falls within a specified range.) With continuous measurements one *must* form grouping intervals (Table 2.4). In Table 2.4 the cumulative relative

Table 2.4 Frequency distribution of age for 1357 male patients with lung cancer.

Age (years)	Frequency (number of patients)	Relative frequency (%)	Cumulative relative frequency (%)
25–	17	1·3	1·3
35–	116	8·5	9·8
45–	493	36·3	46·1
55–	545	40·2	86·3
65–74	186	13·7	100·0
Total	1357	100·0	

frequencies are also tabulated. These give the percentages of the total who are younger than the lower limit of the following interval, that is, 9·8% of the subjects are in the age groups 25–34 and 35–44 and so are younger than 45.

The advantages in presenting numerical data in the form of a frequency distribution rather than a long list of individual observations are too obvious to need stressing. On the other hand, if there are only a few observations, a frequency distribution will be of little value since the number of readings falling into each group will be too small to permit any meaningful pattern to emerge.

We now consider in more detail the practical task of forming a frequency distribution. If the variable is to be grouped, a decision will have to be taken about the end-points of the groups. For convenience these should be chosen, as far as possible, to be 'round' numbers. For distributions of age, for example, it is customary to use multiples of 5 or 10 as the boundaries of the groups. Care should be taken in deciding in which group to place an observation falling on one of the group boundaries, and the decision must be made clear to the reader. Usually such an observation is placed in the group of which the observation is the lower limit. For example, in Table 2.3 a count of 20 lesions would be placed in the group 20–, which includes all counts between 20 and 29, and this convention is indicated by the notation used for the groups.

How many groups should there be? No clear-cut rule can be given. To provide a useful, concise indication of the nature of the distribution, fewer than five groups will usually be too few and more than 20 will usually be too many. Again, if too large a number of groups is chosen, the investigator may find that many of the groups contain frequencies which are too small to provide any regularity in the shape of the distribution. For a given size of grouping interval this difficulty will become more acute as the total number of observations is reduced, and the choice of grouping interval may, therefore, depend on this number. If in doubt, the grouping interval may be chosen smaller than that to be finally used, and groups may be amalgamated in the most appropriate way after the distribution has been formed.

If the original data are contained in a computer file, a frequency distribution can readily be formed by use of a statistical package. If the measurements are available only as a list on paper, the counts should be made by going systematically through the list, 'tallying' each measurement into its appropriate group. The whole process should be repeated as a check. The alternative method of taking each group in turn and counting the observations falling into that group is not to be recommended, as it requires the scanning of the list of observations once for each group (or more than once if a check is required) and thus encourages mistakes.

If the number of observations is not too great (say, fewer than about 50), a frequency distribution can be depicted graphically by a diagram such as Fig. 2.5. Here each individual observation is represented by a dot or some other mark

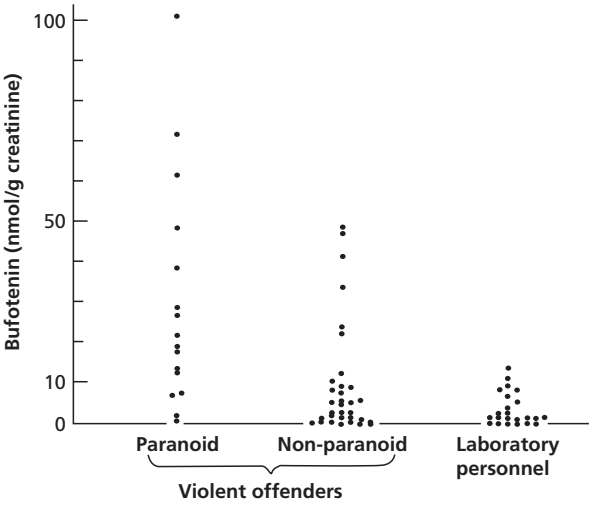


Fig. 2.5 Dot diagram showing the distribution of urinary excretion of bufotenin in three groups of subjects (reprinted from Räisänen *et al*., 1984, by permission of the authors and the editor of *The Lancet*).

opposite the appropriate point on a scale. The general shape of the distribution can be seen at a glance, and it is easy to compare visually two or more distributions of the same variable (Fig. 2.5). With larger numbers of observations this method is unsuitable because the marks tend to become congested, and a *box-and-whisker* plot is more suitable (see p. 38).

When the number of observations is large the original data may be grouped into a frequency distribution table and the appropriate form of diagram is then the *histogram*. Here the values of the variable are by convention represented on the horizontal scale, and the vertical scale represents the frequency, or relative frequency, at each value or in each group. If the variable is discrete and ungrouped (Fig. 2.6), the frequencies may be represented by vertical lines. The more general method, which must be applied if the variable is grouped, is to draw rectangles based on the different groups (Figs 2.7 and 2.8). It may happen that the grouping intervals are not of constant length. In Table 2.3, for example, suppose we decided to pool the groups 60–, 70– and 80–. The total frequency in these groups is 18, but it would clearly be misleading to represent this frequency by a rectangle on a base extending from 60 to 90 and with a height of 18. The correct procedure would be to make the height of the rectangle 6, the average frequency in the three groups (as indicated by the dashed line in Fig. 2.7). One way of interpreting this rule is to say that the height of the rectangle in a histogram is the frequency per standard grouping of the variable (in this example the standard grouping is 10 lesions). Another way is to say that the frequency for

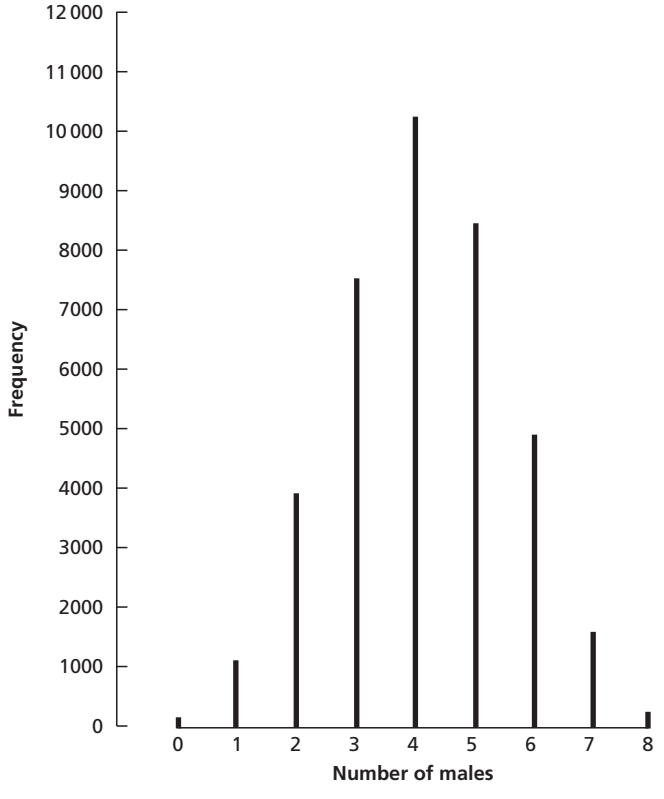


Fig. 2.6 Histogram representing the frequency distribution for an ungrouped discrete variable (number of males in sibships of eight children).

a group is proportional to the *area* rather than the height of the rectangle (in this example the area of any of the original rectangles, or of the composite rectangle formed by the dashed line, is 10 times the frequency for the group). If there is no variation in length of grouping interval, areas are, of course, proportional to heights, and frequencies are represented by either heights or areas.

The cumulative relative frequency may be represented by a line diagram (Fig. 2.9). The positioning of the points on the age axis needs special care, since in the frequency distribution (Table 2.4) the cumulative relative frequencies in the final column are plotted against the start of the age group in the next line. That is, since none of the men are younger than 25, zero is plotted on the vertical axis at age 25, 1·3% are younger than 35 so 1·3% is plotted at age 35, 9·8% at age 45, and so on to 100% at age 75.

The *stem-and-leaf display*, illustrated in Table 2.5, is a useful way of tabulating the original data and, at the same time, depicting the general shape of the frequency distribution. In Table 2.5, the first column lists the initial digit in

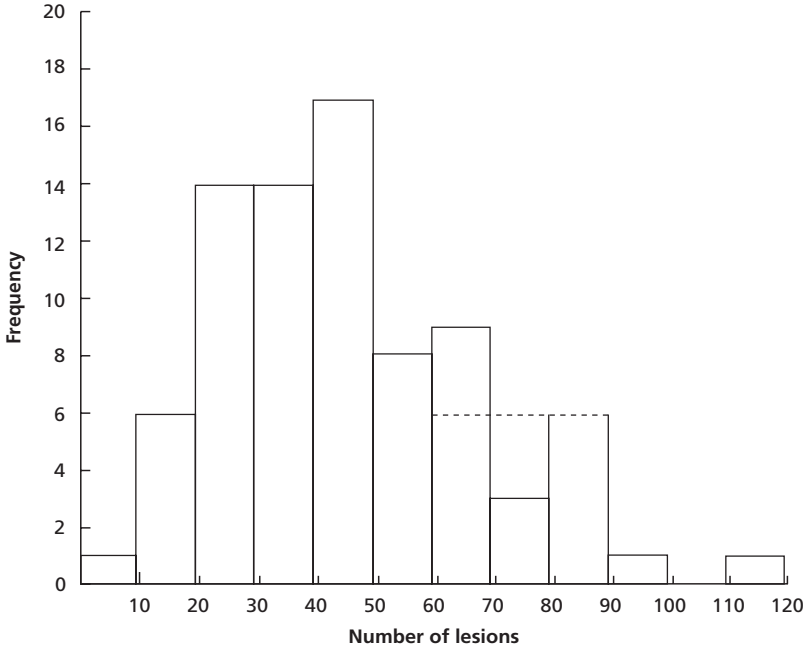


Fig. 2.7 Histogram representing the frequency distribution for a grouped discrete variable (Table 2.3).

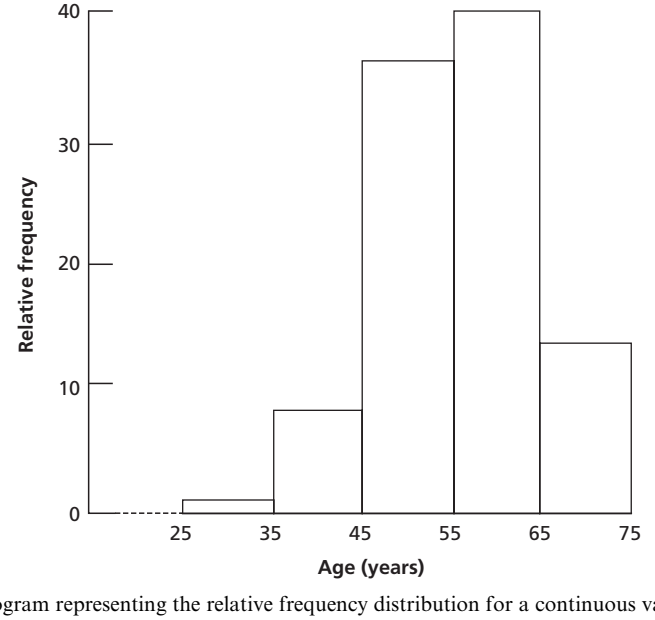


Fig. 2.8 Histogram representing the relative frequency distribution for a continuous variable (age of 1357 men with lung cancer, Table 2.4). Note that the variable shown here is exact age. The age at last birthday is a discrete variable and would be represented by groups displaced half a year to the left from those shown here; see p. 32.

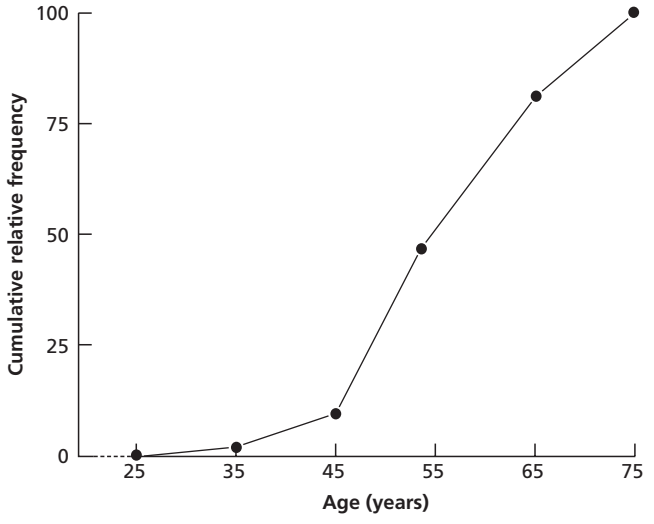


Fig. 2.9 Cumulative frequency plot for age of 1357 men with lung cancer (Table 2.4 and Fig. 2.8).

Table 2.5 Stem-and-leaf display for distribution of number of lesions caused by smallpox virus in egg membranes (see Table 2.3).

Number of lesions	
0*	7
1	024779
2	11122266678999
3	00223456788999
4	01233346677888999
5	11234478
6	014567779
7	057
8	023447
9	8
10	
11	2

the count, and in each row (or 'stem') the numbers to the right (the 'leaves') are the values of the second digit for the various observations in that group. Thus, the single observation in the first group is 7, and the observations in the second group are 10, 12, 14, 17, 17 and 19. The leaves have been ordered on each stem. The similarity in the shape of the stem-and-leaf display and the histogram in Fig. 2.7 is apparent.

The number of asterisks (*) indicates how many digits are required for each leaf. Thus, in Table 2.5, one asterisk is shown because the observations require only one digit from the leaf in addition to the row heading. Suppose that, in the distribution shown in Table 2.5, there had been four outlying values over 100: say, 112, 187, 191 and 248. Rather than having a large number of stems with no leaves and a few with only one leaf, it would be better to use a wider group interval for these high readings. The observations over 100 could be shown as:

```
1**   12,   87,   91
2     48
```

Sometimes it might be acceptable to drop some of the less significant digits. Thus, if such high counts were needed only to the nearest 10 units, they could be displayed as:

```
1**   199
2       5
```

representing 110, 190, 190 and 250.

