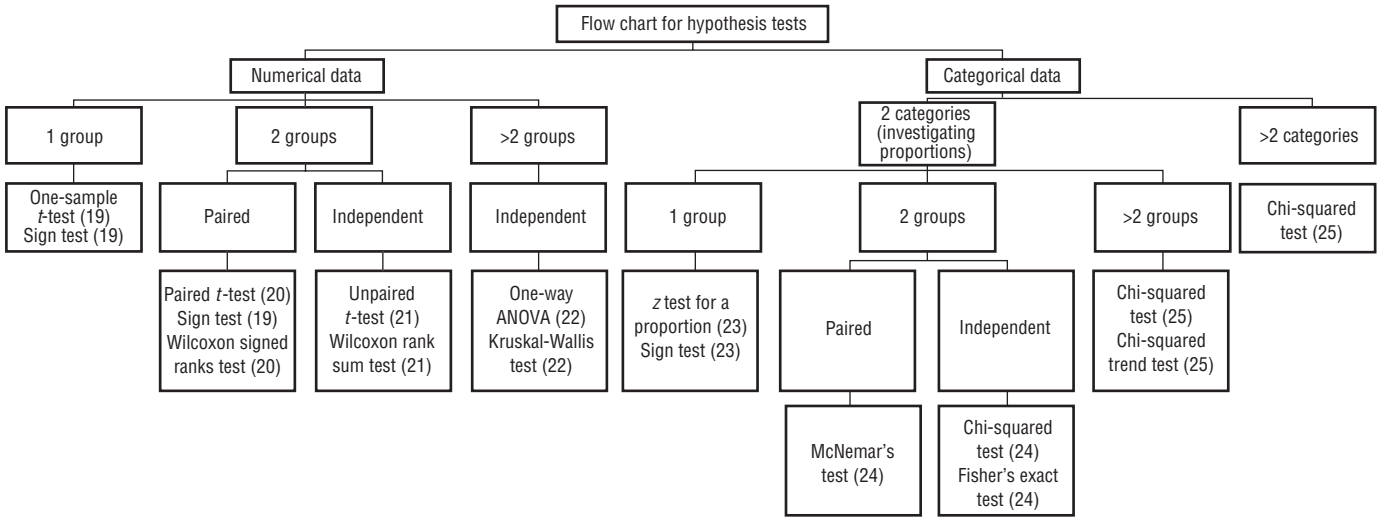
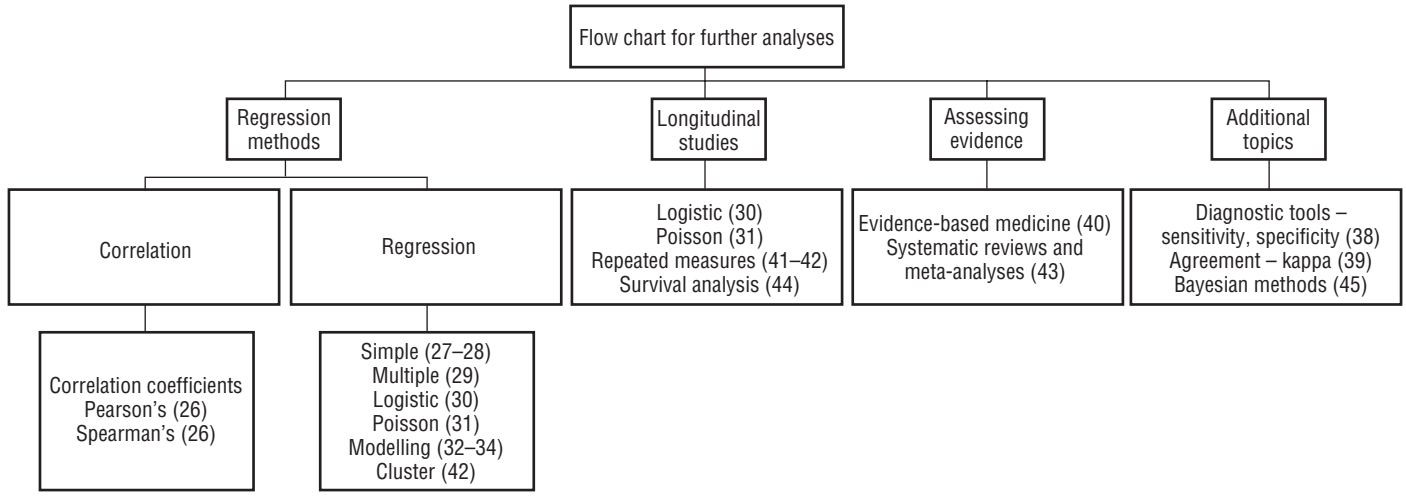
Flow charts indicating appropriate techniques in different circumstances*





*Relevant chapter numbers shown in parentheses.

Medical Statistics at a Glance

Medical Statistics at a Glance

Aviva Petrie

Head of Biostatistics Unit and Senior Lecturer
Eastman Dental Institute
University College London
256 Grays Inn Road
London WC1X 8LD *and*
Honorary Lecturer in Medical Statistics
Medical Statistics Unit
London School of Hygiene and Tropical Medicine
Keppel Street
London WC1E 7HT

Caroline Sabin

Professor of Medical Statistics and Epidemiology
Department of Primary Care and Population Sciences
Royal Free and University College Medical School
Rowland Hill Street
London NW3 2PF

Second edition

Blackwell Publishing

Published by Blackwell Publishing Ltd
Blackwell Publishing, Inc., 350 Main Street, Malden, Massachusetts 02148-5020, USA
Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK
Blackwell Publishing Asia Pty Ltd, 550 Swanston Street, Carlton, Victoria 3053, Australia

First published 2000
Reprinted 2001 (twice), 2002, 2003 (twice), 2004
Second edition 2005

Library of Congress Cataloging-in-Publication Data

Petrie, Aviva.
Medical statistics at a glance / Aviva Petrie, Caroline Sabin.—2nd ed.
p. ; cm.
Includes index.
ISBN-13: 978-1-4051-2780-6 (alk. paper)
ISBN-10: 1-4051-2780-5 (alk. paper)
1. Medical statistics.
[DNLM: 1. Statistics. 2. Research Design. WA 950 P495m 2005] I. Sabin, Caroline. II. Title.

R853.S7P476 2005
610′.72′—dc22

2004026022

ISBN-13: 978-1-4051-2780-6
ISBN-10: 1-4051-2780-5

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

Set in 9/11.5 pt Times by SNP Best-set Typesetter Ltd., Hong Kong
Printed and bound in India by Replika Press Pvt. Ltd.

Commissioning Editor: Martin Sugden
Development Editor: Karen Moore
Production Controller: Kate Charman

For further information on Blackwell Publishing, visit our website: http://www.blackwellpublishing.com

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy, and which has been manufactured from pulp processed using acid-free and elementary chlorine-free practices. Furthermore, the publisher ensures that the text paper and cover board used have met acceptable environmental accreditation standards.

Contents

Visit **www.medstatsaag.com** for further material including an extensive reference list and multiple choice questions (MCQs) with inter-active answers for self-assessment.

Preface

Medical Statistics at a Glance is directed at undergraduate medical students, medical researchers, postgraduates in the biomedical disciplines and at pharmaceutical industry personnel. All of these individuals will, at some time in their professional lives, be faced with quantitative results (their own or those of others) which will need to be critically evaluated and interpreted, and some, of course, will have to pass that dreaded statistics exam! A proper understanding of statistical concepts and methodology is invaluable for these needs. Much as we should like to fire the reader with an enthusiasm for the subject of statistics, we are pragmatic. Our aim in this new edition, as it was in the earlier edition, is to provide the student and the researcher, as well as the clinician encountering statistical concepts in the medical literature, with a book which is sound, easy to read, comprehensive, relevant, and of useful practical application.

We believe *Medical Statistics at a Glance* will be particularly helpful as an adjunct to statistics lectures and as a reference guide. The structure of this second edition is the same as that of the first edition. In line with other books in the *At a Glance* series, we lead the reader through a number of self-contained two-, three- or occasionally four-page chapters, each covering a different aspect of medical statistics. We have learned from our own teaching experiences, and have taken account of the difficulties that our students have encountered when studying medical statistics. For this reason, we have chosen to limit the theoretical content of the book to a level that is sufficient for understanding the procedures involved, yet which does not overshadow the practicalities of their execution.

Medical statistics is a wide-ranging subject covering a large number of topics. We have provided a basic introduction to the underlying concepts of medical statistics and a guide to the most commonly used statistical procedures. Epidemiology is closely allied to medical statistics. Hence some of the main issues in epidemiology, relating to study design and interpretation, are discussed. Also included are chapters which the reader may find useful only occasionally, but which are, nevertheless, fundamental to many areas of medical research; for example, evidence-based medicine, systematic reviews and meta-analysis, survival analysis and Bayesian methods. We have explained the principles underlying these topics so that the reader will be able to understand and interpret the results from them when they are presented in the literature.

The order of the first 30 chapters of this edition corresponds to that of the first edition. Most of these chapters remain unaltered in this new edition: some have relatively minor changes which accommodate recent advances, cross-referencing or re-organization of the new material. Our major amendments relate to comparatively complex forms of regression analysis which are now more widely used than at the time of writing the first edition, partly because the associated software is more accessible and efficient than in the past. We have modified the chapter on binary outcomes and logistic regression (Chapter 30), included a new chapter on rates and Poisson regression (Chapter 31) and have considerably expanded the original statistical modelling chapter so that it now comprises three chapters, entitled 'Generalized linear models (Chapter 32), 'Explanatory variables in statistical models' (Chapter 33) and 'Issues in statistical modelling' (Chapter 34). We have also modified Chapter 41 which describes different approaches to the analysis of clustered data, and added Chapter 42 which outlines the various regression methods that can be used to analyse this type of data. The first edition had a brief description of time series analysis which we decided to omit from this second edition as we felt that it was probably too limited to be of real use, and expanding it would go beyond the bounds of our remit. Because of this omission and the chapters that we have added, the numbering of the chapters in the second edition differs from that of the first edition after Chapter 30. Most of the chapters in this latter section of the book which were also in the first edition are altered only slightly, if at all.