For other variants on stem-and-leaf displays, see Tukey (1977).

If the main purpose of a visual display is to compare two or more distributions, the histogram is a clumsy tool. Superimposed histograms are usually confusing, and spatially separated histograms are often too distant to provide a means of comparison. The dot diagram of Fig. 2.5 or the box-and-whisker plot of Fig. 2.14 (p. 39) is preferable for this purpose.

Alternatively, use may be made of the representation of three-dimensional figures now available in some computer programs; an example is shown in Fig. 2.10 of a bar diagram plotted against two variables simultaneously. With this representation care must be taken not to mislead because of the effects of perspective. Some computer packages produce a three-dimensional effect even for bar charts (such as Fig. 2.1), histograms and pie charts. While the third dimension provides no extra information here, the effect can be very attractive.

The frequency in a distribution or in a histogram is often expressed not as an absolute count but as a relative frequency, i.e. as a proportion or percentage of the total frequency. If the standard grouping of the variable in terms of which the frequencies are expressed is a single unit, the total area under the histogram will be 1 (or 100% if percentage frequencies are used), and the area between any two points will be the relative frequency in this range.

Suppose we had a frequency distribution of heights of 100 men, in 1 cm groups. The relative frequencies would be rather irregular, especially near the extremes of the distribution, owing to the small frequencies in some of the groups. If the number of observations were increased to, say, 1000, the trend of the frequencies would become smoother and we might then reduce the grouping

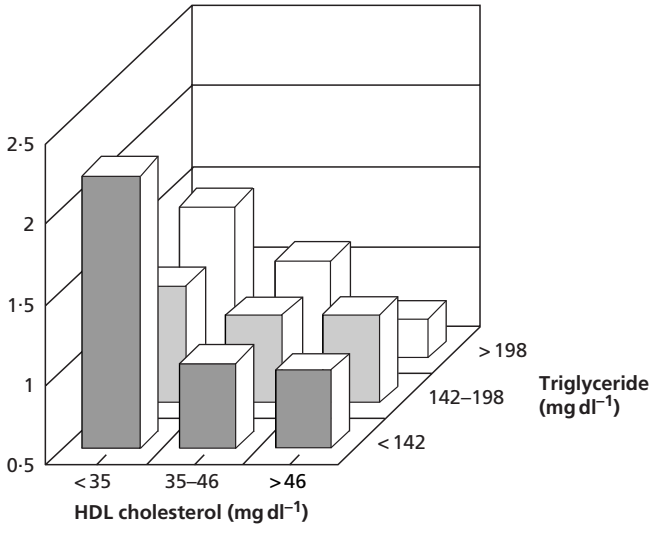


Fig 2.10 A 'three-dimensional' bar diagram showing the relative risk of coronary heart disease in men, according to high-density lipoprotein (HDL) cholesterol and triglyceride (reproduced from Simons *et al*., 1991, by permission of the authors and publishers).

to 0·5 cm, still making the vertical scale in the histogram represent the relative frequency per cm. We could imagine continuing this process indefinitely if there were no limit to the fineness of the measurement of length or to the number of observations we could make. In this imaginary situation the histogram would approach closer and closer to a smooth curve, the *frequency curve*, which can be thought of as an idealized form of histogram (Fig. 2.11). The area between the ordinates erected at any two values of the variable will represent the relative frequency of observations between these two points. These frequency curves are useful as models on which statistical theory is based, and should be regarded as idealized approximations to the histograms which might be obtained in practice with a large number of observations on a variable which can be measured extremely accurately.

We now consider various features which may characterize frequency distributions. Any value of the variable at which the frequency curve reaches a peak is called a *mode*. Most frequency distributions encountered in practice have one peak and are described as *unimodal*. For example, the distribution in Fig. 2.6 has a mode at four males, and that in Table 2.3 at 40–49 lesions. Usually, as in these two examples, the mode occurs somewhere between the two extremes of the distribution. These extreme portions, where the frequency becomes low, are called *tails*. Some unimodal distributions have the mode at one end of the range. For instance, if the emission of γ-particles by some radioactive material

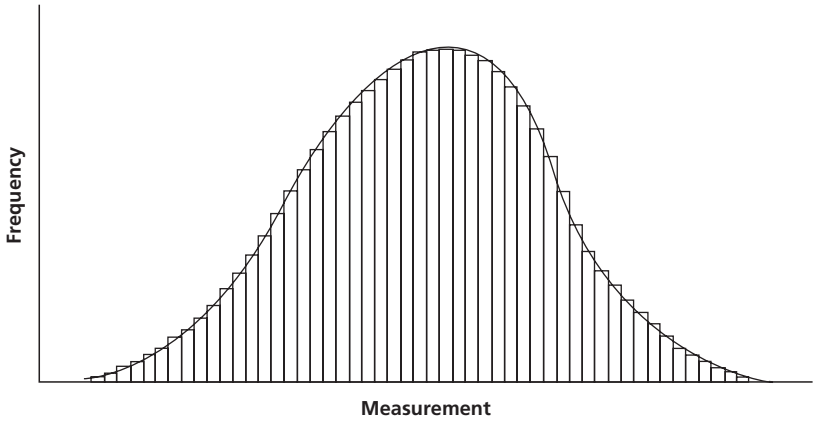


Fig. 2.11 Histogram representing the frequency distribution for a very large number of measurements finely subdivided, with an approximating frequency curve.

is being studied, the frequency distribution of the time interval between successive emissions is shaped like a letter *J* (or rather its mirror image), with a mode at zero. Similarly, if we take families with four children and record the numbers of families in which there have been 0, 1, 2, 3 or 4 cases of poliomyelitis, we shall find a very pronounced mode at zero.

Some distributions will appear to have more than one mode owing to the inevitable random fluctuations of small numbers. In Table 2.3, for example, the observed frequencies show subsidiary modes at 60–69, 80–89 and 110–119, although we should be inclined to pay no great attention to these. Occasionally, even with very large numbers of observations, distributions with more than one mode are found. There has been a great deal of discussion as to whether the distribution of casual blood pressure in a large population free from known circulatory diseases is *bimodal*, i.e. has two modes, because the presence of a second mode at blood pressures higher than the principal mode might indicate that a substantial proportion of the population suffered from essential hypertension.

Another characteristic of some interest is the symmetry or lack of symmetry of the distribution. An asymmetric distribution is called *skew*. The distribution in Fig. 2.6 is fairly symmetrical about the mode. That in Table 2.3 and Fig. 2.7, in which the upper tail is longer than the lower, would be called *positively* skew. The distribution in Table 2.4 and Fig. 2.8 has a slight negative skewness.

Two other characteristics of distributions are of such importance that separate sections will be devoted to them. They are the general location of the distribution on the scale of the variable, and the degree of variation of the observations about the centre of the distribution. Indeed, measures of location and variation are of such general importance that we shall discuss them first in relation to the original, ungrouped, observations, before referring again to frequency distributions.

2.4 **Means and other measures of location**

It is often important to give, in a single figure, some indication of the general level of a series of measurements. Such a figure may be called a *measure of location*, a *measure of central tendency*, a *mean* or an *average*. The most familiar of these measures is the *arithmetic* mean, and is customarily referred to as the 'average'. In statistics the term is often abbreviated to 'mean'.

The mean is the sum of the observations divided by the number of observations. It is awkward to have to express rules of calculation verbally in this way, and we shall therefore digress a little to discuss a convenient notation. A single algebraic symbol, like x or y, will often be used to denote a particular variable. For each variable there may be one or more observations. If there are n observations of variable x they will be denoted by

$$x_1, x_2, x_3, \ldots, x_{n-1}, x_n.$$

The various xs are not necessarily or even usually arranged in order of magnitude. They may be thought of as arranged in the order in which they were calculated or observed. It is often useful to refer to a typical member of the group in order to show some calculation which is to be performed on each member. This is done by introducing a 'dummy' suffix, which will often be i or j. Thus, if x is the height of a schoolchild and $x_1, x_2, \ldots, x_n$ are n values of x, a typical value may be denoted by x_i.

The arithmetic mean of the xs will be denoted by $\bar{x}$ (spoken 'x bar'). Thus,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

The summation occurring in the numerator can be denoted by use of the *summation sign* $\sum$ (the capital Greek letter 'sigma'), which means 'the sum of'. The range of values taken by the dummy suffix is indicated above and below the summation sign. Thus,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

and

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

If, as in this instance, it is clear which values the dummy suffix assumes throughout the summation, the range of the summation, and even occasionally the dummy suffix, may be omitted. Thus,

$$\sum_{i=1}^{n} x_i$$

may be abbreviated to $\sum x_i$ or to $\sum x$. Sometimes the capital letter S is used instead of $\sum$. It is important to realize that $\sum$ stands for an operation (that of obtaining the sum of quantities which follow), rather than a quantity itself.

The mean of a series of observations can be calculated readily on a pocket calculator or computer. Occasionally data may be presented merely in the form of a frequency distribution such as those shown in Tables 2.3 and 2.4. If the original data are available they should certainly be retrieved for calculation of the mean. If the original data are not retrievable, the mean may be estimated (although not with complete accuracy) by assuming that the observations are all clustered at the centres of their grouping intervals. Thus, in Table 2.3, the observations in the group 20–, ranging potentially from 20 to 29, could all be assumed to be 24·5; in Table 2.4, those in the group 45– years, ranging from 45 to 54, could be assumed to be 49·5. The calculated mean is then a *weighted mean* of the mid-points of the age groups (see §8.2). The reference to Table 2.4 draws attention to an unusual feature of ages, which are commonly given as integers, showing the age at last birthday. If the intention were to measure the exact age (which is not usually the case), the mean could be estimated (but again with incomplete accuracy) by adding half a year; in the frequency distribution the mid-point of the group 45– would then be taken to be 50.

Another useful measure of location is the *median*. If the observations are arranged in increasing or decreasing order, the median is the middle observation. If the number of observations, n, is odd, there will be a unique median—the $\frac{1}{2}(n+1)$th observation from either end. If n is even, there is strictly no middle observation, but the median is defined by convention as the mean of the two middle observations—the $\frac{1}{2}n$th and the $(\frac{1}{2}n+1)$th from either end.

The median has several disadvantages in comparison with the mean.

1 It takes no account of the precise magnitude of most of the observations, and is therefore usually less efficient than the mean because it wastes information.

2 If two groups of observations are pooled, the median of the combined group cannot be expressed in terms of the medians of the two component groups. This is not so with the mean. If groups containing n_1 and n_2 observations have means of $\bar{x}_1$ and $\bar{x}_2$, respectively, the mean of the combined group is the weighted mean:

$$(n_1\bar{x}_1 + n_2\bar{x}_2)/(n_1 + n_2).$$

3 The median is much less amenable than the mean to mathematical treatment, and is not much used in the more elaborate statistical techniques.

For descriptive work, however, the median is occasionally useful. Consider the following series of durations (in days) of absence from work owing to sickness:

1, 1, 2, 2, 3, 3, 4, 4, 4, 4, 5, 6, 6, 6, 6, 7, 8, 10, 10, 38, 80.

From a purely descriptive point of view the mean might be said to be misleading. Owing to the highly skew nature of the distribution the mean of 10 days is not really typical of the series as a whole, and the median of 5 days might be a more useful index. Another point is that in skew distributions of this type the mean is very much influenced by the presence of isolated high values. The median is therefore more stable than the mean in the sense that it is likely to fluctuate less from one series of readings to another.

The median of a grouped frequency distribution may be estimated (with incomplete accuracy) from the cumulative frequency plot or from the tabulated cumulative frequencies. In Fig. 2.9, for example, the median is the point on the age axis corresponding to 50% on the cumulative scale. In this example, based on the data from Table 2.4, the mean (of actual age, rather than age last birthday) is estimated as 55·7 years, and the median as 56·0 years.

The median and mean are equal if the series of observations is symmetrically distributed about their common value (as is nearly the case in Table 2.4). For a positively skew distribution (as in Table 2.3) the mean will be greater than the median, while if the distribution is negatively skew the median will be the greater.

A third measure of location, the mode, was introduced in §2.3. It is not widely used in analytical statistics, mainly because of the ambiguity in its definition as the fluctuations of small frequencies are apt to produce spurious modes.

Finally, reference should be made to two other forms of average, which are occasionally used for observations taking positive values only. The *geometric mean* is used extensively in microbiological and serological research. It involves the use of logarithms, and is dealt with in §2.5. The *harmonic mean* is much more rarely used. It requires the replacement of each observation by its reciprocal (i.e. 1 divided by the observation), and the calculation of the arithmetic mean of these reciprocals; the harmonic mean is then the reciprocal of this quantity.

2.5 **Taking logs**

When a variable is restricted to positive values and can vary over a wide range, its distribution is often positively skew. The distribution of numbers of lesions shown in Table 2.3 provides an example. Other examples commonly occurring in medical statistics are concentrations of chemical substances in the blood or urine; and time intervals between two events. Intuitively, the skewness is unsurprising: the lower bound of zero prevents an unduly long left-hand tail, whereas no such constraint exists for the right-hand tail.

We noted in §2.4 that for skewly distributed variables, such as the duration of absence of work used as an illustration on p. 32, the mean may be a less satisfactory measure of location than the median. More generally, many of the methods of analysis to be described later in this book are more appropriate for

variables following an approximately symmetric distribution than for those with skew distributions.

A useful device in such situations is the *logarithmic* (or *log*) *transformation* (or *transform*). Logarithms are readily obtainable by a keystroke on a pocket calculator or as a simple instruction on a computer. In practical work it is customary to use *common* logarithms, to base 10, usually denoted by the key marked 'log' on a calculator. In purely mathematical work, and in many formulae arising in statistics, the *natural* logarithm, denoted by 'ln', may be more convenient (see p. 126).

The essential point about logs is that, when numbers are multiplied together, their logs are added. If a series of numbers increases by a constant multiplying factor, their logs must increase by a constant difference. This is shown by the following series of values of a variable x, and the corresponding values of $y = \log x$.

x	2	20	200	2000
y	0·3	1·3	2·3	3·3

These four values of x are highly skew, with increasing differences between the higher values, whereas the values of y are symmetrically distributed, with equal differences between adjacent values.

Observations made in microbiological and serological research are sometimes expressed as *titres*, which are the dilutions of certain suspensions or reagents at which a specific phenomenon, like agglutination of red cells, first takes place. If repeated observations are made during the same investigation, the possible values of a titre will usually be multiples of the same dilution factor: for example, 2, 4, 8, 16, etc., for twofold dilutions. It is commonly found that a series of titres, obtained, for example, from different sera, is distributed with marked positive skewness on account of the increasingly wide intervals between possible values. A series of titres 2, 4, 8, 16, etc., has logarithms very nearly equal to 0·3, 0·6, 0·9, 1·2, etc., which increase successively by an increment of 0·3 (= log 2). As in the earlier example, the use of log titres is likely to give a series of observations which is more symmetrically distributed than were the original titres.

As an alternative to the median, as a measure of location for this type of variable, the *geometric mean* is widely used. Denote the original readings by x, and the transformed values by y $(= \log x)$. The arithmetic mean, $\bar{y}$, is, like the individual values of y, measured on a logarithmic scale. To get back to the original scale of x, we take $\bar{x}_g$ = antilog $\bar{y}$. (The antilog is often marked '10^x' on a calculator.) The quantity $\bar{x}_g$ is the geometric mean of x. It can never be greater than the arithmetic mean, and the two means will be equal only if all the xs are the same. Note that the geometric mean cannot be used if any of the original observations are negative, since a negative number has no logarithm. Nor can it be used if any of the original readings are zero, since log 0 = minus infinity.

Figures 2.12 and 2.13 illustrate the effect of the log transform in reducing positive skewness. The distribution of serum creatinine levels shown in Fig. 2.12 has considerable positive skewness, as is characteristic of this type of measurement. The distribution of logs shown in Fig. 2.13 is more nearly symmetric. The arithmetic mean of the original data, calculated from the individual readings, is

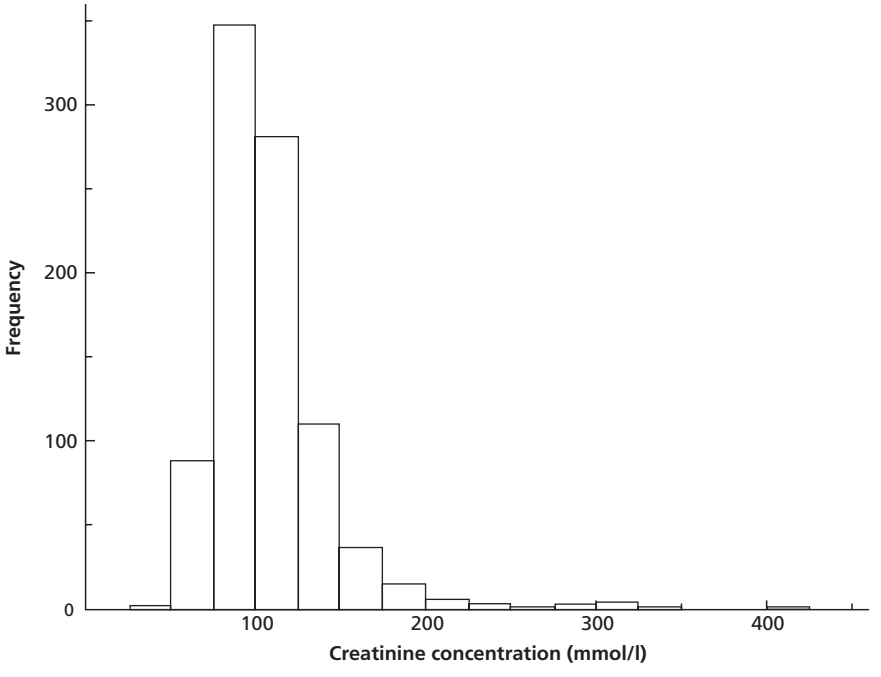


Fig. 2.12 Histogram showing the distribution of serum creatinine levels in a group of 901 subjects.

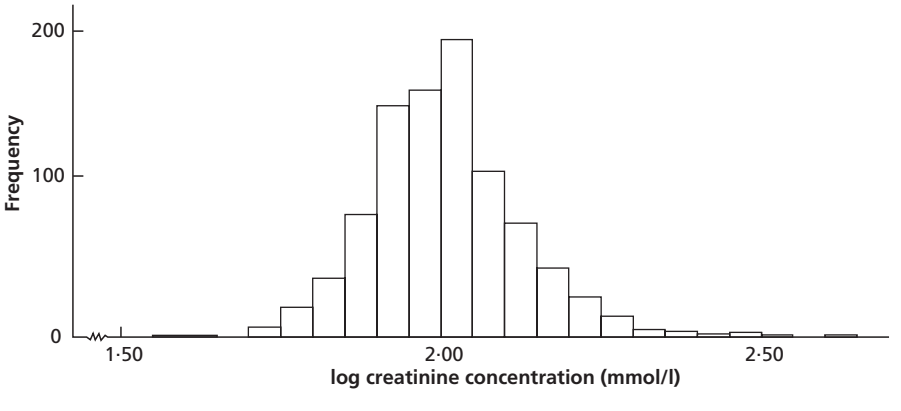


Fig. 2.13 Histogram showing the distribution of the logarithms of creatinine levels, the original values being depicted in Fig. 2.12.

107·01. The geometric mean, calculated from the logs and then transformed back to the original scale, is 102·36. The median is 100·00, closer to the geometric mean than to the arithmetic mean, as one might expect.

The log transform is one example of the way in which raw data can often usefully be transformed to a new scale of measurement before analyses are carried out. The use of transformations is described in more detail in §§10.8 and 11.10. In particular, the log transform will be seen to have other uses besides that of reducing positive skewness.

2.6 **Measures of variation**

When the mean value of a series of measurements has been obtained, it is usually a matter of considerable interest to express the degree of variation or scatter around this mean. Are the readings all rather close to the mean or are some of them scattered widely in each direction? This question is important for purely descriptive reasons, as we shall emphasize below. It is important also since the measurement of variation plays a central part in the methods of statistical inference which are described in this book. To take a simple example, the reliability of the mean of 100 values of some variable depends on the extent to which the 100 readings differ among themselves; if they show little variation the mean value is more reliable, more precisely determined, than if the 100 readings vary widely. The role of variation in statistical inference will be clarified in later chapters of this book. At present we are concerned more with the descriptive aspects.

In works of reference it is common to find a single figure quoted for the value of some biological quantity and the reader may not always realize that the stated figure is some sort of average. In a textbook on nutrition, for example, we might find the vitamin A content of Cheddar cheese given as 390 micrograms per 100 grams of cheese. Clearly, not all specimens of Cheddar cheese contain precisely 390 μg per 100 g; how much variation, then, is there from one piece of cheese to another? To take another example from nutrition, the daily energy requirement of a physically active man aged 25 years, of 180 cm and 73 kg, should be 12·0 megajoules. This requirement must vary from one person to another; how large is the variation?

There is unlikely to be a single answer to questions of this sort, because the amount of variation to be found in a series of measurements will usually depend on the circumstances in which they are made and, in particular, on the way in which these circumstances change from one reading to another. Specimens of Cheddar cheese are likely to vary in their vitamin A content for a number of reasons: major differences in the place and method of manufacture; variation in composition from one specimen to another even within the same batch of manufacture; the age of the cheese, and so on. Variation in the recorded measurement may be partly due to measurement error—in the method of

assay, for example, or because of observer errors. Similarly, if reference is made to the variation in systolic blood pressure it must be made clear what sort of comparison is envisaged. Are we considering differences between various types of individual (for example, groups defined by age or by clinical state); differences between individuals classified in the same group; or variation from one occasion to another in the same individual? And are the instrument and the observer kept constant throughout the series?

We now consider some methods of measuring the variation or scatter of a series of continuous measurements.

This scatter is, of course, one of the features of the data which is elucidated by a frequency distribution. It is, however, convenient to use some single quantity to measure this feature of the data, first for economy of presentation, secondly because the statistical methods to be described later require such an index, and thirdly because the data may be too sparse to enable a distribution to be formed. We therefore require what is variously termed a measure of *variation, scatter, spread* or *dispersion*.

An obvious candidate is the *range*, which is defined as the difference between the maximum value and the minimum value. Note that the range is a definite quantity, measured in the same units as the original observations; if the highest and lowest of a series of diastolic blood pressures are 95 and 65 mmHg, we may say not only (as in conversation) that the readings range from 65 to 95 mmHg, but also that the range is 30 mmHg. There are three main difficulties about the use of the range as a measure of variation. The first is that the numerical value assumed by the range is determined by only two of the original observations. It is true that, if we say that the minimum and maximum readings have the values 65 and 95, we are saying something about the other readings—that they are between these extremes—but apart from this, their exact values have no effect on the range. In this example, the range would be 30 whether: (i) all the other readings were concentrated between 75 and 80; or (ii) they were spread rather evenly between 65 and 95. A desirable measure of the variation of the whole set of readings should be greater in case (ii) than in case (i). Secondly, the interpretation of the range depends on the number of observations. If observations are selected serially from a large group (for example, by taking the blood pressures of one individual after another), the range cannot possibly decrease; it will increase whenever a new reading falls outside the interval between the two previous extremes. The interpretation of the range as a measure of variation of the group as a whole must therefore depend on a knowledge of the number of observations on which it is based. This is an undesirable feature; no such allowance is required, for instance, in the interpretation of a mean value as a measure of location. Thirdly, calculations based on extreme values are rather unreliable because big differences in these extremes are liable to occur between two similar investigations.

If the number of observations is not too small, a modification may be introduced which avoids the use of the absolute extreme values. If the readings are arranged in ascending or descending order, two values may be ascertained which cut off a small fraction of the observations at each end, just as the median breaks the distribution into two equal parts. The value below which a quarter of the observations fall is called the *lower quartile*, that which is exceeded by a quarter of the observations is called the *upper quartile*, and the distance between them is called the *interquartile range*. This measure is not subject to the second disadvantage of the range and is less subject to the other disadvantages.

The evaluation of the quartiles for a set of n values is achieved by first calculating the corresponding ranks by

$$r_l = \tfrac{1}{4}n + \tfrac{1}{2}$$

and

$$r_u = \tfrac{3}{4}n + \tfrac{1}{2}$$

and then calculating the quartiles as the corresponding values in the orderd set of values, using interpolation if necessary. For example, in the data on duration of absence from work (p. 32) where n is 21, $r_l = 5\frac{3}{4}$ and $r_u = 16\frac{1}{4}$. The lower quartile is then obtained by interpolation between the 5th and 6th values; these are both 3 days, so the lower quartile is also 3 days. The upper quartile is obtained by interpolation between the 16th and 17th values, 7 and 8 days. The interpolation involves moving $\frac{1}{4}$ of the way from the 16th value towards the 17th value, to give $7 + \frac{1}{4}(8 - 7) = 7\frac{1}{4}$ days. It should be noted that there is not a single standard convention for calculating the quartiles. Some authors define the quartiles in terms of the ranks $r_l = \frac{1}{4}(n + 1)$ and $r_u = \frac{3}{4}(n + 1)$. It is also common to round $\frac{1}{4}$ and $\frac{3}{4}$ to the nearest integer and to use interpolation only when this involves calculation of the mid-point between two values. Differences between the results using the different conventions are usually small and unimportant in practice.

The quartiles are particular examples of a more general index, the *percentile* (or *centile*). The value below which P% of the values fall is called the Pth percentile. Thus, the lower and upper quartiles are the 25th and 75th percentiles, respectively. The quartiles or any other percentiles could be read off a plot of cumulative relative frequency such as Fig. 2.9. The term *quantile* is used when the proportion is expressed as a fraction rather than a percentage: thus, the 30th percentile is the same as the 0·3 quantile.

A convenient method of displaying the location and variability of a set of data is the box-and-whisker plot. The basic form of this plot shows a box defined by the lower and upper quartiles and with the median marked by a subdivision of the box. The whiskers extend from both ends of the box to the minimum and

maximum values. Elaborations of this plot show possible outlying values (§2.7) separately beyond the ends of the whiskers by redefining the whiskers to have a maximum length in terms of the interquartile range. Box-and-whisker plots facilitate a visual comparison of groups. Figure 2.14 shows a comparison of birth weight between two groups of migrant Vietnamese mothers, one group following a traditional diet and the other a non-traditional diet. In this plot the whiskers extend to the most extreme observations within $\pm 1{\cdot}5$ interquartile ranges of the quartiles and more extreme points are plotted individually. The higher median and quartiles for the group with the non-traditional diet show clearly, as also does the lower variability associated with this diet.

An alternative approach is to make some use of all the deviations from the mean, $x_i - \bar{x}$. Clearly, the greater the scatter of the observations the greater will the magnitude of these deviations tend to be. It would be of no use to take the mean of the deviations $x_i - \bar{x}$, since some of these will be negative and some positive. In fact,

$$\begin{aligned} \sum(x_i - \bar{x}) &= \sum x_i - \sum \bar{x} \\ &= \sum x_i - n\bar{x} \\ &= 0 \quad \text{since } \bar{x} = \sum x_i / n. \end{aligned}$$

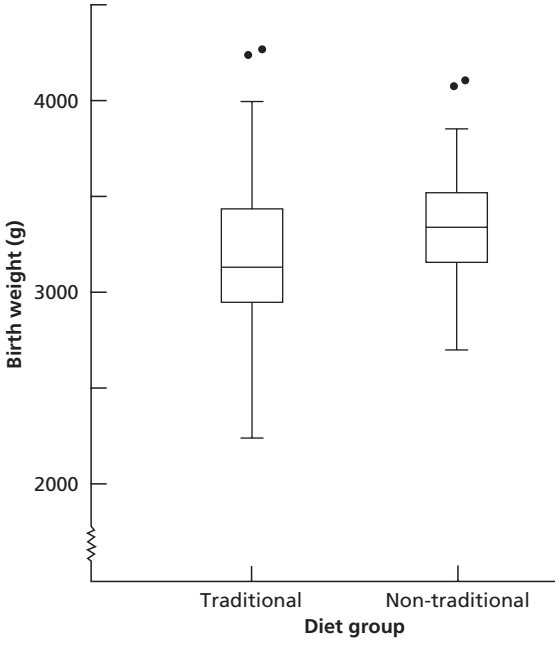


Fig. 2.14 Box-and-whisker plot comparing the distributions of birth weight of babies between two groups of Vietnamese mothers (data of Mitchell and Mackerras, 1995).