The description of every statistical technique is accompanied by an example illustrating its use. We have generally obtained the data for these examples from collaborative studies in which we or colleagues have been involved; in some instances, we have used real data from published papers. Where possible, we have used the same data set in more than one chapter to reflect the reality of data analysis which is rarely restricted to a single technique or approach. Although we believe that formulae should be provided and the logic of the approach explained as an aid to understanding, we have avoided showing the details of complex calculations—most readers will have access to computers and are unlikely to perform any but the simplest calculations by hand.

We consider that it is particularly important for the reader to be able to interpret output from a computer package. We have therefore chosen, where applicable, to show results using extracts from computer output. In some instances, where we believe individuals may have difficulty with its interpretation, we have included (Appendix C) and annotated the complete computer output from an analysis of a data set. There are many statistical packages in common use; to give the reader an indication of how output can vary, we have not restricted the output to a particular package and have, instead, used three well known ones—SAS, SPSS and Stata.

There is extensive cross-referencing throughout the text to help the reader link the various procedures. A basic set of statistical tables is contained in Appendix A. Neave, H.R. (1981) *Elememen-tary Statistical Tables* Routledge, and Diem, K. (1970) *Documenta Geigy Scientific Tables*, 7th Edn, Blackwell Publishing: Oxford, amongst others, provide fuller versions if the reader requires more precise results for hand calculations. The Glossary of terms (Appendix D) provides readily accessible explanations of commonly used terminology.

We know that one of the greatest difficulties facing non-statisticians is choosing the appropriate technique. We have therefore produced two flow charts which can be used both to aid the decision as to what method to use in a given situation and to locate a particular technique in the book easily. These flow charts are displayed prominently on the inside cover for easy access.

The reader may find it helpful to assess his/her progress in self-directed learning by attempting the interactive exercises on our Website (www.medstatsaag.com). This Website also contains a full set of references (some of which are linked directly to Medline) to supplement the references quoted in the text and provide useful background information for the examples. For those readers

who wish to gain a greater insight into particular areas of medical statistics, we can recommend the following books:

Altman, D.G. (1991). *Practical Statistics for Medical Research*. Chapman and Hall, London.

Armitage, P., Berry, G. and Matthews, J.F.N. (2001). *Statistical Methods in Medical Research, 4th Edn*. Blackwell Science, Oxford.

Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. Wiley, Chichester.

We are extremely grateful to Mark Gilthorpe and Jonathan Sterne who made invaluable comments and suggestions on aspects of this second edition, and to Richard Morris, Fiona Lampe, Shak Hajat and Abul Basar for their counsel on the first edition. We wish to thank everyone who has helped us by providing data for the examples. Naturally, we take full responsibility for any errors that remain in the text or examples. We should also like to thank Mike, Gerald, Nina, Andrew and Karen who tolerated, with equanimity, our preoccupation with the first edition and lived with us through the trials and tribulations of this second edition.

Aviva Petrie
Caroline Sabin
London

1 Types of data

Data and statistics

The purpose of most studies is to collect **data** to obtain information about a particular area of research. Our data comprise **observations** on one or more variables; any quantity that varies is termed a **variable**. For example, we may collect basic clinical and demographic information on patients with a particular illness. The variables of interest may include the sex, age and height of the patients.

Our data are usually obtained from a **sample** of individuals which represents the **population** of interest. Our aim is to condense these data in a meaningful way and extract useful information from them. **Statistics** encompasses the methods of collecting, summarizing, analysing and drawing conclusions from the data: we use statistical techniques to achieve our aim.

Data may take many different forms. We need to know what form every variable takes before we can make a decision regarding the most appropriate statistical methods to use. Each variable and the resulting data will be one of two types: **categorical** or **numerical** (Fig. 1.1).

Categorical (qualitative) data

These occur when each individual can only belong to one of a number of distinct categories of the variable.

- **Nominal data**—the categories are not ordered but simply have names. Examples include blood group (A, B, AB, and O) and marital status (married/widowed/single etc.). In this case, there is no reason to suspect that being married is any better (or worse) than being single!
- **Ordinal data**—the categories are ordered in some way. Examples include disease staging systems (advanced, moderate, mild, none) and degree of pain (severe, moderate, mild, none).

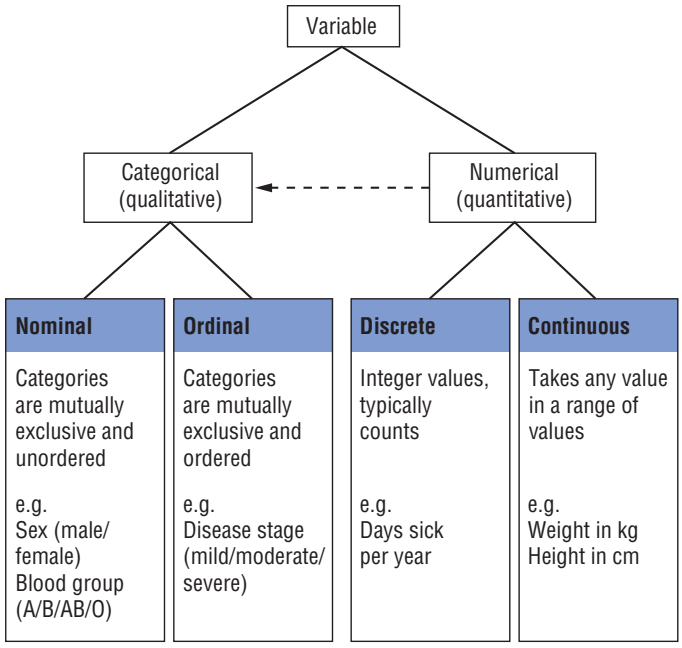


Figure 1.1 Diagram showing the different types of variable.

A categorical variable is **binary** or **dichotomous** when there are only two possible categories. Examples include 'Yes/No', 'Dead/Alive' or 'Patient has disease/Patient does not have disease'.

Numerical (quantitative) data

These occur when the variable takes some numerical value. We can subdivide numerical data into two types.

- **Discrete data**—occur when the variable can only take certain whole numerical values. These are often counts of numbers of events, such as the number of visits to a GP in a year or the number of episodes of illness in an individual over the last five years.
- **Continuous data**—occur when there is no limitation on the values that the variable can take, e.g. weight or height, other than that which restricts us when we make the measurement.

Distinguishing between data types

We often use very different statistical methods depending on whether the data are categorical or numerical. Although the distinction between categorical and numerical data is usually clear, in some situations it may become blurred. For example, when we have a variable with a large number of ordered categories (e.g. a pain scale with seven categories), it may be difficult to distinguish it from a discrete numerical variable. The distinction between discrete and continuous numerical data may be even less clear, although in general this will have little impact on the results of most analyses. Age is an example of a variable that is often treated as discrete even though it is truly continuous. We usually refer to 'age at last birthday' rather than 'age', and therefore, a woman who reports being 30 may have just had her 30th birthday, or may be just about to have her 31st birthday.

Do not be tempted to record numerical data as categorical at the outset (e.g. by recording only the range within which each patient's age falls rather than his/her actual age) as important information is often lost. It is simple to convert numerical data to categorical data once they have been collected.

Derived data

We may encounter a number of other types of data in the medical field. These include:

- **Percentages**—These may arise when considering improvements in patients following treatment, e.g. a patient's lung function (forced expiratory volume in 1 second, FEV1) may increase by 24% following treatment with a new drug. In this case, it is the level of improvement, rather than the absolute value, which is of interest.
- **Ratios** or **quotients**—Occasionally you may encounter the ratio or quotient of two variables. For example, body mass index (BMI), calculated as an individual's weight (kg) divided by her/his height squared (m^2), is often used to assess whether s/he is over- or under-weight.
- **Rates**—Disease rates, in which the number of disease events occurring among individuals in a study is divided by the total number of years of follow-up of all individuals in that

study (Chapter 31), are common in epidemiological studies (Chapter 12).

• **Scores**—We sometimes use an arbitrary value, i.e. a score, when we cannot measure a quantity. For example, a series of responses to questions on quality of life may be summed to give some overall quality of life score on each individual.

All these variables can be treated as numerical variables for most analyses. Where the variable is derived using more than one value (e.g. the numerator and denominator of a percentage), it is important to record all of the values used. For example, a 10% improvement in a marker following treatment may have different clinical relevance depending on the level of the marker before treatment.

Censored data

We may come across **censored** data in situations illustrated by the following examples.

• If we measure laboratory values using a tool that can only detect levels above a certain cut-off value, then any values below this cut-off will not be detected. For example, when measuring virus levels, those below the limit of detectability will often be reported as 'undetectable' even though there may be some virus in the sample.

• We may encounter censored data when following patients in a trial in which, for example, some patients withdraw from the trial before the trial has ended. This type of data is discussed in more detail in Chapter 44.

2 Data entry

When you carry out any study you will almost always need to enter the data onto a computer package. Computers are invaluable for improving the accuracy and speed of data collection and analysis, making it easy to check for errors, produce graphical summaries of the data and generate new variables. It is worth spending some time planning data entry—this may save considerable effort at later stages.

Formats for data entry

There are a number of ways in which data can be entered and stored on a computer. Most statistical packages allow you to enter data directly. However, the limitation of this approach is that often you cannot move the data to another package. A simple alternative is to store the data in either a spreadsheet or database package. Unfortunately, their statistical procedures are often limited, and it will usually be necessary to output the data into a specialist statistical package to carry out analyses.

A more flexible approach is to have your data available as an **ASCII** or **text** file. Once in an ASCII format, the data can be read by most packages. ASCII format simply consists of rows of text that you can view on a computer screen. Usually, each variable in the file is separated from the next by some **delimiter**, often a space or a comma. This is known as **free format**.

The simplest way of entering data in ASCII format is to type the data directly in this format using either a word processing or editing package. Alternatively, data stored in spreadsheet packages can be saved in ASCII format. Using either approach, it is customary for each row of data to correspond to a different individual in the study, and each column to correspond to a different variable, although it may be necessary to go on to subsequent rows if data from a large number of variables are collected on each individual.

Planning data entry

When collecting data in a study you will often need to use a form or questionnaire for recording the data. If these forms are designed carefully, they can reduce the amount of work that has to be done when entering the data. Generally, these forms/questionnaires include a series of boxes in which the data are recorded—it is usual to have a separate box for each possible digit of the response.

Categorical data

Some statistical packages have problems dealing with non-numerical data. Therefore, you may need to assign numerical codes to categorical data before entering the data onto the computer. For example, you may choose to assign the codes of 1, 2, 3 and 4 to categories of 'no pain', 'mild pain', 'moderate pain' and 'severe pain', respectively. These codes can be added to the forms when collecting the data. For binary data, e.g. yes/no answers, it is often convenient to assign the codes 1 (e.g. for 'yes') and 0 (for 'no').

- **Single-coded** variables—there is only one possible answer to a question, e.g. 'is the patient dead?'. It is not possible to answer both 'yes' and 'no' to this question.
- **Multi-coded** variables—more than one answer is possible for each respondent. For example, 'what symptoms has this patient experienced?'. In this case, an individual may have experienced any of a number of symptoms. There are two ways to deal with this type of data depending upon which of the two following situations applies.
  - **There are only a few possible symptoms, and individuals may have experienced many of them.** A number of different binary variables can be created which correspond to whether the patient has answered yes or no to the presence of each possible symptom. For example, 'did the patient have a cough?' 'Did the patient have a sore throat?'
  - **There are a very large number of possible symptoms but each patient is expected to suffer from only a few of them.** A number of different nominal variables can be created; each successive variable allows you to name a symptom suffered by the patient. For example, 'what was the first symptom the patient suffered?' 'What was the second symptom?' You will need to decide in advance the maximum number of symptoms you think a patient is likely to have suffered.

Numerical data

Numerical data should be entered with the same precision as they are measured, and the unit of measurement should be consistent for all observations on a variable. For example, weight should be recorded in kilograms or in pounds, but not both interchangeably.

Multiple forms per patient

Sometimes, information is collected on the same patient on more than one occasion. It is important that there is some unique identifier (e.g. a serial number) relating to the individual that will enable you to link all of the data from an individual in the study.

Problems with dates and times

Dates and times should be entered in a consistent manner, e.g. either as day/month/year or month/day/year, but not interchangeably. It is important to find out what format the statistical package can read.

Coding missing values

You should consider what you will do with missing values before you enter the data. In most cases you will need to use some symbol to represent a missing value. Statistical packages deal with missing values in different ways. Some use special characters (e.g. a full stop or asterisk) to indicate missing values, whereas others require you to define your own code for a missing value (commonly used values are 9, 999 or –99). The value that is chosen should be one that is not possible for that variable. For example, when entering a categorical variable with four categories (coded 1, 2, 3 and 4), you may choose the value 9 to represent missing values. However, if the variable is 'age of child' then a different code should be chosen. Missing data are discussed in more detail in Chapter 3.

Example

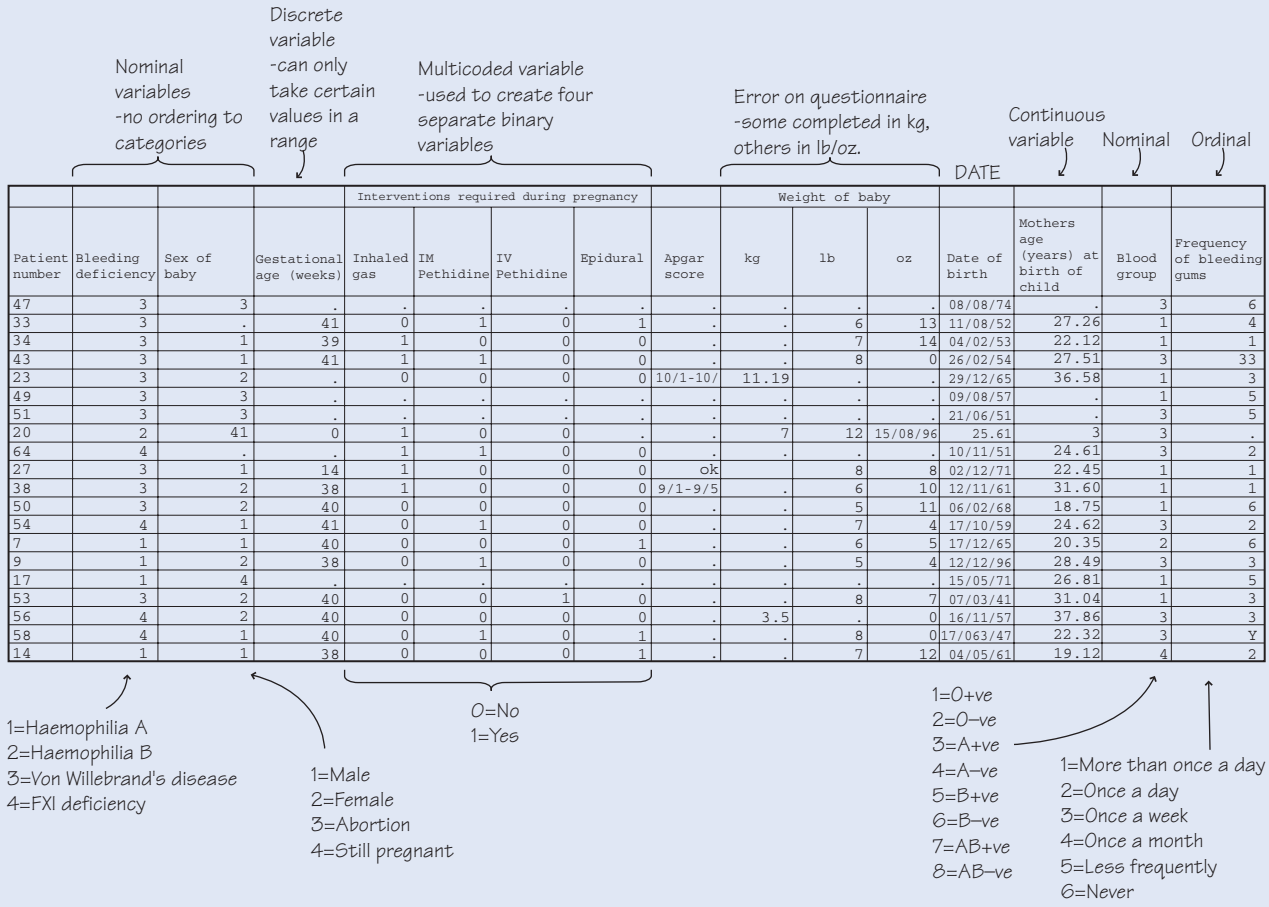


Figure 2.1 Portion of a spreadsheet showing data collected on a sample of 64 women with inherited bleeding disorders.

As part of a study on the effect of inherited bleeding disorders on pregnancy and childbirth, data were collected on a sample of 64 women registered at a single haemophilia centre in London. The women were asked questions relating to their bleeding disorder and their first pregnancy (or their current pregnancy if they were pregnant for the first time on the date of interview). Fig. 2.1 shows the data from a small selection of the women after the data have been entered onto a spreadsheet, but before they have been checked for errors. The coding schemes for the categorical variables are shown at the bottom of Fig. 2.1. Each row of the spreadsheet represents a separate individual in the study; each column represents a different variable. Where the woman is still pregnant, the age of the woman at the time of birth has been calculated from the estimated date of the baby's delivery. Data relating to the live births are shown in Chapter 37.

Data kindly provided by Dr R. A. Kadir, University Department of Obstetrics and Gynaecology, and Professor C. A. Lee, Haemophilia Centre and Haemostasis Unit, Royal Free Hospital, London.

3 Error checking and outliers

In any study there is always the potential for errors to occur in a data set, either at the outset when taking measurements, or when collecting, transcribing and entering the data onto a computer. It is hard to eliminate all of these errors. However, you can reduce the number of typing and transcribing errors by checking the data carefully once they have been entered. Simply scanning the data by eye will often identify values that are obviously wrong. In this chapter we suggest a number of other approaches that you can use when checking data.

Typing errors

Typing mistakes are the most frequent source of errors when entering data. If the amount of data is small, then you can check the typed data set against the original forms/questionnaires to see whether there are any typing mistakes. However, this is time-consuming if the amount of data is large. It is possible to type the data in twice and compare the two data sets using a computer program. Any differences between the two data sets will reveal typing mistakes. Although this approach does not rule out the possibility that the same error has been incorrectly entered on both occasions, or that the value on the form/questionnaire is incorrect, it does at least minimize the number of errors. The disadvantage of this method is that it takes twice as long to enter the data, which may have major cost or time implications.

Error checking

- **Categorical data**—It is relatively easy to check categorical data, as the responses for each variable can only take one of a number of limited values. Therefore, values that are not allowable must be errors.
- **Numerical data**—Numerical data are often difficult to check but are prone to errors. For example, it is simple to transpose digits or to misplace a decimal point when entering numerical data. Numerical data can be **range checked**—that is, upper and lower limits can be specified for each variable. If a value lies outside this range then it is flagged up for further investigation.
- **Dates**—It is often difficult to check the accuracy of dates, although sometimes you may know that dates must fall within certain time periods. Dates can be checked to make sure that they are valid. For example, 30th February must be incorrect, as must any day of the month greater than 31, and any month greater than 12. Certain logical checks can also be applied. For example, a patient's date of birth should correspond to his/her age, and patients should usually have been born before entering the study (at least in most studies). In addition, patients who have died should not appear for subsequent follow-up visits!

With all error checks, a value should only be corrected if there is evidence that a mistake has been made. You should not change values simply because they look unusual.

Handling missing data

There is always a chance that some data will be missing. If a very large proportion of the data is missing, then the results are unlikely to be reliable. The reasons why data are missing should always be investigated—if missing data tend to cluster on a particular variable and/or in a particular sub-group of individuals, then it may indicate that the variable is not applicable or has never been measured for that group of individuals. If this is the case, it may be necessary to exclude that variable or group of individuals from the analysis. We may encounter particular problems when the chance that data are missing is strongly related to the variable of greatest interest in our study (e.g. the outcome in a regression analysis—Chapter 27). In this situation, our results may be severely biased (Chapter 12). For example, suppose we are interested in a measurement which reflects the health status of patients and this information is missing for some patients because they were not well enough to attend their clinic appointments: we are likely to get an overly optimistic overall view of the patients' health if we take no account of the missing data in the analysis. It may be possible to reduce this bias by using appropriate statistical methods[1] or by estimating the missing data in some way[2], but a preferable option is to minimize the amount of missing data at the outset.

Outliers

What are outliers?