Therefore the mean of the deviations $x_i - \bar{x}$ will always be zero. We could, however, take the mean of the deviations ignoring their sign, i.e. counting them all as positive. These quantities are called the absolute values of the deviations and are denoted by $|x_i - \bar{x}|$. Their mean, $\sum |x_i - \bar{x}|/n$, is called the *mean deviation*. This measure has the drawback of being difficult to handle mathematically, and we shall not consider it any further in this book.

Another way of getting over the difficulty caused by the positive and negative signs is to square them. The mean value of the squared deviations is called the *variance* and is a most important measure in statistics. Its formula is

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n}. \tag{2.1}$$

The numerator is often called the *sum of squares about the mean*. The variance is measured in the square of the units in which x is measured. For example, if x is height in cm, the variance will be measured in cm^2. It is convenient to have a measure of variation expressed in the original units of x, and this can be easily done by taking the square root of the variance. This quantity is known as the *standard deviation*, and its formula is

$$\text{Standard deviation} = \sqrt{\left[\frac{\sum (x_i - \bar{x})^2}{n}\right]}. \tag{2.1a}$$

In practice, in calculating variances and standard deviations, the n in the denominator is almost always replaced by $n - 1$. The reason for this is that in applying the methods of statistical inference, developed later in this book, it is useful to regard the collection of observations as being a *sample* drawn from a much larger group of possible readings. The large group is often called a *population*. When we calculate a variance or a standard deviation we may wish not merely to describe the variation in the sample with which we are dealing, but also to estimate as best we can the variation in the population from which the sample is supposed to have been drawn. In a certain respect (see §5.1) a better estimate of the population variance is obtained by using a divisor $n - 1$ instead of n. Thus, we shall almost always use the formula for the *estimated variance* or *sample variance*:

$$\text{Estimated variance}, s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}, \tag{2.2}$$

and, similarly,

$$\text{Estimated standard deviation}, s = \sqrt{\left[\frac{\sum (x_i - \bar{x})^2}{n - 1}\right]}. \tag{2.2a}$$

Having established the convention we shall very often omit the word 'estimated' and refer to s^2 and s as 'variance' and 'standard deviation', respectively.

The modification of the divisor from n to $n-1$ is clearly not very important when n is large. It is more important for small values of n. Although the theoretical justification will be discussed more fully in §5.1, two heuristic arguments may be used now which may make the divisor $n-1$ appear more plausible. First, consider the case when $n=1$; that is, there is a single observation. Formula (2.1) with a divisor n gives a variance $0/1=0$. Now, this is a reasonable expression of the complete absence of variation in the available observation: it cannot differ from itself. On the other hand, a single observation provides no information at all about the variation in the population from which it is drawn, and this fact is reflected in the calculation of the estimated variance, s^2, from (2.2), which becomes 0/0, an indeterminate quantity.

Secondly, in the general case when n takes any value, we have already seen that $\sum(x_i-\bar{x})=0$. This means that if $n-1$ of these deviations $x_i-\bar{x}$ are chosen arbitrarily, the nth is determined automatically. (It is the sum of the $n-1$ chosen values of $x_i-\bar{x}$ with the sign changed.) In other words, only $n-1$ of the n deviations which are squared in the numerator of (2.1) or (2.2) are *independent*. The divisor $n-1$ in (2.2) may be regarded as the number of independent quantities among the sum of squared deviations in the numerator. The divisor $n-1$ is, in fact, a particular case of a far-reaching concept known as the *degrees of freedom* of an estimate of variance, which will be developed in §5.1.

The direct calculation of the estimated variance is illustrated in Table 2.6. In this particular example the calculation is fairly straightforward. In general, two features of the method are likely to cause trouble. Errors can easily arise in the subtraction of the mean from each reading. Further, if the mean is not a 'round' number, as it was in this example, it will need to be rounded off. The deviations

Table 2.6 Calculation of estimated variance and standard deviation: direct formula.

x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$
8	0	0
5	−3	9
4	−4	16
12	4	16
15	7	49
5	−3	9
7	−1	1
$\sum x_i=56$		$\sum(x_i-\bar{x})^2=100$
$n=7$		
$\bar{x}=56/7=8$		
$s^2=100/6=16{\cdot}67$		
$s=\sqrt{16{\cdot}67}=4{\cdot}08$		

$x_i - \bar{x}$ will then need to be written with several significant digits and doubt will arise as to whether an adequate number of significant digits was retained for $\bar{x}$. These difficulties have led to the widespread use of an alternative method of calculating the sum of squares about the mean, $\sum(x_i - \bar{x})^2$. It is based on the fact that

$$\sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}. \tag{2.3}$$

This is called the short-cut formula for the sum of squares about the mean.

The important point about (2.3) is that the computation is performed without the need to calculate individual deviations from the mean, $x_i - \bar{x}$. The sum of squares of the original observations, x_i, is corrected by subtraction of a quantity dependent only on the mean (or, equivalently, the total) of the x_i. This second term is therefore often called the *correction term*, and the whole expression a *corrected* sum of squares.

The previous example is reworked in Table 2.7. We again have the result $\sum(x_i - \bar{x})^2 = 100$, and the subsequent calculations follow as in Table 2.6.

The short-cut formula avoids the need to square individual deviations with many significant digits, but involves the squares of the x_i, which may be large numbers. This rarely causes trouble using a calculator, although care must be taken to carry sufficient digits in the correction term to give the required number of digits in the difference between the two terms. (For example, if $\sum x_i^2 = 2025$ and $(\sum x_i)^2/n = 2019{\cdot}3825$, the retention of all these decimals will give

Table 2.7 Calculation of estimated variance and standard deviation: short-cut formula (same data as in Table 2.6).

x_i	x_i^2
8	64
5	25
4	16
12	144
15	225
5	25
7	49
—	—
56	548

$$\begin{aligned}\sum(x_i - \bar{x})^2 &= \sum x_i^2 - (\sum x_i)^2/n \\ &= 548 - 56^2/7 \\ &= 548 - 448 \\ &= 100\end{aligned}$$

Subsequent steps as in Table 2.6

$\sum(x_i - \bar{x})^2 = 5{\cdot}6175$; if the correction term had been rounded off to the nearest whole number it would have given $\sum(x_i - \bar{x})^2 = 6$—an accuracy of only 1 significant digit.) Indeed, this rounding error can cause problems in high-speed computing, and computer programs should use the direct rather than the short-cut formula for the sum of squares about the mean.

Most scientific calculators have a summation key which accumulates n, $\sum x$ and $\sum x^2$ in stores and then the mean and standard deviation are each calculated by successive keystrokes, the latter invoking calculation of (2.3). Calculators usually have separate keys for (2.1a) and (2.2a), often called σ_n and σ_{n-1}, respectively.

If the observations are presented in the form of a frequency distribution, and the raw data are not retrievable, the standard deviation may be obtained, as indicated earlier for the calculation of the mean, by assuming that all observations are clustered at the mid-points of their grouping intervals. As with the mean, the true standard deviation can then only be estimated, with some loss of accuracy. There is an additional problem with the standard deviation and variance, in that the calculation from grouped data tends to give somewhat too high a value. An appropriate correction for this effect, called *Sheppard's correction*, is to subtract $\frac{1}{12}h^2$ from the calculated variance, h being the size of the grouping interval. (If the grouping interval is not constant an average value may be used.) The correction is rather small unless the grouping is quite crude, and for this reason is often ignored. In the example based on the data of Table 2.4, for which $h = 10$, the standard deviation is estimated as 8·76 years, or 8·27 years after applying Sheppard's correction.

The standard deviation of a set of measurements is expressed in the same units as the measurements and hence in the same units as the mean. It is occasionally useful to describe the variability by expressing the standard deviation as a proportion, or a percentage, of the mean. The resulting measure, called the *coefficient of variation*, is thus a dimensionless quantity—a pure number. In symbols,

$$\mathrm{CV}(x) = \frac{s}{\bar{x}} \times 100\%. \tag{2.4}$$

The coefficient of variation is most useful as a descriptive tool in situations in which a change in the conditions under which measurements are made alters the standard deviation in the same proportion as it alters the mean. The coefficient of variation then remains unchanged and is a useful single measure of variability. It is mentioned again in a more substantive context in §5.3.

In §2.5 we described the geometric mean. By analogy, the *geometric standard deviation* may be obtained by calculating the standard deviation of the logarithmically transformed measurements and converting back to the original scale by

taking the antilog. The resulting quantity is a factor by which the geometric mean may be multiplied or divided to give a typically deviant value, and is most useful for skew distributions where the variation may be more symmetric after logarithmic transformation (as in Figs 2.12 and 2.13).

2.7 **Outlying observations**

Occasionally, as noted in §2.2, a single observation is affected by a gross error, either of measurement or recording, or due to a sudden lapse from the general standards of investigation applying to the rest of the data. It is important to detect such errors if possible, partly because they are likely to invalidate the assumptions underlying standard methods of analysis and partly because gross errors may seriously distort estimates such as mean values.

Some errors will result in recorded observations that are entirely plausible, and they are likely to be very difficult to detect. The position is more hopeful if (as is often the case) the error leads to an outlying observation, far distant from other comparable observations. Of course, such outliers are not necessarily the result of error. They may give valuable information about a patient's health at the time of the observation. In any case, outliers should be investigated and the reasons for their appearance should be sought. We therefore need methods for detecting outlying observations so that appropriate action can be taken.

Recording errors can often be reduced by careful checking of all the steps at which results are copied on to paper, and by minimizing the number of transcription steps between the original record and the final data. Frequently the original measurement will come under suspicion. The measurement may be repeatable (for example, the height of an individual if a short time has elapsed since the first measurement), or one may be able to check it by referring to an authoritative document (such as a birth certificate, if age is in doubt). A much more difficult situation arises when an observation is strongly suspected of being erroneous but no independent check is available. Possible courses of action are discussed below, at (b) and (c). First we discuss some methods of detecting gross errors.

(a) Logical checks

Certain values may be seen to be either impossible or extremely implausible by virtue of the meaning of the variable. Frequently the range of variation is known sufficiently well to enable upper and/or lower limits to be set; for example, an adult man's height below, say, 140 cm or above 205 cm would cause suspicion. Other results would be impossible because of relationships between variables; for example, in the UK a child aged 10 years cannot be married. Checks of this sort

can be carried out routinely, once the rules of acceptability have been defined. They can readily be performed by computer, possibly when the data are being entered (p.17); indeed, editing procedures of this sort should form a regular part of the analysis of large-scale bodies of data by computer.

(b) Statistical checks

Certain observations may be found to be unusual, not necessarily on a priori grounds, but at least by comparison with the rest of the data. Whether or not they are to be rejected or amended is a controversial matter to be discussed below; at any rate, the investigator will probably wish to have his/her attention drawn to them.

A good deal of statistical checking can be done quite informally by graphical exploration and the formation of frequency distributions. If most of the observations follow a simple bell-shaped distribution such as the *normal* (§3.8), with one or two aberrant values falling well away from the main distribution, the *normal plot* described in §11.9 is a useful device. Sometimes, observations are unusual only when considered in relation to other variables; for example, in an anthropometric survey of schoolchildren, a weight measurement may be seen to be unusually low or high in relation to the child's height; checking rules for weights in terms of heights will be much more effective than rules based on weights alone (Healy, 1952; see also §12.3). The detection of outliers that are apparent only when other variables are considered forms part of the diagnostic methods used in multiple regression (§11.9).

Should a statistically unusual reading be rejected or amended? If there is some external reason for suspicion, for example that an inexperienced technician made the observation in question, or that the air-conditioning plant failed at a certain point during an experiment, common sense would suggest the omission of the observation and (where appropriate) the use of special techniques for making adjustments for the missing data (§18.6). When there is no external reason for suspicion there are strong arguments for retaining the original observations. Clearly there are some observations which no reasonable person would retain—for example, an adult height recorded as 30 cm. However, the decision will usually have to be made as a subjective judgement on ill-defined criteria which depend on one's knowledge of the data under study, the purpose of the analysis and so on. A full treatment of this topic, including descriptions of more formal methods of dealing with outliers, is given by Barnett and Lewis (1994).

(c) Robust estimation

In some analyses the question of rejection or correction may not arise, and yet it may be suspected or known that occasional outliers occur, and some safeguard

against their effect may be sought. We shall discuss in §4.2 the idea of estimating the mean of a large population from a sample of observations drawn from it. In most situations we should be content to do this by calculating the mean of the sample values, but we might sometimes seek an estimator that is less influenced than the sample mean by occasional outliers. This approach is called *robust estimation*, and a wide range of such estimators has been suggested. In §2.4 we commended the sample median as a measure of location on the grounds that it is less influenced by outliers than is the mean. For positive-valued observations, the logarithmic transformation described in §2.5, and the use of the geometric mean, would have a similar effect in damping down the effect of outlying high values, but unfortunately it would have the opposite effect of exaggerating the effect of outlying low values. One of the most widely used robust measures of location is the *trimmed mean*, obtained by omitting some of the most extreme observations (for example, a fixed proportion in each tail) and taking the mean of the rest. These estimators are remarkably efficient for samples from normal distributions (§3.8), and better than the sample mean for distributions 'contaminated' with a moderate proportion of outliers. The choice of method (e.g. the proportion to be trimmed from the tails) is not entirely straightforward, and the precision of the resulting estimator may be difficult to determine.