Outliers are observations that are distinct from the main body of the data, and are incompatible with the rest of the data. These values may be genuine observations from individuals with very extreme levels of the variable. However, they may also result from typing errors or the incorrect choice of units, and so any suspicious values should be checked. It is important to detect whether there are outliers in the data set, as they may have a considerable impact on the results from some types of analyses (Chapter 29).

For example, a woman who is 7 feet tall would probably appear as an outlier in most data sets. However, although this value is clearly very high, compared with the usual heights of women, it may be genuine and the woman may simply be very tall. In this case, you should investigate this value further, possibly checking other variables such as her age and weight, before making any decisions about the validity of the result. The value should only be changed if there really is evidence that it is incorrect.

Checking for outliers

A simple approach is to print the data and visually check them by eye. This is suitable if the number of observations is not too large and if the potential outlier is much lower or higher than the rest of the data. Range checking should also identify possible outliers. Alternatively, the data can be plotted in some way (Chapter 4)—outliers can be clearly identified on histograms and scatter plots (see also Chapter 29 for a discussion of outliers in regression analysis).

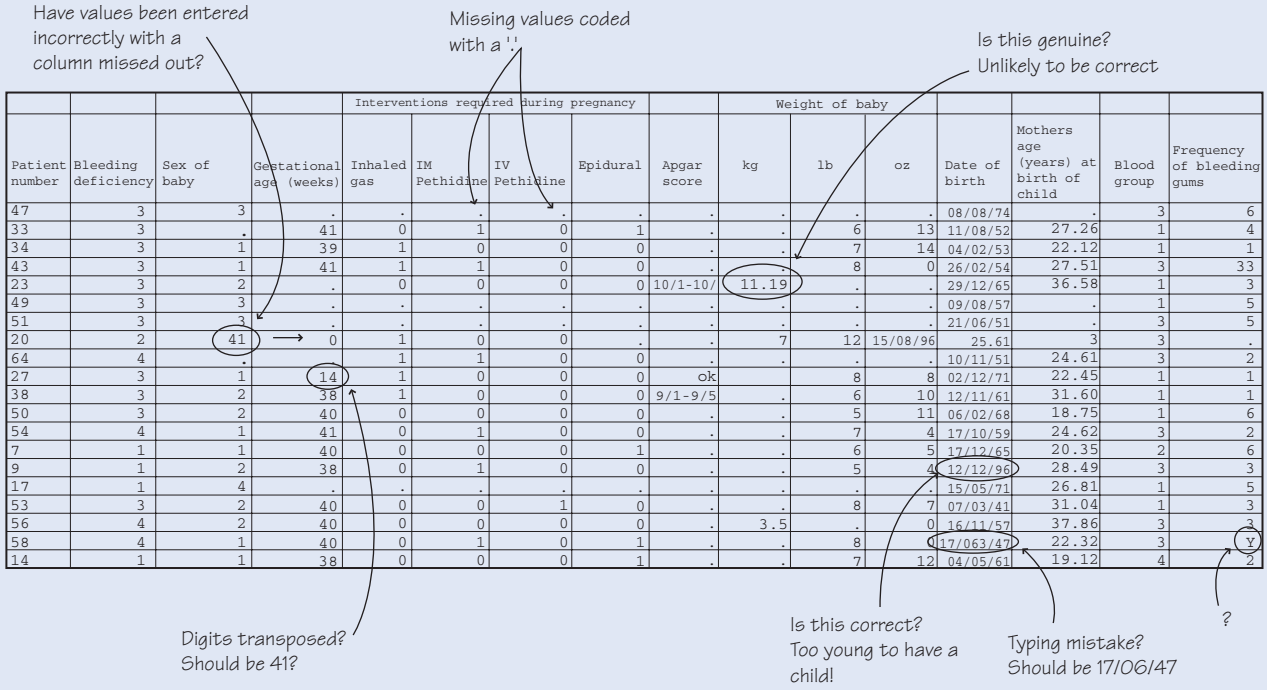
Handling outliers

It is important not to remove an individual from an analysis simply because his/her values are higher or lower than might be expected.

[1] Laird, N.M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, **7**, 305–315.
[2] Engels, J.M. and Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology*, **56**: 968–976.

However, the inclusion of outliers may affect the results when some statistical techniques are used. A simple approach is to repeat the analysis both including and excluding the value. If the results are similar, then the outlier does not have a great influence on the result. However, if the results change drastically, it is important to use appropriate methods that are not affected by outliers to analyse the data. These include the use of transformations (Chapter 9) and non-parametric tests (Chapter 17).

Example



				Interventions required during pregnancy					Weight of baby							
Patient number	Bleeding deficiency	Sex of baby	Gestational age (weeks)	Inhaled gas	IM Pethidine	IV Pethidine	Epidural	Apgar score	kg	lb	oz	Date of birth	Mothers age (years) at birth of child	Blood group	Frequency of bleeding gums
47	3	3	.	.	.	.	.	.	.	.	.	08/08/74	.	3	6
33	3	.	41	0	1	0	1	.	.	6	13	11/08/52	27.26	1	4
34	3	1	39	1	0	0	0	.	.	7	14	04/02/53	22.12	1	1
43	3	1	41	1	1	0	0	.		8	0	26/02/54	27.51	3	33
23	3	2	.	0	0	0	0	10/1-10/	11.19	.	.	29/12/65	36.58	1	3
49	3	3	.	.	.	.	.	.	.	.	.	09/08/57	.	1	5
51	3	3	.	.	.	.	.	.	.	.	.	21/06/51	.	3	5
20	2	41	0	1	0	0	.	.	7	12	15/08/96	25.61	3	3	.
64	4	.		1	1	0	0	.	.	.	.	10/11/51	24.61	3	2
27	3	1	14	1	0	0	0	ok		8	8	02/12/71	22.45	1	1
38	3	2	38	1	0	0	0	9/1-9/5	.	6	10	12/11/61	31.60	1	1
50	3	2	40	0	0	0	0	.	.	5	11	06/02/68	18.75	1	6
54	4	1	41	0	1	0	0	.	.	7	4	17/10/59	24.62	3	2
7	1	1	40	0	0	0	1	.	.	6	5	17/12/65	20.35	2	6
9	1	2	38	0	1	0	0	.	.	5	4	12/12/96	28.49	3	3
17	1	4	.	.	.	.	.	.	.	.	.	15/05/71	26.81	1	5
53	3	2	40	0	0	1	0	.	.	8	7	07/03/41	31.04	1	3
56	4	2	40	0	0	0	0	.	3.5	.	0	16/11/57	37.86	3	3
58	4	1	40	0	1	0	1	.	.	8	0	17/063/47	22.32	3	Y
14	1	1	38	0	0	0	1	.	.	7	12	04/05/61	19.12	4	2

Figure 3.1 Checking for errors in a data set.

After entering the data described in Chapter 2, the data set is checked for errors. Some of the inconsistencies highlighted are simple data entry errors. For example, the code of '41' in the 'sex of baby' column is incorrect as a result of the sex information being missing for patient 20; the rest of the data for patient 20 had been entered in the incorrect columns. Others (e.g. unusual values in the gestational age and weight columns) are likely to be errors, but the notes should be checked before any decision is made, as these may reflect genuine outliers. In this case, the gestational age of patient number 27 was 41 weeks, and it was decided that a weight of 11.19 kg was incorrect. As it was not possible to find the correct weight for this baby, the value was entered as missing.

4 Displaying data graphically

One of the first things that you may wish to do when you have entered your data onto a computer is to summarize them in some way so that you can get a 'feel' for the data. This can be done by producing diagrams, tables or summary statistics (Chapters 5 and 6). Diagrams are often powerful tools for conveying information about the data, for providing simple summary pictures, and for spotting outliers and trends before any formal analyses are performed.

One variable

Frequency distributions

An **empirical frequency distribution** of a variable relates each possible observation, class of observations (i.e. range of values) or category, as appropriate, to its observed **frequency** of occurrence. If we replace each frequency by a **relative frequency** (the percentage of the total frequency), we can compare frequency distributions in two or more groups of individuals.

Displaying frequency distributions

Once the frequencies (or relative frequencies) have been obtained for *categorical* or some *discrete numerical* data, these can be displayed visually.

- **Bar or column chart**—a separate horizontal or vertical bar is drawn for each category, its length being proportional to the frequency in that category. The bars are separated by small gaps to indicate that the data are categorical or discrete (Fig. 4.1a).
- **Pie chart**—a circular 'pie' is split into sections, one for each category, so that the area of each section is proportional to the frequency in that category (Fig. 4.1b).

It is often more difficult to display *continuous numerical* data, as the data may need to be summarized before being drawn. Commonly used diagrams include the following:

- **Histogram**—this is similar to a bar chart, but there should be no gaps between the bars as the data are continuous (Fig. 4.1d). The width of each bar of the histogram relates to a range of values for the variable. For example, the baby's weight (Fig. 4.1d) may be categorized into 1.75–1.99 kg, 2.00–2.24 kg, . . . , 4.25–4.49 kg. The area of the bar is proportional to the frequency in that range. Therefore, if one of the groups covers a wider range than the others, its base will be wider and height shorter to compensate. Usually, between