Similar methods are available for more complex problems, although the choice of method is, as in the simpler case of the mean, often arbitrary, and the details of the analysis may be complicated. See Draper and Smith (1998, Chapter 25) for further discussion in the case of multiple regression (§11.6).

3 Probability

3.1 The meaning of probability

A clinical trial shows that 50 patients receiving treatment A for a certain disease fare better, on the average, than 50 similar patients receiving treatment B. Is it safe to assume that treatment A is really better than treatment B for this condition? Should the investigator use A rather than B for future patients? These are questions typical of those arising from any statistical investigation. The first is one of inference: what conclusions can reasonably be drawn from this investigation? The second question is one of decision: what is the rational choice of future treatment, taking into account the information provided by the trial and the known or unknown consequences of using an inferior treatment? The point to be emphasized here is that the answers to both questions, and indeed those to almost all questions asked about statistical data, are in some degree couched in uncertainty. There may be a very strong suggestion indeed that A is better than B, but can we be entirely sure that the patients receiving B were not more severely affected than those on A and that this variability between the patients was not a sufficient reason for their different responses to treatment? This possibility may, in any particular instance, seem unlikely, but it can rarely, if ever, be completely ruled out. The questions that have to be asked, therefore, must receive an answer phrased in terms of uncertainty. If the uncertainty is low, the conclusion will be firm, the decision will be safe. If the uncertainty is high, the investigation must be regarded as inconclusive. It is thus important to consider the measurement of uncertainty, and the appropriate tool for this purpose is the *theory of probability*. Initially the approach will be rather formal; later chapters are concerned with the application of probability theory to statistical problems of various types.

If a coin is tossed a very large number of times and the result of each toss written down, the results may be something like the following (*H* standing for heads and *T* for tails):

$$TTHTHHTHTTTHTHHTHHHHTTH\ldots$$

Such a sequence will be called a *random sequence* or *random series*, each place in the sequence will be called a *trial*, and each result will often be called an *event* or *outcome*. A random sequence of binary outcomes, such as *H* and *T* in this

example, is sometimes called a *Bernoulli sequence* (James Bernoulli, 1654–1705). A random sequence is characterized by a complete lack of pattern or of predictability. In coin tossing the chance of finding H at any one stage is just the same as at any other stage, and is quite uninfluenced by the outcomes of the previous tosses. (Contrary to some people's intuition, the chance of getting a head would be neither raised nor lowered by a knowledge that there had just occurred a run of, say, six tails.)

In such a sequence it will be found that as the sequence gets larger and larger the proportion of trials resulting in a particular outcome becomes less and less variable and settles down closer to some limiting value. This long-run proportion is called the *probability* of the particular outcome. Figure 3.1 shows the proportion of heads after various numbers of tosses, in an actual experiment. Clearly the proportion is settling down close to $\frac{1}{2}$, and it would be reasonable to say that the probability of a head is about $\frac{1}{2}$. Considerations of symmetry would, of course, have led us to this conclusion before seeing the experimental results. The slight differences between the indentations on the two sides of a coin, possibly variations in density, and even some minor imbalance in the tossing method, might make the probability very slightly different from $\frac{1}{2}$, but we should be unlikely ever to do a tossing experiment sufficiently long to distinguish between a probability of 0·5 and one of, say, 0·5001.

The reader will observe that this definition of probability is rather heuristic. We can never observe a sequence of trials and say unambiguously 'This is a

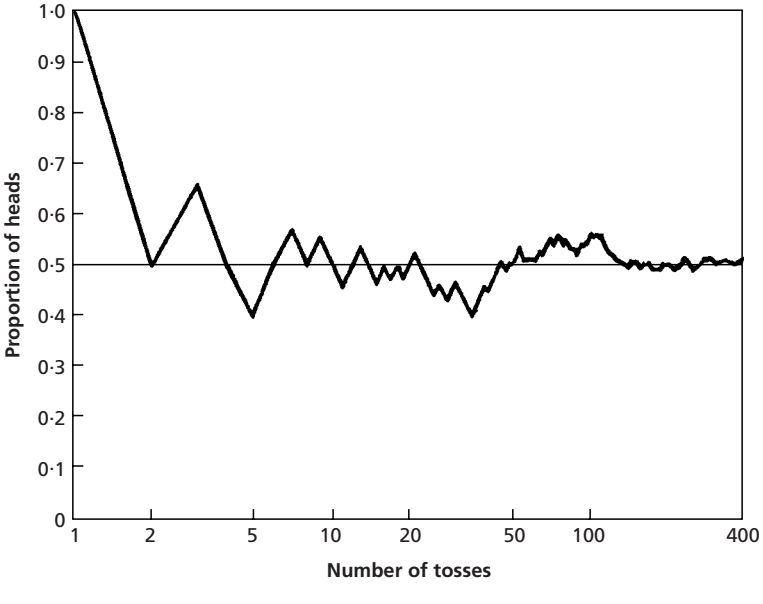


Fig. 3.1 Proportion of heads in a sequence of tosses of a coin with a logarithmic scale for the number of tosses (reprinted from Cramér, 1946, by permission of the author and publishers).

random sequence'; we observe only a finite portion of the sequence and there may be underlying patterns in the outcomes which cannot readily be discerned. Nor can we observe a sequence and state precisely the probability of a certain outcome; the probability is a long-run property and again an insufficient portion of the sequence is observed. Nevertheless, there are many phenomena which apparently behave in this way, and the concept of a random sequence should be regarded as an idealistic concept which apparently describes these phenomena very faithfully indeed. Here are some other examples of empirical 'random' (Bernoulli) sequences.

1 *The throws of a dice*. If the dice is well made, the probability of each outcome, 1–6, will be very close to $\frac{1}{6}$.
2 *The sex of successive live births occurring in a large human population*. The probability of a male is known to vary from population to population (for example, it depends on the stillbirth rate) but it is usually a little over $\frac{1}{2}$. In England and Wales it is currently about 0·515; the probability of a female birth is correspondingly about $1 - 0{\cdot}515 = 0{\cdot}485$.
3 *A sequence of outcomes of antenatal screening examinations, each classified by whether or not a specific fetal abnormality is present*. Thus, the probability that open spina bifida is present might be about 0·0012.

A consequence of this definition of probability is that it is measured by a number between 0 and 1. If an event never occurs in any of the trials in a random sequence, its probability is zero. If it occurs in every trial, its probability is unity.

A further consequence is that probability is not defined for sequences in which the succession of events follows a manifestly non-random pattern. If a machine were invented which tossed heads and tails in strict alternation (H, T, H, T, H, T, ..., etc.), the long-run proportions of heads and tails would both be $\frac{1}{2}$, but it would be incorrect to say that the probabilities of these events were $\frac{1}{2}$. The sequence is non-random because the behaviour of the odd-numbered trials is different from that of the even-numbered trials. It would be better to think of the series as a mixture of two separate series: the odd-numbered trials, in which the probability of H is 1, and the even-numbered trials, in which this probability is 0.

This concept of probability provides a measure of uncertainty for certain types of phenomena which occur in nature. There is a fairly high degree of certainty that any one future birth will not exhibit spina bifida because the probability for that event is low. There is considerable uncertainty about the sex of a future birth because the probabilities of both outcomes are about $\frac{1}{2}$. The definition is, however, much more restrictive than might be wished. What is the probability that smoking is a contributory cause of lung cancer? This question uses the word 'probability' in a perfectly natural conversational way. It does not, however, accord with our technical definition, for it is impossible to think of a random sequence of trials, in some of which smoking is a contributory cause of lung cancer and in some of which it is not.

It will appear in due course that the so-called 'frequency' definition of probability, which has been put forward above, can be used as the basis of statistical inference, and often some rewording of the question can shed some light on the plausibility of hypotheses such as 'Smoking is a contributory cause of lung cancer.' However, many theoretical statisticians, probabilitists and logicians advocate a much wider interpretation of the concept of probability than is permitted in the frequency definition outlined above. On this broader view, one should interpret probability as a measure of one's degree of belief in a proposition, and direct statements about the probability that a certain scientific hypothesis is true are quite in order. We return to this point of view in Chapter 6, but until that chapter is reached we shall restrict our attention to the frequency definition.

3.2 **Probability calculations**

The main purpose of allotting numerical values to probabilities is to allow calculations to be performed on these numbers. The two basic operations which concern us here are *addition* and *multiplication*, and we consider first the addition of probabilities.

Consider a random sequence of trials with more than one possible outcome for each trial. In a series of throws of a dice, for example, we might ask for the probability of *either* a 1 *or* a 3 being thrown. The answer is fairly clear. If the dice is perfectly formed, the probability of a 1 is $\frac{1}{6}$, and the probability of a 3 is $\frac{1}{6}$. That is, a 1 will appear in $\frac{1}{6}$ of trials in the long run, and a 3 will appear in the same proportion of a long series of trials. In no trial will a 1 *and* a 3 appear together. Therefore the compound event 'either a 1 or a 3' will occur in $\frac{1}{6} + \frac{1}{6}$, or $\frac{1}{3}$, of the trials in the long run. The probabilities for the two separate events have been added together.

Note the importance of the observation that a 1 and a 3 cannot both occur together; they are, in other words, *mutually exclusive*. Without this condition the simple form of the addition rule could not be valid. For example, if a doctor's name is chosen haphazardly from the *British Medical Register*, the probability that the doctor is male is about 0·8. The probability that the doctor qualified at an English medical school is about 0·6. What is the probability that the doctor either is male or qualified in England, or both? If the two separate probabilities are added the result is $0{\cdot}8 + 0{\cdot}6 = 1{\cdot}4$, clearly a wrong answer since probabilities cannot be greater than 1. The trouble is that the probability of the double event—male and qualified in England—has been counted twice, once as part of the probability of being male and once as part of the probability of being qualified in England. To obtain the right answer, the probability of the double event must be subtracted. Thus, denoting the two events by A and B, we have the more general form of the *addition rule*:

$$\begin{aligned}\text{Probability of } A \text{ or } B \text{ or both} = &(\text{Probability of } A)\\ &+ (\text{Probability of } B)\\ &- (\text{Probability of } A \text{ and } B).\end{aligned}$$

It will be convenient to write this as

$$\mathrm{P}(A \text{ or } B \text{ or both}) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \text{ and } B). \tag{3.1}$$

In this particular example the probability of the double event has not been given, but it must clearly be greater than 0·4, to ensure that the right side of the equation (3.1) is less than 1.

If the two events are mutually exclusive, the last term on the right of (3.1) is zero, and we have the *simple form of the addition rule*:

$$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B).$$

Suppose now that two random sequences of trials are proceeding simultaneously; for example, at each stage a coin may be tossed and a dice thrown. What is the *joint probability* of a particular combination of results, for example a head (H) on the coin and a 5 on the dice? The result is given by the *multiplication rule*:

$$\mathrm{P}(H \text{ and } 5) = \mathrm{P}(H) \times \mathrm{P}(5, \text{ given } H). \tag{3.2}$$

That is, the long-run proportion of pairs of trials in which both H and 5 occur is equal to the long-run proportion of trials in which H occurs on the coin, multiplied by the long-run proportion of those trials which occur with 5 on the dice. The second term on the right of (3.2) is an example of a *conditional probability*, the first event, 5, being 'conditional' on the second, H. A common notation is to replace 'given' by a vertical line, that is, $\mathrm{P}(5 \mid H)$.

In this particular example, there would be no reason to suppose that the probability of 5 on the dice was in the least affected by whether or not H occurred on the coin. In other words,

$$\mathrm{P}(5, \text{ given } H) = \mathrm{P}(5).$$

The conditional probability is equal to the unconditional probability. The two events are now said to be *independent*, and we have the *simple form of the multiplication rule*:

$$\begin{aligned}\mathrm{P}(H \text{ and } 5) &= \mathrm{P}(H) \times \mathrm{P}(5)\\ &= \tfrac{1}{2} \times \tfrac{1}{6}\\ &= \tfrac{1}{12}.\end{aligned}$$

Effectively, in this example, there are 12 combinations which occur equally often in the long run: *H1*, *H2*, ..., *H6*, *T1*, *T2*, ..., *T6*.

In general, pairs of events need not be independent, and the general form of the multiplication rule (3.2) must be used. In the earlier example we referred to the probability of a doctor being male and having qualified in England. If these events were independent, we could calculate this as

$$0{\cdot}8 \times 0{\cdot}6 = 0{\cdot}48,$$

and, denoting the events by A and B, (3.1) would give

$$\begin{aligned} \mathrm{P}(A \text{ or } B \text{ or both}) &= 0{\cdot}80 + 0{\cdot}60 - 0{\cdot}48 \\ &= 0{\cdot}92. \end{aligned}$$

These events may not be independent, however, since some medical schools are more likely to accept women than others. The correct value for P(A and B) could only be ascertained by direct investigation.

As another example of the lack of independence, suppose that in a certain large community 30% of individuals have blue eyes. Then

$$\begin{aligned} \mathrm{P}(\text{blue right eye}) &= 0{\cdot}3 \\ \mathrm{P}(\text{blue left eye}) &= 0{\cdot}3. \end{aligned}$$

P(blue right eye and blue left eye) is not given by

$$\mathrm{P}(\text{blue right eye}) \times \mathrm{P}(\text{blue left eye}) = 0{\cdot}09.$$

It is obtained by the general formula (3.2) as

$$\begin{aligned} &\mathrm{P}(\text{blue right eye}) \times \mathrm{P}(\text{blue left eye, given blue right eye}) \\ &\quad = 0{\cdot}3 \times 1{\cdot}0 \\ &\quad = 0{\cdot}3. \end{aligned}$$

Addition and multiplication may be combined in the same calculation. In the double sequence with a coin and a dice, what is the probability of getting either heads and 2 *or* tails and 4? Each of these combinations has a probability of $\frac{1}{12}$ (by the multiplication rule for independent events). Each combination is a possible outcome in the double sequence and the outcomes are mutually exclusive. The two probabilities of $\frac{1}{12}$ may therefore be added to give a final probability of $\frac{1}{6}$ that either one or the other combination occurs.