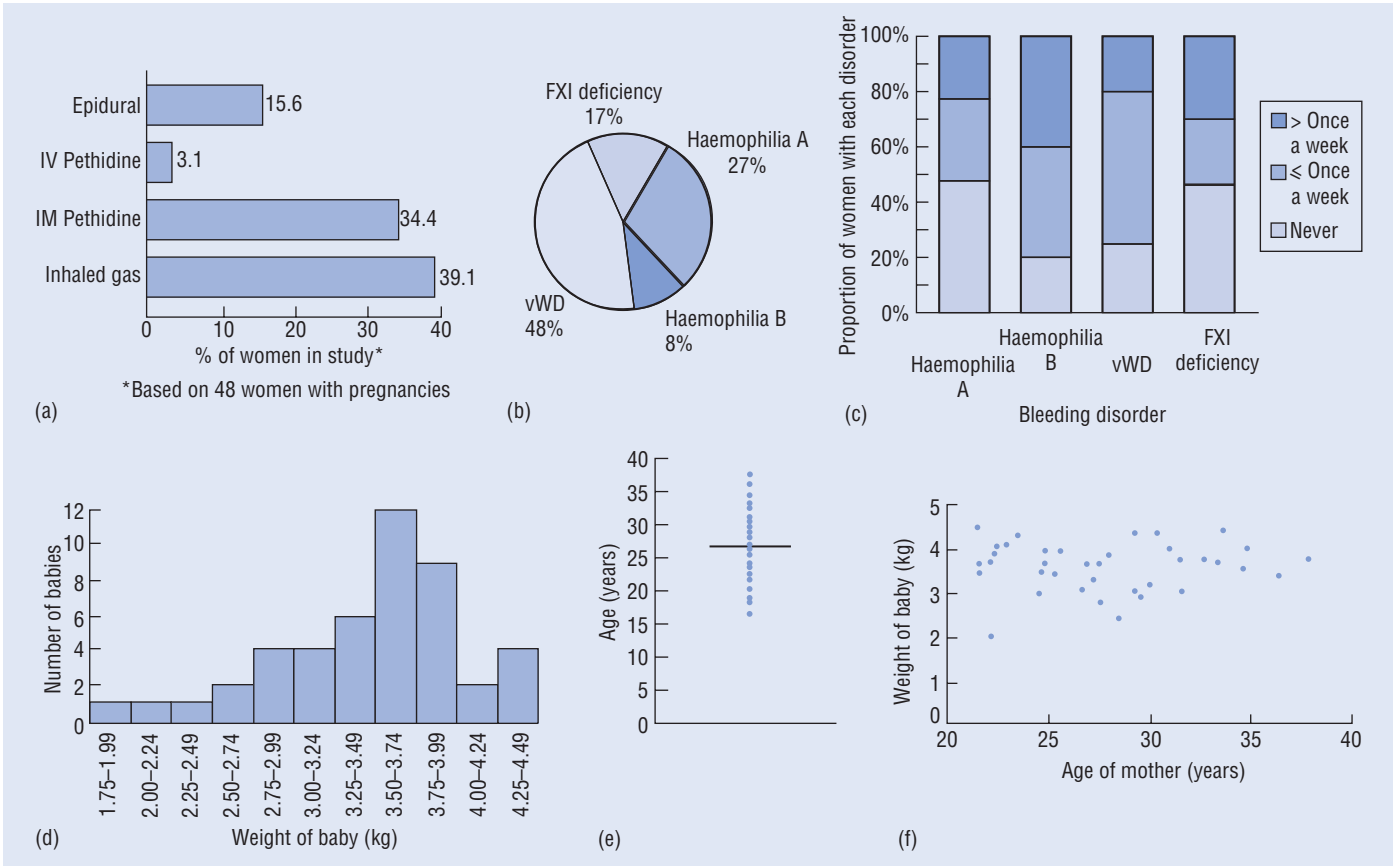


Figure 4.1 A selection of graphical output which may be produced when summarizing the obstetric data in women with bleeding disorders (Chapter 2). (a) **Bar chart** showing the percentage of women in the study who required pain relief from any of the listed interventions during labour. (b) **Pie chart** showing the percentage of women in the study with each bleeding disorder. (c) **Segmented column chart** showing the frequency with which women with different bleeding disorders experience bleeding gums. (d) **Histogram** showing the weight of the baby at birth. (e) **Dot plot** showing the mother's age at the time of the baby's birth, with the median age marked as a horizontal line. (f) **Scatter diagram** showing the relationship between the mother's age at delivery (on the horizontal or *x*-axis) and the weight of the baby (on the vertical or *y*-axis).

Beclomethasone dipropionate		Placebo
3	1.0	04
665	1.1	39
53	1.2	99
9751	1.3	1135677999
955410	1.4	0148
987655	1.5	00338899
9531100	1.6	0001355
731	1.7	00114569
99843110	1.8	6
654400	1.9	01
6	2.0	
7	2.1	19
10	2.2	

Figure 4.2 Stem-and-leaf plot showing the FEV1 (litres) in children receiving inhaled beclomethasone dipropionate or placebo (Chapter 21).

five and 20 groups are chosen; the ranges should be narrow enough to illustrate patterns in the data, but should not be so narrow that they are the raw data. The histogram should be labelled carefully to make it clear where the boundaries lie.

- **Dot plot**—each observation is represented by one dot on a horizontal (or vertical) line (Fig. 4.1e). This type of plot is very simple to draw, but can be cumbersome with large data sets. Often a summary measure of the data, such as the mean or median (Chapter 5), is shown on the diagram. This plot may also be used for discrete data.
- **Stem-and-leaf plot**—This is a mixture of a diagram and a table; it looks similar to a histogram turned on its side, and is effectively the data values written in increasing order of size. It is usually drawn with a vertical **stem**, consisting of the first few digits of the values, arranged in order. Protruding from this stem are the **leaves**—i.e. the final digit of each of the ordered values, which are written horizontally (Fig. 4.2) in increasing numerical order.
- **Box plot** (often called a **box-and-whisker plot**)—This is a vertical or horizontal rectangle, with the ends of the rectangle corresponding to the upper and lower quartiles of the data values (Chapter 6). A line drawn through the rectangle corresponds to the median value (Chapter 5). Whiskers, starting at the ends of the rectangle, usually indicate minimum and maximum values but sometimes relate to particular percentiles, e.g. the 5th and 95th percentiles (Chapter 6, Fig. 6.1). Outliers may be marked.

The 'shape' of the frequency distribution

The choice of the most appropriate statistical method will often depend on the shape of the distribution. The distribution of the data is usually **unimodal** in that it has a single 'peak'. Sometimes the distribution is **bimodal** (two peaks) or **uniform** (each value is equally likely and there are no peaks). When the distribution is unimodal, the main aim is to see where the majority of the data values lie, relative to the maximum and minimum values. In particular, it is important to assess whether the distribution is:

- **symmetrical**—centred around some mid-point, with one side being a mirror-image of the other (Fig. 5.1);
- **skewed to the right (positively skewed)**—a long tail to the right with one or a few high values. Such data are common in medical research (Fig. 5.2);
- **skewed to the left (negatively skewed)**—a long tail to the left with one or a few low values (Fig. 4.1d).

Two variables

If one variable is categorical, then separate diagrams showing the distribution of the second variable can be drawn for each of the categories. Other plots suitable for such data include **clustered** or **segmented** bar or column charts (Fig. 4.1c).

If both of the variables are numerical or ordinal, then the relationship between the two can be illustrated using a **scatter diagram** (Fig. 4.1f). This plots one variable against the other in a two-way diagram. One variable is usually termed the x variable and is represented on the horizontal axis. The second variable, known as the y variable, is plotted on the vertical axis.

Identifying outliers using graphical methods

We can often use single variable data displays to identify outliers. For example, a very long tail on one side of a histogram may indicate an outlying value. However, outliers may sometimes only become apparent when considering the relationship between two variables. For example, a weight of 55 kg would not be unusual for a woman who was 1.6 m tall, but would be unusually low if the woman's height was 1.9 m.

5 Describing data: the 'average'

Summarizing data

It is very difficult to have any 'feeling' for a set of numerical measurements unless we can summarize the data in a meaningful way. A diagram (Chapter 4) is often a useful starting point. We can also condense the information by providing measures that describe the important characteristics of the data. In particular, if we have some perception of what constitutes a representative value, and if we know how widely scattered the observations are around it, then we can formulate an image of the data. The **average** is a general term for a measure of **location**; it describes a typical measurement. We devote this chapter to averages, the most common being the mean and median (Table 5.1). We introduce measures that describe the scatter or **spread** of the observations in Chapter 6.

The arithmetic mean

The **arithmetic mean**, often simply called the mean, of a set of values is calculated by adding up all the values and dividing this sum by the number of values in the set.

It is useful to be able to summarize this verbal description by an algebraic formula. Using mathematical notation, we write our set of n observations of a variable, x, as $x_1, x_2, x_3, \ldots, x_n$. For example, x might represent an individual's height (cm), so that x_1 represents the height of the first individual, and x_i the height of the ith individual, etc. We can write the formula for the arithmetic mean of the observations, written $\bar{x}$ and pronounced 'x bar', as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Using mathematical notation, we can shorten this to:

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

where Σ (the Greek uppercase 'sigma') means 'the sum of', and the sub- and super-scripts on the Σ indicate that we sum the values from $i = 1$ to n. This is often further abbreviated to

$$\bar{x} = \frac{\sum x_i}{n} \quad \text{or to } \bar{x} = \frac{\sum x}{n}$$

The median

If we arrange our data in order of magnitude, starting with the smallest value and ending with the largest value, then the **median** is the middle value of this ordered set. The median divides the ordered values into two halves, with an equal number of values both above and below it.

It is easy to calculate the median if the number of observations, n, is **odd**. It is the $(n + 1)/2$th observation in the ordered set. So, for example, if $n = 11$, then the median is the $(11 + 1)/2 = 12/2 = 6$th observation in the ordered set. If n is **even** then, strictly, there is no median. However, we usually calculate it as the arithmetic mean of the two middle observations in the ordered set [i.e. the $n/2$th and the $(n/2 + 1)$th]. So, for example, if $n = 20$, the median is the arithmetic mean of the $20/2 = 10$th and the $(20/2 + 1) = (10 + 1) = 11$th observations in the ordered set.

The median is similar to the mean if the data are symmetrical (Fig. 5.1), less than the mean if the data are skewed to the right (Fig. 5.2), and greater than the mean if the data are skewed to the left.

The mode

The **mode** is the value that occurs most frequently in a data set; if the data are continuous, we usually group the data and calculate the modal group. Some data sets do not have a mode because each value only occurs once. Sometimes, there is more than one mode; this is when two or more values occur the same number of times, and the frequency of occurrence of each of these values is greater than that of any other value. We rarely use the mode as a summary measure.

The geometric mean

The arithmetic mean is an inappropriate summary measure of location if our data are skewed. If the data are skewed to the right, we can produce a distribution that is more symmetrical if we take the logarithm (to base 10 or to base e) of each value of the variable in this data set (Chapter 9). The arithmetic mean of the log values is a measure of location for the transformed data. To obtain a measure that has the same units as the original observations, we have to back-transform (i.e. take the antilog of) the mean of the log data; we call this the **geometric mean**. Provided the distribution of the log data is approximately symmetrical, the geometric mean is similar to the median and less than the mean of the raw data (Fig. 5.2).

The weighted mean

We use a **weighted mean** when certain values of the variable of interest, x, are more important than others. We attach a weight, w_i, to each of the values, x_i, in our sample, to reflect this importance. If the values $x_1, x_2, x_3, \ldots, x_n$ have corresponding weights $w_1, w_2, w_3, \ldots, w_n$, the weighted arithmetic mean is:

$$\frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum w_i x_i}{\sum w_i}$$

For example, suppose we are interested in determining the average length of stay of hospitalized patients in a district, and we know the average discharge time for patients in every hospital. To take account of the amount of information provided, one approach might be to take each weight as the number of patients in the associated hospital.

The weighted mean and the arithmetic mean are identical if each weight is equal to one.

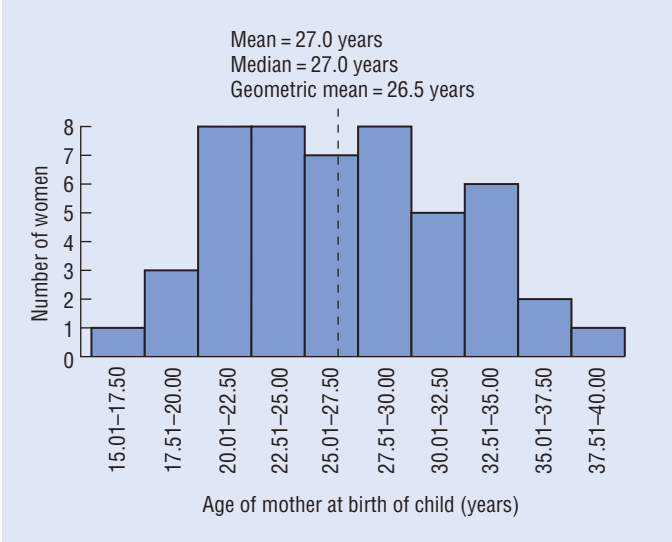


Figure 5.1 The mean, median and geometric mean age of the women in the study described in Chapter 2 at the time of the baby's birth. As the distribution of age appears reasonably symmetrical, the three measures of the 'average' all give similar values, as indicated by the dotted line.

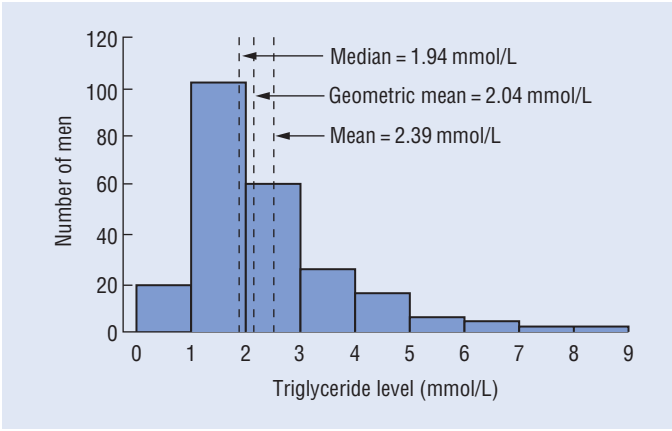


Figure 5.2 The mean, median and geometric mean triglyceride level in a sample of 232 men who developed heart disease (Chapter 19). As the distribution of triglyceride levels is skewed to the right, the mean gives a higher 'average' than either the median or geometric mean.

Table 5.1 Advantages and disadvantages of averages.

Type of average	Advantages	Disadvantages
Mean	• Uses all the data values • Algebraically defined and so mathematically manageable • Known sampling distribution (Chapter 9)	• Distorted by outliers • Distorted by skewed data
Median	• Not distorted by outliers • Not distorted by skewed data	• Ignores most of the information • Not algebraically defined • Complicated sampling distribution
Mode	• Easily determined for categorical data	• Ignores most of the information • Not algebraically defined • Unknown sampling distribution
Geometric mean	• Before back-transformation, it has the same advantages as the mean • Appropriate for right skewed data	• Only appropriate if the log transformation produces a symmetrical distribution
Weighted mean	• Same advantages as the mean • Ascribes relative importance to each observation • Algebraically defined	• Weights must be known or estimated

6 Describing data: the 'spread'

Summarizing data

If we are able to provide two summary measures of a continuous variable, one that gives an indication of the 'average' value and the other that describes the 'spread' of the observations, then we have condensed the data in a meaningful way. We explained how to choose an appropriate average in Chapter 5. We devote this chapter to a discussion of the most common measures of **spread** (**dispersion** or **variability**) which are compared in Table 6.1.

The range

The **range** is the difference between the largest and smallest observations in the data set; you may find these two values quoted instead of their difference. Note that the range provides a misleading measure of spread if there are outliers (Chapter 3).

Ranges derived from percentiles

What are percentiles?

Suppose we arrange our data in order of magnitude, starting with the smallest value of the variable, x, and ending with the largest value. The value of x that has 1% of the observations in the ordered set lying below it (and 99% of the observations lying above it) is called the first **percentile**. The value of x that has 2% of the observations lying below it is called the second percentile, and so on. The values of x that divide the ordered set into 10 equally sized groups, that is the 10th, 20th, 30th, . . . , 90th percentiles, are called **deciles**. The values of x that divide the ordered set into four equally sized groups, that is the 25th, 50th, and 75th percentiles, are called **quartiles**. The 50th percentile is the **median** (Chapter 5).

Using percentiles

We can obtain a measure of spread that is not influenced by outliers by excluding the extreme values in the data set, and determining the range of the remaining observations. The **interquartile range** is the difference between the first and the third quartiles, i.e. between the 25th and 75th percentiles (Fig. 6.1). It contains the central 50% of the observations in the ordered set, with 25% of the observations lying below its lower limit, and 25% of them lying above its upper limit. The **interdecile range** contains the central 80% of the observations, i.e. those lying between the 10th and 90th percentiles. Often we use the range that contains the central 95% of the observations, i.e. it excludes 2.5% of the observations above its upper limit and 2.5% below its lower limit (Fig. 6.1). We may use this interval, provided it is calculated from enough values of the variable in healthy individuals, to diagnose disease. It is then called the **reference interval**, **reference range** or **normal range** (see Chapter 38).

The variance

One way of measuring the spread of the data is to determine the extent to which each observation deviates from the arithmetic mean. Clearly, the larger the deviations, the greater the variability of the observations. However, we cannot use the mean of these deviations as a measure of spread because the positive differences exactly cancel out the negative differences. We overcome this problem by squaring each deviation, and finding the mean of these squared deviations (Fig. 6.2); we call this the **variance**. If we have a sample of n observations, $x_1, x_2, x_3, \ldots, x_n$, whose mean is $\bar{x} = (\Sigma x_i)/n$, we calculate the variance, usually denoted by s^2, of these observations as:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

We can see that this is not quite the same as the arithmetic mean of the squared deviations because we have divided by $n - 1$ instead

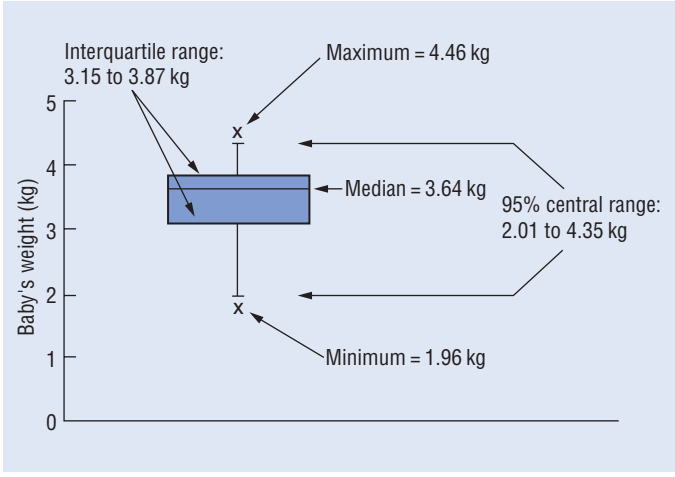


Figure 6.1 A box-and-whisker plot of the baby's weight at birth (Chapter 2). This figure illustrates the median, the interquartile range, the range that contains the central 95% of the observations and the maximum and minimum values.

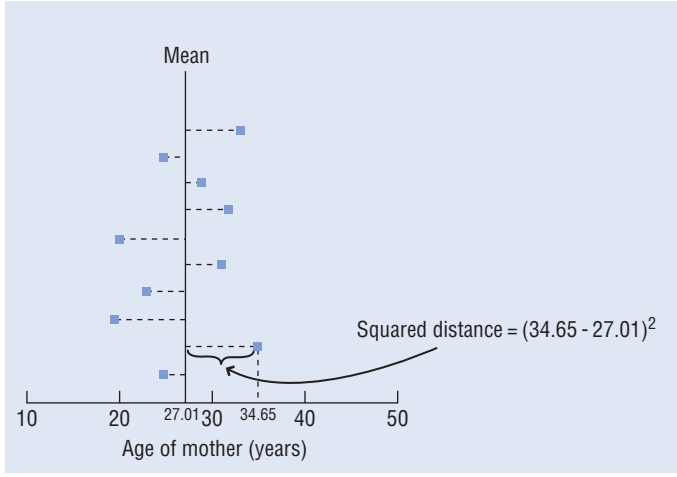


Figure 6.2 Diagram showing the spread of selected values of the mother's age at the time of baby's birth (Chapter 2) around the mean value. The variance is calculated by adding up the squared distances between each point and the mean, and dividing by $(n - 1)$.

of n. The reason for this is that we almost always rely on *sample* data in our investigations (Chapter 10). It can be shown theoretically that we obtain a better sample estimate of the population variance if we divide by $(n-1)$.

The units of the variance are the square of the units of the original observations, e.g. if the variable is weight measured in kg, the units of the variance are kg^2.

The standard deviation

The **standard deviation** is the square root of the variance. In a sample of n observations, it is:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

We can think of the standard deviation as a sort of average of the deviations of the observations from the mean. It is evaluated in the same units as the raw data.

If we divide the standard deviation by the mean and express this quotient as a percentage, we obtain the **coefficient of variation**. It is a measure of spread that is independent of the units of measurement, but it has theoretical disadvantages so is not favoured by statisticians.