As a slightly more complicated example, consider the sex composition of families of four children. As an approximation, let us assume that the proportion of males at birth is 0·51, that all the children in the families may be considered as independent random selections from a sequence in which the probability of a boy is 0·51, and that the question relates to all liveborn infants so that differential survival does not concern us. The question is, what are the probabilities that a family of four contains no boys, one boy, two boys, three boys and four boys?

The probability that there will be no boys is the probability that each of the four children will be a girl. The probability that the first child is a girl is $1 - 0{\cdot}51 = 0{\cdot}49$. By successive applications of the multiplication rule for independent events, the probability that the first two are girls is $(0{\cdot}49)^2$; the probability that the first three are girls is $(0{\cdot}49)^3$; and the probability that all four are girls is $(0{\cdot}49)^4 = 0{\cdot}0576$. About 1 in 17 of all families of four will consist of four girls. Write this

$$\mathrm{P}(GGGG) = 0{\cdot}0576.$$

A family with one boy and three girls might arise in any of the following ways, *BGGG*, *GBGG*, *GGBG*, *GGGB*, according to which of the four children is the boy. Each of these ways has a probability of $(0{\cdot}49)^3(0{\cdot}51) = 0{\cdot}0600$. The total probability of one boy is therefore, by the addition rule,

$$\begin{aligned} &0{\cdot}0600 + 0{\cdot}0600 + 0{\cdot}0600 + 0{\cdot}0600 \\ &\quad = 4(0{\cdot}0600) \\ &\quad = 0{\cdot}2400. \end{aligned}$$

A family with two boys and two girls might arise in any of the following ways: *BBGG*, *BGBG*, *BGGB*, *GBBG*, *GBGB*, *GGBB*. Each of these has a probability of $(0{\cdot}49)^2(0{\cdot}51)^2$, and the total probability of two boys is

$$6(0{\cdot}49)^2(0{\cdot}51)^2 = 0{\cdot}3747.$$

Similarly for the other family composition types. The complete results are shown in Table 3.1.

Note that the five probabilities total to 1, as they should since this total is the probability that one or other of the five family composition types arises (these being mutually exclusive). Since these five types exhaust all the possibilities, the total probability must be unity. If one examined the records of a very large number of families experiencing four live births, would the proportions of

Table 3.1 Calculation of probabilities of families with various sex compositions.

Composition		
Boys	Girls	Probability
0	4	$(0{\cdot}49)^4 = 0{\cdot}0576$
1	3	$4(0{\cdot}49)^3(0{\cdot}51) = 0{\cdot}2400$
2	2	$6(0{\cdot}49)^2(0{\cdot}51)^2 = 0{\cdot}3747$
3	1	$4(0{\cdot}49)(0{\cdot}51)^3 = 0{\cdot}2600$
4	0	$(0{\cdot}51)^4 = 0{\cdot}0677$
		1·0000

the five types be close to the values shown in the last column? Rather close, perhaps, but it would not be surprising to find some slight but systematic discrepancies because the formal assumptions underlying our argument may not be strictly correct. For one thing, the probability of a male birth may vary slightly from family to family (as pointed out in §3.1). More importantly, families that start in an unbalanced way, with several births of the same sex, are more likely to be continued than those that are better balanced. The first two births in families that are continued to the third stage will then not be representative of all two-birth families. A similar bias may exist in the progression from three births to four. The extent of these biases would be expected to differ from one community to another, and this seems to be borne out by actual data, some of which agree more closely than others with the theoretical probabilities.

It is sometimes convenient to express a probability in terms of the *odds*, which equal the probability that the event occurs divided by the probability that it does not occur. Thus, the odds of throwing a 6 with a dice are $\frac{1}{5}$.

3.3 **Bayes' theorem**

It was pointed out in §3.1 that the frequency definition of probability does not normally permit one to allot a numerical value to the probability that a certain proposition or hypothesis is true. We shall discuss in Chapter 6 the 'Bayesian' approach of assigning numerical values to probabilities of hypotheses, to represent degrees of belief. In the present section we introduce one of the basic tools of Bayesian statistics, Bayes' theorem, entirely within a frequentist framework.

There are some situations in which the relevant alternative hypotheses can be thought of as presenting themselves in a random sequence so that numerical probabilities can be associated with them. For instance, a doctor in charge of a clinic may be interested in the hypothesis: 'This patient has disease A.' By regarding the patient as a random member of a large collection of patients presenting themselves at the clinic the doctor may be able to associate with the hypothesis a certain probability, namely the long-run proportion of patients with disease A. This may be regarded as a *prior probability*, since it can be ascertained (or at least estimated roughly) from retrospective observations.

Suppose the doctor now makes certain new observations, after which the probability of the hypothesis: 'This patient has disease A' is again considered. The new value may be called a *posterior probability* because it refers to the situation after the new observations have been made. Intuitively one would expect the posterior probability to exceed the prior probability if the new observations were particularly common on the hypothesis in question and relatively uncommon on any alternative hypothesis. Conversely, the posterior probability would be expected to be less than the prior probability if the observations were not often observed in disease A but were common in other situations.

Consider a simple example in which there are only three possible diseases (A, B and C), with prior probabilities π_A, π_B and π_C (with $\pi_A + \pi_B + \pi_C = 1$). Suppose that the doctor's observations fall conveniently into one of four categories 1, 2, 3, 4, and that the probability distributions of the various outcomes for each disease are as follows:

	Outcome				
Disease	1	2	3	4	Total
A	$l_{1\|A}$	$l_{2\|A}$	$l_{3\|A}$	$l_{4\|A}$	1
B	$l_{1\|B}$	$l_{2\|B}$	$l_{3\|B}$	$l_{4\|B}$	1
C	$l_{1\|C}$	$l_{2\|C}$	$l_{3\|C}$	$l_{4\|C}$	1

Suppose the doctor observes outcome 2. The total probability of this outcome is

$$\pi_A l_{2|A} + \pi_B l_{2|B} + \pi_C l_{2|C}.$$

The three terms in this expression are in fact the probabilities of disease A and outcome 2, disease B and outcome 2, disease C and outcome 2. Once the doctor has observed outcome 2, therefore, the posterior probabilities of A, B and C, $\pi_{A|2}$, $\pi_{B|2}$ and $\pi_{C|2}$, are

$$\frac{\pi_A l_{2|A}}{\pi_A l_{2|A} + \pi_B l_{2|B} + \pi_C l_{2|C}}, \frac{\pi_B l_{2|B}}{\pi_A l_{2|A} + \pi_B l_{2|B} + \pi_C l_{2|C}}, \frac{\pi_C l_{2|C}}{\pi_A l_{2|A} + \pi_B l_{2|B} + \pi_C l_{2|C}}.$$

The prior probabilities have been multiplied by factors proportional to $l_{2|A}$, $l_{2|B}$ and $l_{2|C}$. Although these three quantities are straightforward probabilities, they do not form part of the same distribution, being entries in a column rather than a row of the table above. Probabilities of a particular outcome on different hypotheses are called *likelihoods* of these hypotheses.

This is an example of the use of Bayes' theorem (named after an English clergyman, Thomas Bayes, *c*. 1701–61). More generally, if the hypothesis H_i has a prior probability π_i and outcome y has a probability $l_{y|i}$ when H_i is true, the posterior probability of H_i after outcome y has been observed is

$$\pi_{i|y} = \frac{\pi_i l_{y|i}}{\sum_h \pi_h l_{y|h}}. \tag{3.3}$$

An alternative form of this equation concerns the ratio of probabilities of two hypotheses, H_i and H_j:

$$\frac{\pi_{i|y}}{\pi_{j|y}} = \frac{\pi_i}{\pi_j} \times \frac{l_{y|i}}{l_{y|j}}. \tag{3.4}$$

In (3.4), the left-hand side is the ratio of posterior probabilities, and the terms on the right-hand side are the ratio of prior probabilities and the likelihood ratio for the outcome y. In other words, a ratio of prior probabilities is

converted to a ratio of posterior probabilities by multiplication by the likelihood ratio.

A third form of Bayes' theorem concerns the odds in favour of hypothesis H_i:

$$\frac{\pi_{i|y}}{1-\pi_{i|y}} = \frac{\pi_i}{1-\pi_i} \times \frac{l_{y|i}}{\mathrm{P}(y\,|\,\text{not } H_i)}. \tag{3.5}$$

In (3.5), the posterior odds are derived from the prior odds by multiplication by a ratio called the *Bayes factor*. The denominator $\mathrm{P}(y\,|\text{not } H_i)$ is not a pure likelihood, since it involves the prior probabilities for hypotheses other than H_i.

In some examples, the outcomes will be continuous variables such as weight or blood pressure, in which case the likelihoods take the form of *probability densities*, to be described in the next section. The hypotheses H_i may form a continuous set (for example, H_i may specify that the mean of a population is some specific quantity that can take any value over a wide range, so that there is an infinite number of hypotheses). In that case the summation in the denominator of (3.3) must be replaced by an integral. But Bayes' theorem always takes the same basic form: prior probabilities are converted to posterior probabilities by multiplication in proportion to likelihoods.

The example provides an indication of the way in which Bayes' theorem may be used as an aid to diagnosis. In practice there are severe problems in estimating the probabilities appropriate for the population of patients under treatment; for example, the distribution of diseases observed in a particular centre is likely to vary with time. The determination of the likelihoods will involve extensive and carefully planned surveys and the definition of the outcome categories may be difficult.

One of the earliest applications of Bayes' theorem to medical diagnosis was that of Warner *et al.* (1961). They examined data from a large number of patients with congenital heart disease. For each of 33 different diagnoses they estimated the prior probability, π_i, and the probabilities $l_{y|i}$ of various combinations of symptoms. Altogether 50 symptoms, signs and other variables were measured on each individual. Even if all these had been dichotomies there would have been 2^{50} possible values of y, and it would clearly be impossible to get reliable estimates of all the $l_{y|i}$. Warner *et al.* overcame this problem by making an assumption which has often been made by later workers in this field, namely that the symptoms and other variables are statistically independent. The probability of any particular combination of symptoms, y, can then be obtained by multiplying together the separate, or *marginal*, probabilities of each. In this study firm diagnoses for certain patients could be made by intensive investigation and these were compared with the diagnoses given by Bayes' theorem and also with those made by experienced cardiologists using the same information. Bayes' theorem seems to emerge well from the comparison. Nevertheless, the assumption of independence

of symptoms is potentially dangerous and should not be made without careful thought.

A number of papers illustrating the use of Bayesian methods and related techniques of decision theory to problems in medical diagnosis, prognosis and decision-making are to be found in the journal *Medical Decision Making*. For a review of decision analysis see Glasziou and Schwartz (1991). The aim here is to provide rules for the choice among decisions to be made during the course of medical treatment. These may involve such questions as whether to proceed immediately with an operation or whether to delay the decision while the patient is kept under observation or until the results of laboratory tests become available. Such choices depend not only on assessments of the probabilities of life-threatening diseases, based on the evidence currently available, but also on assessments of the expected gain (or *utility*) to be derived from various outcomes. Bayesian methods are central to the calculation of probabilities, but the assignment of utilities may be difficult and indeed highly subjective.

For further discussion see Bailey (1977, §4.7), Weinstein and Fineberg (1980), Spiegelhalter and Knill-Jones (1984) and Pauker and Kassirer (1992).

Example 3.1

Fraser and Franklin (1974) studied 700 case records of patients with liver disease. The initial set of over 300 symptoms, signs and test results was reduced to 97 variables for purposes of analysis. In particular, groups of symptoms and signs recognized to be clinically interdependent were amalgamated so as to minimize the danger inherent in the independence assumption discussed above. Patients with rare diagnoses (those represented by less than 12 patients) or multiple diagnoses were omitted, as were those with incomplete records. There remained 480 cases. Prior probabilities were estimated from the relative frequencies of the various diagnoses, and likelihoods for particular combinations of symptoms, signs and test results were obtained by multiplication of marginal frequencies.

Application of Bayes' theorem to each of the patient records gave posterior probabilities. For example, the posterior probabilities for one patient were:

Acute infective hepatitis	0·952
Cholestatic infective hepatitis	0·048
All other diagnoses	0·000

In a large number of cases the first or first two diagnoses accounted for a high probability, as in this instance.

This first analysis led to some revisions of the case records, and the exercise was repeated on a reduced set of 419 patients and the predicted diagnoses compared with the true diagnoses. A result was categorized as 'equivocal' if the highest posterior probability was less than three times as great as the second highest. Similar predictions were done on the basis of the likelihoods alone (i.e. ignoring the prior probabilities). The results were:

	Correct	Equivocal	Incorrect
Likelihood	316 (75%)	51 (12%)	52 (12%)
Bayesian	325 (78%)	48 (11%)	46 (11%)

It seems likely that the Bayesian results would have shown a greater improvement over the likelihood results if the rare diseases had not been excluded (because the priors would then have been relatively more important).