Variation within- and between-subjects

If we take repeated measurements of a continuous variable on an individual, then we expect to observe some variation (**intra**- or **within-subject** variability) in the responses on that individual. This may be because a given individual does not always respond in exactly the same way and/or because of measurement error. However, the variation within an individual is usually less than the variation obtained when we take a single measurement on every individual in a group (**inter**- or **between-subject** variability). For example, a 17-year-old boy has a lung vital capacity that ranges between 3.60 and 3.87 litres when the measurement is repeated 10 times; the values for single measurements on 10 boys of the same age lie between 2.98 and 4.33 litres. These concepts are important in study design (Chapter 13).

Table 6.1 Advantages and disadvantages of measures of spread.

Measure of spread	Advantages	Disadvantages
Range	• Easily determined	• Uses only two observations • Distorted by outliers • Tends to increase with increasing sample size
Ranges based on percentiles	• Usually unaffected by outliers • Independent of sample size • Appropriate for skewed data	• Clumsy to calculate • Cannot be calculated for small samples • Uses only two observations • Not algebraically defined
Variance	• Uses every observation • Algebraically defined	• Units of measurement are the square of the units of the raw data • Sensitive to outliers • Inappropriate for skewed data
Standard deviation	• Same advantages as the variance • Units of measurement are the same as those of the raw data • Easily interpreted	• Sensitive to outliers • Inappropriate for skewed data

7 Theoretical distributions: the Normal distribution

In Chapter 4 we showed how to create an **empirical frequency distribution** of the observed data. This contrasts with a theoretical **probability distribution** which is described by a mathematical model. When our empirical distribution approximates a particular probability distribution, we can use our theoretical knowledge of that distribution to answer questions about the data. This often requires the evaluation of probabilities.

Understanding probability

Probability measures uncertainty; it lies at the heart of statistical theory. A probability measures the chance of a given event occurring. It is a positive number that lies between zero and one. If it is equal to zero, then the event *cannot* occur. If it is equal to one, then the event *must* occur. The probability of the **complementary** event (the event *not* occurring) is one minus the probability of the event occurring. We discuss **conditional probability**, the probability of an event, given that another event has occurred, in Chapter 45.

We can calculate a probability using various approaches.

- **Subjective**—our personal degree of belief that the event will occur (e.g. that the world will come to an end in the year 2050).
- **Frequentist**—the proportion of times the event would occur if we were to repeat the experiment a large number of times (e.g. the number of times we would get a 'head' if we tossed a fair coin 1000 times).
- ***A priori***—this requires knowledge of the theoretical *model*, called the **probability distribution**, which describes the probabilities of all possible outcomes of the 'experiment'. For example, genetic theory allows us to describe the probability distribution for eye colour in a baby born to a blue-eyed woman and brown-eyed man by initially specifying all possible genotypes of eye colour in the baby and their probabilities.

The rules of probability

We can use the rules of probability to add and multiply probabilities.

- **The addition rule**—if two events, A and B, are *mutually exclusive* (i.e. each event precludes the other), then the probability that either one or the other occurs is equal to the sum of their probabilities.

$$\text{Prob}(\text{A } \textit{or } \text{B}) = \text{Prob}(\text{A}) + \text{Prob}(\text{B})$$

e.g. if the probabilities that an adult patient in a particular dental practice has no missing teeth, some missing teeth or is edentulous (i.e. has no teeth) are 0.67, 0.24 and 0.09, respectively, then the probability that a patient has some teeth is $0.67 + 0.24 = 0.91$.

- **The multiplication rule**—if two events, A and B, are *independent* (i.e. the occurrence of one event is not contingent on the other), then the probability that both events occur is equal to the product of the probability of each:

$$\text{Prob}(\text{A } \textit{and } \text{B}) = \text{Prob}(\text{A}) \times \text{Prob}(\text{B})$$

e.g. if two unrelated patients are waiting in the dentist's surgery, the probability that both of them have no missing teeth is $0.67 \times 0.67 = 0.45$.

Probability distributions: the theory

A **random variable** is a quantity that can take any one of a set of mutually exclusive values with a given probability. A **probability distribution** shows the probabilities of all possible values of the random variable. It is a theoretical distribution that is expressed mathematically, and has a mean and variance that are analogous to those of an empirical distribution. Each probability distribution is defined by certain **parameters** which are summary measures (e.g. mean, variance) characterizing that distribution (i.e. knowledge of them allows the distribution to be fully described). These parameters are estimated in the sample by relevant **statistics**. Depending on whether the random variable is discrete or continuous, the probability distribution can be either discrete or continuous.

- **Discrete** (e.g. Binomial, Poisson)—we can derive probabilities corresponding to every possible value of the random variable. ***The sum of all such probabilities is one***.
- **Continuous** (e.g. Normal, Chi-squared, t and F)—we can only derive the probability of the random variable, x, taking values in certain ranges (because there are infinitely many values of x). If the horizontal axis represents the values of x, we can draw a curve from the equation of the distribution (the **probability density function**); it resembles an empirical relative frequency distribution (Chapter 4). ***The total area under the curve is one***; this area represents the probability of all possible events. The probability that x lies between two limits is equal to the area under the curve between these values (Fig. 7.1). For convenience, tables (Appendix A) have been produced to enable us to evaluate probabilities of interest for commonly used continuous probability distributions. These are particularly useful in the context of confidence intervals (Chapter 11) and hypothesis testing (Chapter 17).

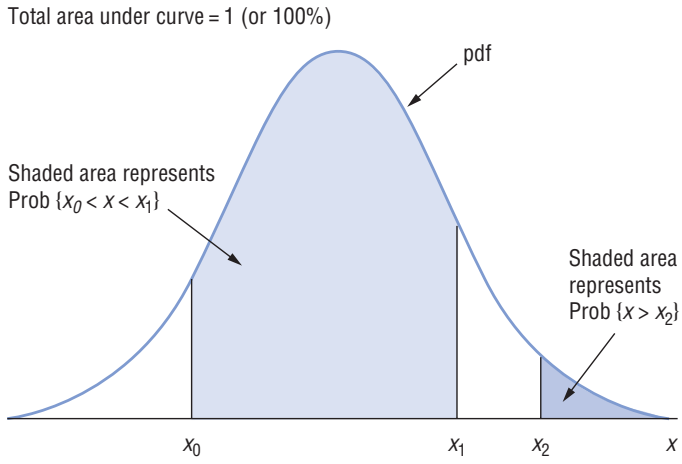


Figure 7.1 The probability density function, pdf, of x.

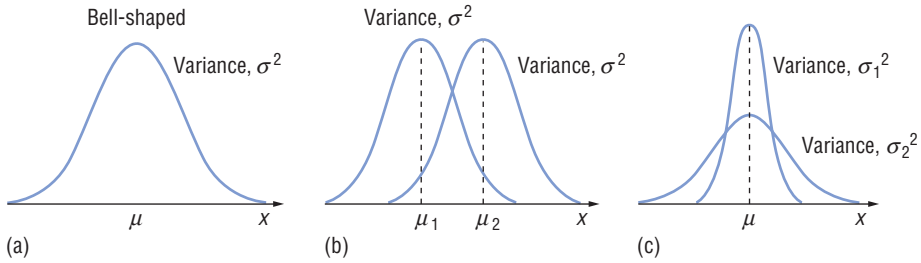


Figure 7.2 The probability density function of the Normal distribution of the variable, x. (a) Symmetrical about mean, μ: variance = σ^2. (b) Effect of changing mean ($\mu_2 > \mu_1$). (c) Effect of changing variance ($\sigma_1^2 < \sigma_2^2$).

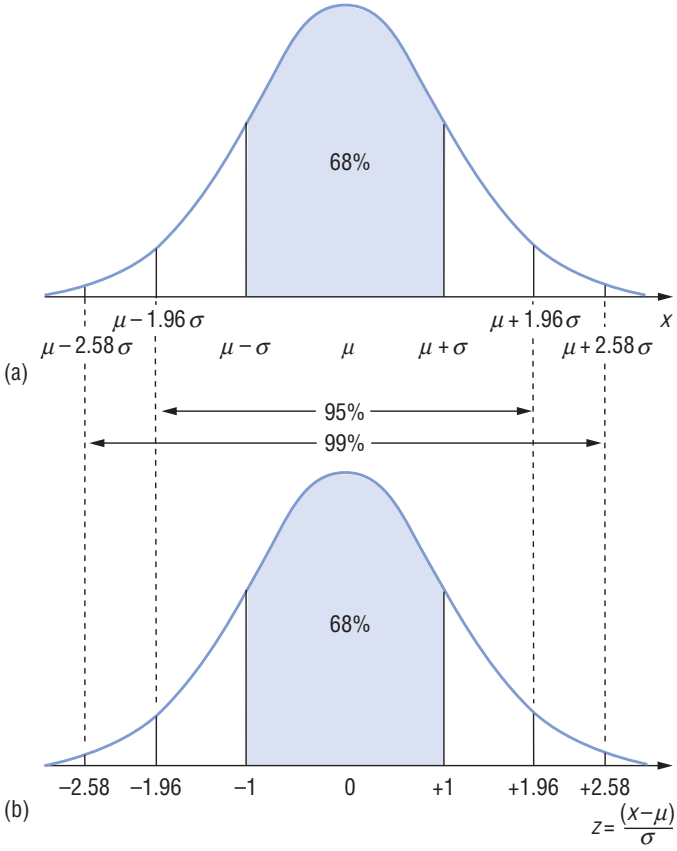


Figure 7.3 Areas (percentages of total probability) under the curve for (a) Normal distribution of x, with mean μ and variance σ^2, and (b) Standard Normal distribution of z.

The Normal (Gaussian) distribution

One of the most important distributions in statistics is the **Normal distribution**. Its probability density function (Fig. 7.2) is:

- completely described by two parameters, the *mean* (μ) and the *variance* (σ^2);
- bell-shaped (unimodal);
- symmetrical about its mean;
- shifted to the right if the mean is increased and to the left if the mean is decreased (assuming constant variance);
- flattened as the variance is increased but becomes more peaked as the variance is decreased (for a fixed mean).

Additional properties are that:

- the mean and median of a Normal distribution are equal;
- the probability (Fig. 7.3a) that a Normally distributed random variable, x, with mean, μ, and standard deviation, σ, lies between:

$(\mu - \sigma)$ and $(\mu + \sigma)$ is 0.68

$(\mu - 1.96\sigma)$ and $(\mu + 1.96\sigma)$ is 0.95

$(\mu - 2.58\sigma)$ and $(\mu + 2.58\sigma)$ is 0.99

These intervals may be used to define **reference intervals** (Chapters 6 and 38).