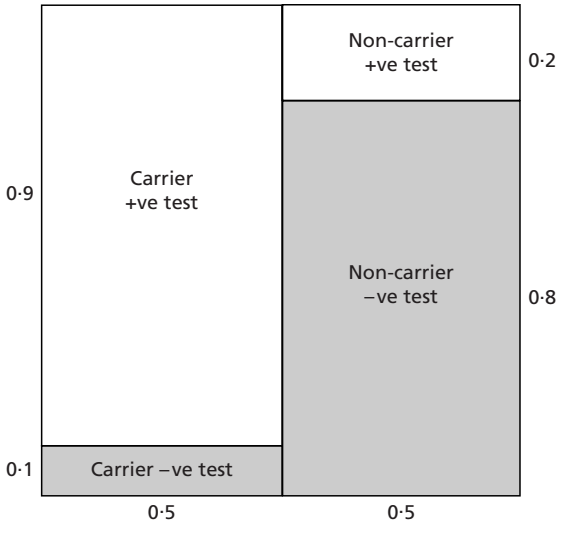
There is a danger in validating such a diagnostic procedure on the data set from which the method has been derived, because the estimates of probability are 'best' for this particular set and less appropriate for other sets. Fraser and Franklin therefore checked the method on 70 new cases with diagnoses falling into the group previously considered. The Bayesian method gave 44 (63%) correct, 12 (17%) equivocal and 14 (20%) incorrect results, with the likelihood method again slightly worse.

The following example illustrates the application of Bayes' theorem to genetic counselling.

Example 3.2

From genetic theory it is known that a woman with a haemophiliac brother has a probability of $\frac{1}{2}$ of being a carrier of a haemophiliac gene. A recombinant DNA diagnostic probe test provides information that contributes towards discriminating between carriers and non-carriers of a haemophiliac gene. It has been observed that 90% of women known to be carriers give a positive test result, whilst 20% of non-carriers have a positive result. The woman has the test. What is the probability that she is a carrier if the result is negative?

The probability may be evaluated using (3.3). Let H_1 be the hypothesis that the woman is a carrier and H_2 that she is not. Then $\pi_1 = \pi_2 = 0{\cdot}5$. If outcome y represents a negative test then $l_{y|1}$, the probability of a negative test result if the woman is a carrier,



equals $1 - 0{\cdot}9 = 0{\cdot}1$, and $l_{y|2}$, the probability of a negative test result if the woman is not a carrier, equals 0·8. Therefore,

$$\begin{aligned} \text{P(carrier given } -\text{ve test)} &= \frac{0{\cdot}5 \times 0{\cdot}1}{0{\cdot}5 \times 0{\cdot}1 + 0{\cdot}5 \times 0{\cdot}8} \\ &= 0{\cdot}11. \end{aligned}$$

Thus, the posterior probability is less than the prior probability, as would be expected, since the extra information, a negative test, is more probable if the woman is not a carrier.

A diagrammatic representation of the calculation is as follows. A square with sides of length 1 unit is divided vertically according to the prior probabilities of each hypothesis, and then horizontally within each column according to the probabilities of outcome. Then by the multiplication rule (3.2) each of the four rectangles has an area equal to the probability of the corresponding hypothesis and outcome.

Once it is known that the outcome is negative, then only the shaded area is relevant and the probability that the woman is a carrier is the proportion of the shaded area that is in the carrier column, leading to the same numerical answer as direct application of (3.3).

3.4 **Probability distributions**

Table 3.1 provides our first example of a *probability distribution*. That is, it shows how the total probability, equal to 1, is distributed among the different types of family. A variable whose different values follow a probability distribution is known as a *random variable*. In Table 3.1 the number of boys in a family is a random variable. So is the number of girls.

If a random variable can be associated with different points on a scale, the probability distribution can be represented visually by a histogram, just as for frequency distributions. We shall consider first some examples of ungrouped discrete random variables. Here the vertical scale of the histogram measures the probability for each value of the random variable, and each probability is represented by a vertical line.

Example 3.3

In repeated tosses of an unbiased coin, the outcome is a random variable with two values, H and T. Each value has a probability $\frac{1}{2}$. The distribution is shown in Fig. 3.2, where the two outcomes, H and T, are allotted to arbitrary points on the horizontal axis.

Example 3.4

In a genetic experiment we may cross two heterozygotes with genotypes Aa (that is, at a particular gene locus, each parent has one gene of type A and one of type a). The progeny will be homozygotes (aa or AA) or heterozygotes (Aa), with the probabilities shown below.

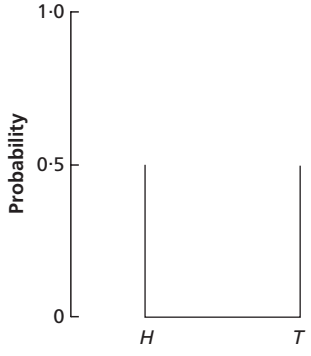


Fig. 3.2 Probability distribution for random variable with two values: the results of tossing a coin with equal probabilities of heads and tails.

Genotype	No. of *A* genes	Probability
aa	0	$\frac{1}{4}$
Aa	1	$\frac{1}{2}$
AA	2	$\frac{1}{4}$
		1

The three genotypes may be allotted to points on a scale by using as a random variable the number of *A* genes in the genotype. This random variable takes the values 0, 1 and 2, and the probability distribution is depicted in Fig. 3.3.

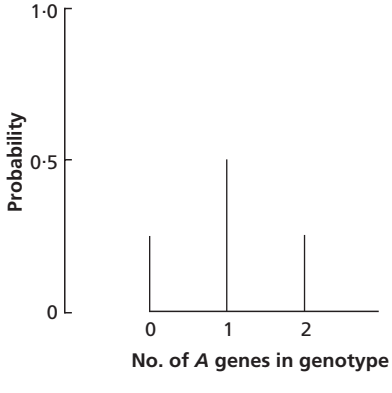


Fig. 3.3 Probability distribution for random variable with three values: the number of *A* genes in the genotype of progeny of an $Aa \times Aa$ cross.

Example 3.5

A third example is provided by the characterization of families of four children by the number of boys. The probabilities, in the particular numerical case considered in §3.2, are given in Table 3.1, and they are depicted in Fig. 3.4.

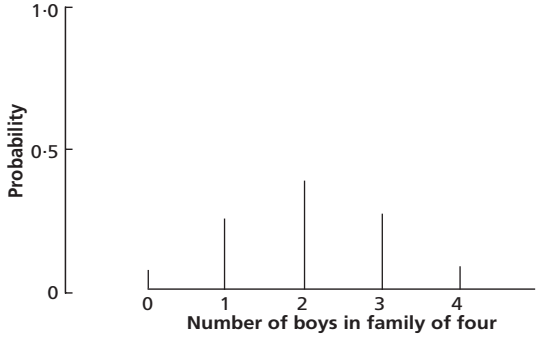


Fig. 3.4 Probability distribution of number of boys in family of four children if male births occur independently with probability 0.51 (Table 3.1).

When the random variable is continuous, it is of little use to refer to the probabilities of particular values of the variable, because these probabilities are in general zero. For example, the probability that the exact height of a male adult is 70 in. is zero, because in the virtually infinite population of exact heights of adult males a negligible proportion will be exactly 70 in. If, however, we consider a small interval centred at 70 in., say $70 - h$ to $70 + h$, where h is very small, there will be a small non-zero probability associated with this interval. Furthermore, the probability will be very nearly proportional to h. Thus, the probability of a height between 69·98 and 70·02 in. will be very nearly double the probability for the interval 69·99 to 70·01. It is therefore a reasonable representation of the situation to suppose that there is a *probability density* characteristic of the value 70 in., which can be denoted by $f(70)$, such that the probability for a small interval $70 - h$ to $70 + h$ is very close to

$$2hf(70).$$

The probability distribution for a continuous random variable, x, can therefore be depicted by a graph of the probability density $f(x)$ against x, as in Fig. 3.5. This is, in fact, the frequency curve discussed in §2.3. The reader familiar with the calculus will recognize $f(x)$ as the derivative with respect to x of the probability $F(x)$ that the random variable assumes a value less than or equal to x. $F(x)$ is called the *distribution function* and is represented by the area underneath the curve in Fig. 3.5 from the left end of the distribution (which may be at minus infinity) up to the value x. The distribution function corresponding to the density function of Fig. 3.5 is shown in Fig. 3.6. Note that the height of the density function is proportional to the slope of the distribution function; in the present example both these quantities are zero at the lower and upper extremes of the variables and attain a maximum at an intermediate point.

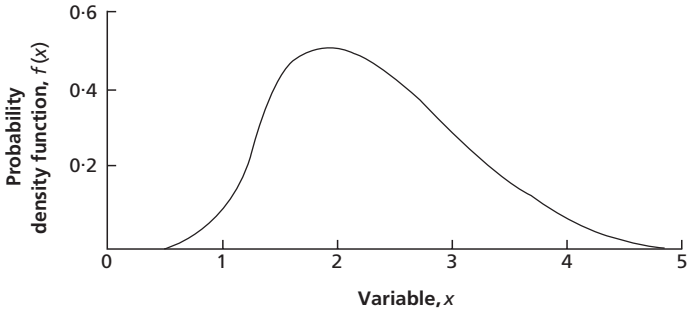


Fig. 3.5 Probability density function for a continuous random variable.

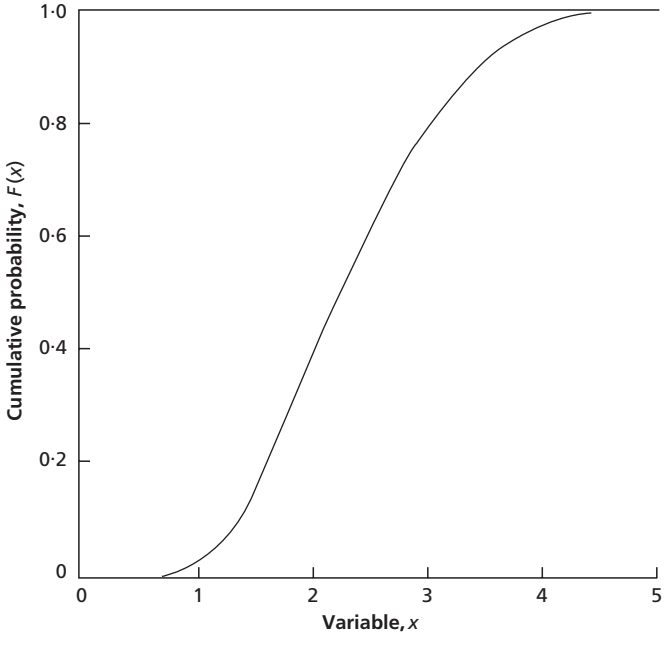


Fig. 3.6 Distribution function corresponding to the density function shown in Fig. 3.5.

The shape of a probability distribution may be characterized by the features already used for frequency distributions. In particular, we may be concerned with the number and position of the modes, the values of the random variable at which the probability, or (for continuous variables) the probability density, reaches a maximum. We may be interested too in the skewness of a probability distribution.

By analogy with §§2.4 and 2.6 we are particularly interested in the mean and standard deviation of a random variable, and these concepts are discussed in the next section.

3.5 **Expectation**

There is some difficulty in deciding what is meant by the phrase 'mean of a random variable'. The mean has been defined earlier only for a finite number, n, of observations. With a probability distribution such as that in Table 3.1 the number of observations must be thought of as infinite. How, then, is the mean to be calculated?

Suppose n is very large—so large that the relative frequencies of the different values of a discrete random variable like that in Table 3.1 can be taken to be very nearly equal to the probabilities. If they were exactly equal to the probabilities, the frequency distribution of the number of boys would be as follows:

x No. of boys	Frequency
0	$0{\cdot}0576n$
1	$0{\cdot}2400n$
2	$0{\cdot}3747n$
3	$0{\cdot}2600n$
4	$0{\cdot}0677n$
	n

The mean value of x would be

$$\frac{(0 \times 0{\cdot}0576n) + (1 \times 0{\cdot}2400n) + (2 \times 0{\cdot}3747n) + (3 \times 0{\cdot}2600n) + (4 \times 0{\cdot}0677n)}{n}.$$

The factor n may be cancelled from the numerator and denominator of the expression, to give the numerical result 2·04. The arbitrary sample size, n, does not appear.

If the probabilities of $0, 1, \ldots, 4$ boys are denoted by $P_0, P_1, \ldots, P_4$, the formula for the mean is clearly

$$(0 \times P_0) + (1 \times P_1) + (2 \times P_2) + (3 \times P_3) + (4 \times P_4).$$

In general, if x is a discrete random variable taking values $x_0, x_1, x_2, \ldots$, with probabilities $P_0, P_1, P_2, \ldots$, the mean value of x is calculated as

$$\sum_i x_i P_i. \tag{3.6}$$

The mean value of a random variable, calculated in this way, is often called the *expected value*, the *mathematical expectation* or simply the *expectation* of x, and the operation involved (multiplying each value of x by its probability, and then adding) is denoted by $\mathrm{E}(x)$. The expectation of a random variable is often allotted a Greek symbol like μ (lower-case Greek letter 'mu'), to distinguish it from a mean value calculated from a finite number of observations (denoted usually by symbols such as $\bar{x}$ or $\bar{y}$).

As a second example, consider the probability distribution of x, the number of A genes in the genotypes shown in Example 3.4. Here,

$$\begin{aligned} \mathrm{E}(x) &= (0 \times 0{\cdot}25) + (1 \times 0{\cdot}50) + (2 \times 0{\cdot}25) \\ &= 1. \end{aligned}$$

If x follows a continuous distribution, the formula given above for $\mathrm{E}(x)$ cannot be used. However, one could consider a discrete distribution in which possible values of x differed by a small interval $2h$. To any value X_0 we could allot the probability given by the continuous distribution for values of x between $X_0 - h$ and $X_0 + h$ (which, as we have seen, will be close to $2hf(X_0)$ if h is small enough). The expectation of x in this discrete distribution can be calculated by the general rule. As the interval h gets smaller and smaller, the discrete distribution will approach more and more closely the continuous distribution, and in general the expectation will approach a quantity which formally is given by the expression

$$\mu = \int_{-\infty}^{\infty} xf(x)\ \mathrm{d}x.$$

This provides a definition of the expectation of x for a continuous distribution.

The *variance* of a random variable is defined as

$$\mathrm{var}(x) = \mathrm{E}(x - \mu)^2;$$

that is, as the expectation of the squared difference from the mean. This is an obvious development from the previous formula $\sum(x_i - \bar{x})^2/n$ for the variance of a finite number, n, of observations, since this quantity is the mean value of the squared difference from the sample mean, $\bar{x}$. The distinction between the divisor of n and that of $n - 1$ becomes of no importance when we are dealing with probability distributions since n is effectively infinite.

The variance of a random variable is customarily given the symbol σ^2 (σ being the lower-case Greek letter 'sigma'). The standard deviation is again defined as σ, the square root of the variance.

By analogy with the short-cut formula (2.3) for the sample variance, we shall find it convenient to use the following relationship:

$$\sigma^2 = \mathrm{E}(x^2) - \mu^2. \tag{3.7}$$

The two formulae for the variance may be illustrated by the distribution in Example 3.4, for which we have already obtained $\mu = 1$. With the direct formula, we proceed as follows:

x	P	$x - \mu$	$(x - \mu)^2$
0	0·25	−1	1
1	0·50	0	0
2	0·25	1	1

$$\begin{aligned}\sigma^2 &= \mathrm{E}(x-\mu)^2 \\ &= (1 \times 0{\cdot}25) + (0 \times 0{\cdot}50) + (1 \times 0{\cdot}25) \\ &= 0{\cdot}5.\end{aligned}$$

With the short-cut formula,

x	P	x^2
0	0·25	0
1	0·50	1
2	0·25	4

$$\begin{aligned}\mathrm{E}(x^2) &= (0 \times 0{\cdot}25) + (1 \times 0{\cdot}50) + (4 \times 0{\cdot}25) \\ &= 1{\cdot}5; \\ \sigma^2 &= \mathrm{E}(x^2) - \mu^2 \\ &= 1{\cdot}5 - 1^2 \\ &= 0{\cdot}5,\end{aligned}$$

as before.

We have so far discussed probability distributions in rather general terms. In the next three sections we consider three specific forms of distribution which play a very important role in statistical theory and practice.

3.6 **The binomial distribution**

We have already met a particular case of this form of distribution in the example of §3.2 on the sex distribution in families of four.

In general, suppose we have a random sequence in which the outcome of each individual trial is of one of two types, A or B, these outcomes occurring with probabilities π and $1-\pi$, respectively. (The symbol π, the lower-case Greek letter 'pi', is used merely as a convenient Greek letter and has no connection at all with the mathematical constant $\pi = 3{\cdot}14159\ldots$) In the previous example, A and B were boys and girls, and π was 0·51.

Consider now a group of n observations from this random sequence (in the example $n = 4$). It will be convenient to refer to each such group as a 'sample' of n observations. What is the probability distribution of the number of As in the sample? This number we shall call r, and clearly r must be one of the numbers $0, 1, 2, \ldots, n-1, n$. Define also $p = r/n$, the *proportion* of As in the sample, and $q = (n-r)/n = 1-p$, the proportion of Bs.

As in the example, we argue that the probability of r As and $n-r$ Bs is

$$\pi^r(1-\pi)^{n-r}$$

multiplied by the number of ways in which one can choose r out of the n sample members to receive a lable 'A'. This multiplying factor is called a *binomial*

coefficient. In the example the binomial coefficients were worked out by simple enumeration, but clearly this could be tedious with large values of n and r. The binomial coefficient is usually denoted by

$$\binom{n}{r}$$

(referred to in speaking as 'n binomial r'), or

$${}^{n}\mathrm{C}_{r}.$$

Tables of binomial coefficients are provided in most books of mathematical tables. For moderate values of n and r they can be calculated directly from:

$$\binom{n}{r} = \frac{n(n-1)(n-2)\ldots(n-r+1)}{1\cdot 2\cdot 3\ldots r} \tag{3.8}$$

(where the single dots are multiplication signs and the rows of dots mean that all the intervening integers are used). The quantity $1\cdot 2\cdot 3\ldots r$ is called 'factorial r' or 'r factorial' and is usually written $r!$. Since the expression $n(n-1)\ldots(n-r+1)$, which occurs in the numerator of

$$\binom{n}{r},$$

can be written as

$$\frac{n!}{(n-r)!},$$

it follows that

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{3.9}$$

This formula involves unnecessarily heavy multiplication, but it draws attention to the symmetry of the binomial coefficients:

$$\binom{n}{r} = \binom{n}{n-r}. \tag{3.10}$$

This is, indeed, obvious from the definition. Any selection of r objects out of n is automatically a selection of the $n-r$ objects which remain.

If we put $r = 0$ in (3.8), both the numerator and the denominator are meaningless. Putting $r = n$ would give

$$\binom{n}{n} = \frac{n!}{n!} = 1,$$

and it would accord with the symmetry result to put

$$\binom{n}{0} = 1. \tag{3.11}$$

This is clearly the correct result, since there is precisely one way of selecting 0 objects out of n to be labelled as As: namely to select all the n objects to be labelled as Bs. Note that (3.11) accords with (3.9) if we agree to call $0! = 1$; this is merely a convention since 0! is strictly not covered by our previous definition of the factorial, but it provides a useful extension of the definition which is used generally in mathematics.

The binomial coefficients required in the example of §3.2 could have been obtained from (3.8) as follows:

$$\begin{aligned}
\binom{4}{0} &= 1 \\
\binom{4}{1} &= \frac{4}{1} = 4 \\
\binom{4}{2} &= \frac{4\cdot 3}{1\cdot 2} = 6 \\
\binom{4}{3} &= \frac{4\cdot 3\cdot 2}{1\cdot 2\cdot 3} = 4 \\
\binom{4}{4} &= \frac{4\cdot 3\cdot 2\cdot 1}{1\cdot 2\cdot 3\cdot 4} = 1.
\end{aligned}$$

A useful way to obtain binomial coefficients for small values of n, without any multiplication, is by means of Pascal's triangle:

n				1			
1			1		1		
2		1		2		1	
3	1		3		3		1
4		1		4		6		4		1
5	1		5		10		10		5		1
etc.				etc.			

In this triangle of numbers, which can be extended downwards indefinitely, each entry is obtained as the sum of the two adjacent numbers on the line above. Thus, in the fifth row (for $n = 4$),

$$4 = 1 + 3, \quad 6 = 3 + 3, \quad \text{etc.}$$

Along each row are the binomial coefficients

$$\binom{n}{0}, \binom{n}{1}, \ldots \quad \text{up to} \binom{n}{n-1}, \binom{n}{n}.$$

The probability that the sample of n individuals contains r As and $n - r$ Bs, then, is

$$\binom{n}{r}\pi^r(1-\pi)^{n-r}. \tag{3.12}$$

If this expression is evaluated for each value of r from 0 to n, the sum of these $n+1$ values will represent the probability of obtaining 0 As or 1 A or 2 As, etc., up to n As. These are the only possible results from the whole sequence and they are mutually exclusive; the sum of the probabilities is, therefore, 1. That this is so follows algebraically from the classical binomial theorem, for

$$\begin{aligned}\binom{n}{0}\pi^0(1-\pi)^n + \binom{n}{1}\pi^1(1-\pi)^{n-1} + \ldots + \binom{n}{n}\pi^n(1-\pi)^0 &= [\pi + (1-\pi)]^n \\ &= 1^n \\ &= 1.\end{aligned}$$

This result was verified in the particular example of Table 3.1.

The expectation and variance of r can now be obtained by applying the general formulae (3.6) and (3.7) to the probability distribution (3.12) and using standard algebraic results on the summation of series. Alternative derivations are given in §4.4. The results are:

$$\mathrm{E}(r) = n\pi \tag{3.13}$$

and

$$\mathrm{var}(r) = n\pi(1-\pi). \tag{3.14}$$

The formula for the expectation is intuitively acceptable. The mean number of As is equal to the number of observations multiplied by the probability that an individual result is an A. The expectation of the number of boys out of four, in our previous example, was shown in §3.5 to be 2·04. We now see that this result could have been obtained from (3.13):

$$\mathrm{E}(r) = 4 \times 0{\cdot}51 = 2{\cdot}04.$$

The formula (3.14) for the variance is less obvious. For a given value of n, var(r) reaches a maximum value when $\pi = 1 - \pi = \frac{1}{2}$ (when $\mathrm{var}(r) = \frac{1}{4}n$), and falls off markedly as π approaches 0 or 1. If π is very small, the factor $1 - \pi$ in (3.14) is very close to 1, and var(r) becomes very close to $n\pi$, the value of E(r).

We shall often be interested in the probability distribution of p, the *proportion* of As in the sample. Now $p = r \times (1/n)$, and the multiplying factor $1/n$ is constant from one sample to another. It follows that

$$\mathrm{E}(p) = \mathrm{E}(r) \times (1/n) = \pi \tag{3.15}$$

and

$$\mathrm{var}(p) = \mathrm{var}(r) \times (1/n)^2 = \frac{\pi(1-\pi)}{n}. \tag{3.16}$$

The square in the multiplying factor for the variance arises because the units in which the variance is measured are the squares of the units of the random variable.

It will sometimes be convenient to refer to the standard deviations of r or of p. These are the square roots of the corresponding variances:

$$\mathrm{SD}(r) = \sqrt{[n\pi(1-\pi)]}$$

and

$$\mathrm{SD}(p) = \sqrt{\left[\frac{\pi(1-\pi)}{n}\right]}.$$

Some further properties of the binomial distribution are given in §4.4. Binomial probabilities can be evaluated in many statistical packages.

Example 3.6

Table 3.2 is given by Lancaster (1965) from data published by Roberts *et al.* (1939). These authors observed 551 crosses between rats, with one parent heterozygous for each of five factors and the other parent homozygous recessive for each. The distribution is that of the number of dominant genes out of five, for each offspring. The theoretical distribution is the binomial with $n = 5$ and $\pi = \frac{1}{2}$, and the 'expected' frequencies, obtained by multiplying the binomial probabilities by 551, are shown in the table. The agreement between observed and expected frequencies is satisfactory.

The binomial distribution is characterized by the mathematical variables π and n. Variables such as these which partly or wholly characterize a probability distribution are known as *parameters*. They are, of course, entirely distinct from

Table 3.2 Distribution of number of dominant genes at five loci, in crosses between parents heterozygous for each factor and those homozygous recessive for each (Lancaster, 1965).

Number of dominant genes	Number of offspring	
	Observed	Expected
0	17	17·2
1	81	86·1
2	152	172·2
3	180	172·2
4	104	86·1
5	17	17·2
	551	551·0

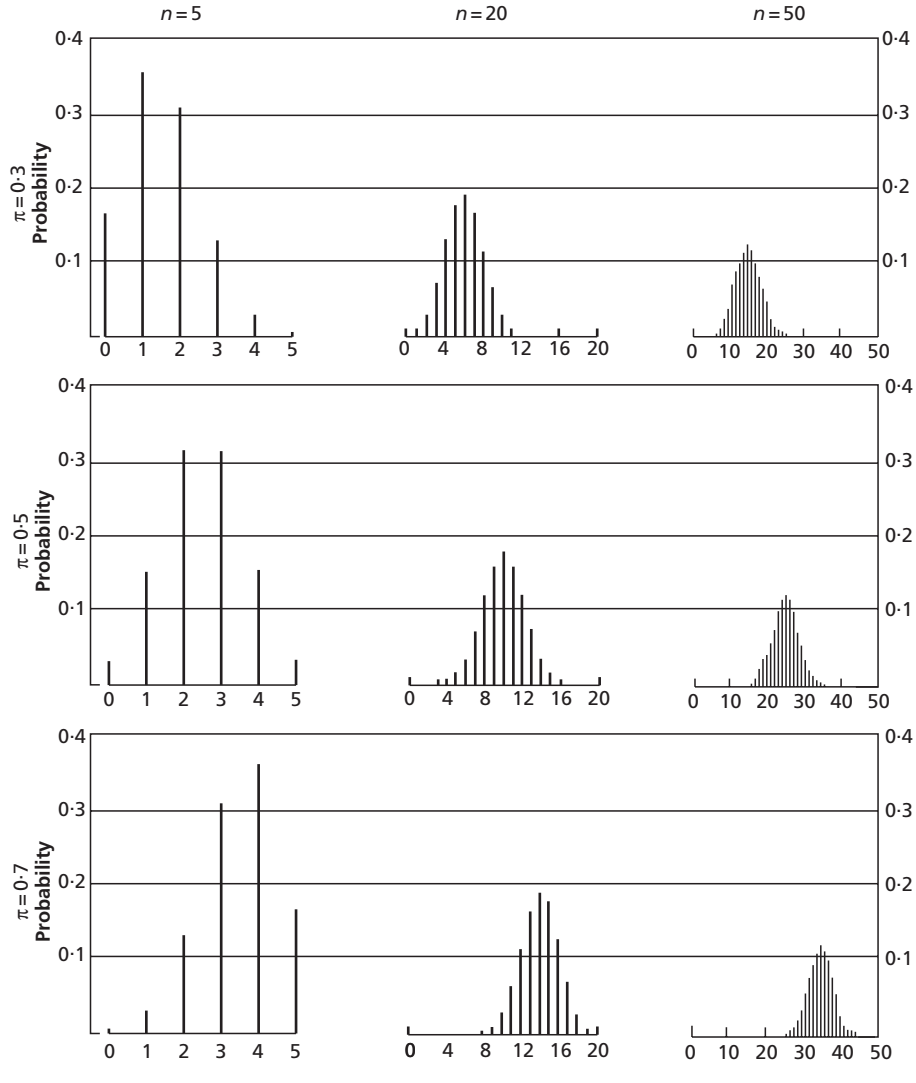


Fig. 3.7 Binomial distribution for various values of π and n. The horizontal scale in each diagram shows values of r.

random variables. Figure 3.7 illustrates the shape of the distribution for various combinations of π and n. Note that, for a particular value of n, the distribution is symmetrical for $\pi = \frac{1}{2}$ and asymmetrical for $\pi < \frac{1}{2}$ or $\pi > \frac{1}{2}$; and that for a particular value of π the asymmetry decreases as n increases.

Statistical methods based on the binomial distribution are described in detail in §§4.4, 4.5 and 8.5.