We show how to assess Normality in Chapter 35.

The Standard Normal distribution

There are infinitely many Normal distributions depending on the values of μ and σ. The Standard Normal distribution (Fig. 7.3b) is a particular Normal distribution for which probabilities have been tabulated (Appendix A1, A4).

- The Standard Normal distribution has a **mean of zero** and a **variance of one**.
- If the random variable, x, has a Normal distribution with mean, μ, and variance, σ^2, then the **Standardized Normal Deviate (SND)**, $z = \frac{x - \mu}{\sigma}$, is a random variable that has a Standard Normal distribution.

8 Theoretical distributions: other distributions

Some words of comfort

Do not worry if you find the theory underlying probability distributions complex. Our experience demonstrates that you want to know only when and how to use these distributions. We have therefore outlined the essentials, and omitted the equations that define the probability distributions. You will find that you only need to be familiar with the basic ideas, the terminology and, perhaps (although infrequently in this computer age), know how to refer to the tables.

More continuous probability distributions

These distributions are based on continuous random variables. Often it is not a measurable variable that follows such a distribution, but a statistic derived from the variable. The total area under the probability density function represents the probability of all possible outcomes, and is equal to one (Chapter 7). We discussed the Normal distribution in Chapter 7; other common distributions are described in this chapter.

The *t*-distribution (Appendix A2, Fig. 8.1)

- Derived by W.S. Gossett, who published under the pseudonym 'Student', it is often called Student's *t*-distribution.
- The parameter that characterizes the *t*-distribution is the **degrees of freedom**, so we can draw the probability density function if we know the equation of the *t*-distribution and its degrees of freedom. We discuss degrees of freedom in Chapter 11; note that they are often closely affiliated to sample size.
- Its shape is similar to that of the Standard Normal distribution, but it is more spread out with longer tails. Its shape approaches Normality as the degrees of freedom increase.
- It is particularly useful for calculating confidence intervals for and testing hypotheses about one or two means (Chapters 19–21).

The Chi-squared (χ^2) distribution (Appendix A3, Fig. 8.2)

- It is a right skewed distribution taking positive values.
- It is characterized by its **degrees of freedom** (Chapter 11).
- Its shape depends on the degrees of freedom; it becomes more symmetrical and approaches Normality as the degrees of freedom increases.
- It is particularly useful for analysing categorical data (Chapters 23–25).

The *F*-distribution (Appendix A5)

- It is skewed to the right.
- It is defined by a ratio. The distribution of a ratio of two estimated variances calculated from Normal data approximates the *F*-distribution.
- The two parameters which characterize it are the **degrees of freedom** (Chapter 11) of the numerator and the denominator of the ratio.
- The *F*-distribution is particularly useful for comparing two variances (Chapter 18), and more than two means using the analysis of variance (ANOVA) (Chapter 22).

The Lognormal distribution

- It is the probability distribution of a random variable whose log (to base 10 or *e*) follows the Normal distribution.
- It is highly skewed to the right (Fig. 8.3a).
- If, when we take logs of our raw data that are skewed to the right, we produce an empirical distribution that is nearly Normal (Fig. 8.3b), our data approximate the Lognormal distribution.
- Many variables in medicine follow a Lognormal distribution. We can use the properties of the Normal distribution (Chapter 7) to make inferences about these variables after transforming the data by taking logs.
- If a data set has a Lognormal distribution, we can use the geometric mean (Chapter 5) as a summary measure of location.

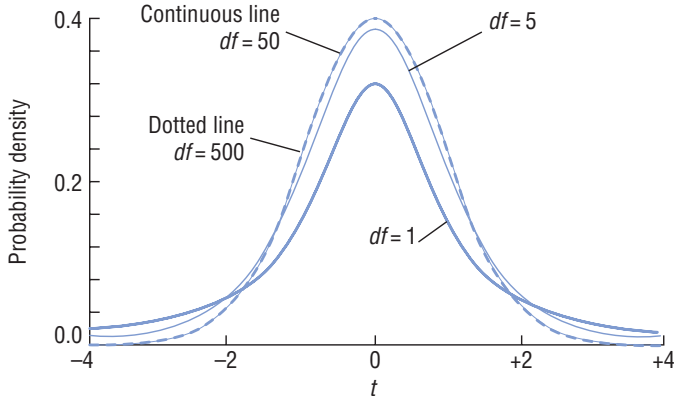


Figure 8.1 *t*-distributions with degrees of freedom (*df*) = 1, 5, 50, and 500.

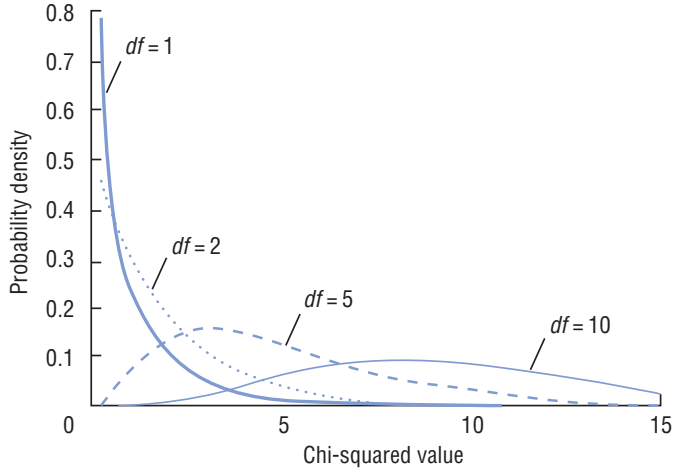


Figure 8.2 Chi-squared distributions with degrees of freedom (*df*) = 1, 2, 5, and 10.

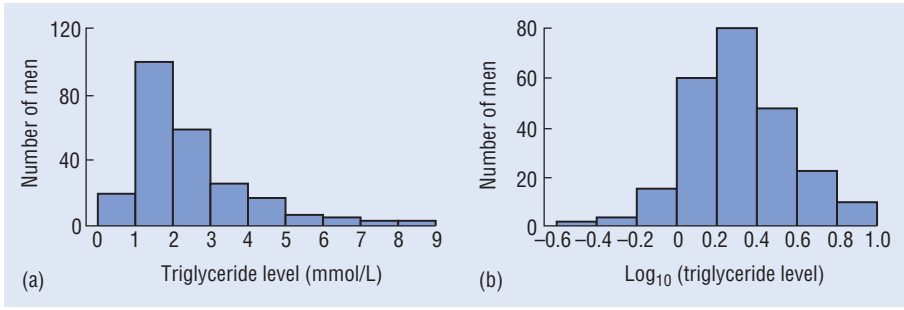


Figure 8.3 (a) The Lognormal distribution of triglyceride levels in 232 men who developed heart disease (Chapter 19). (b) The approximately Normal distribution of $\log_{10}$ (triglyceride level).

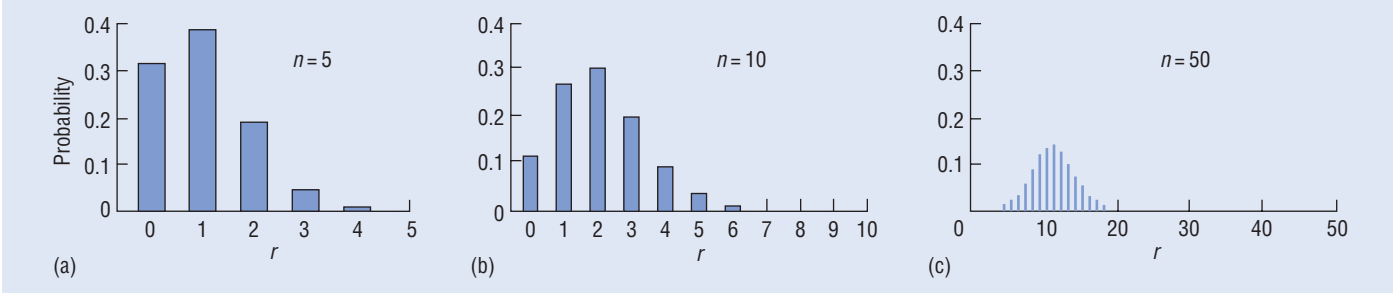


Figure 8.4 Binomial distribution showing the number of successes, r, when the probability of success is $\pi = 0.20$ for sample sizes (a) $n = 5$, (b) $n = 10$, and (c) $n = 50$. (N.B. in Chapter 23, the observed seroprevalence of HHV-8 was $p = 0.187 \approx 0.2$, and the sample size was 271: the proportion was assumed to follow a Normal distribution.)

Discrete probability distributions

The random variable that defines the probability distribution is discrete. The sum of the probabilities of all possible mutually exclusive events is one.

The Binomial distribution

- Suppose, in a given situation, there are only two outcomes, 'success' and 'failure'. For example, we may be interested in whether a woman conceives (a success) or does not conceive (a failure) after *in vitro* fertilization (IVF). If we look at $n = 100$ unrelated women undergoing IVF (each with the same probability of conceiving), the Binomial random variable is the observed number of conceptions (successes). Often this concept is explained in terms of n independent repetitions of a trial (e.g. 100 tosses of a coin) in which the outcome is either success (e.g. head) or failure.
- The two parameters that describe the Binomial distribution are ***n***, the number of individuals in the sample (or repetitions of a trial) and π, the true probability of success for each individual (or in each trial).
- Its **mean** (the value for the random variable that we *expect* if we look at n individuals, or repeat the trial n times) is $n\pi$. Its **variance** is $n\pi(1 - \pi)$.
- When n is small, the distribution is skewed to the right if $\pi < 0.5$ and to the left if $\pi > 0.5$. The distribution becomes more symmetrical as the sample size increases (Fig. 8.4) and approximates the Normal distribution if both $n\pi$ and $n(1 - \pi)$ are greater than 5.
- We can use the properties of the Binomial distribution when making inferences about **proportions**. In particular, we often use the Normal approximation to the Binomial distribution when analysing proportions.

The Poisson distribution

- The Poisson random variable is the **count** of the number of events that occur independently and randomly in time or space at some average rate, μ. For example, the number of hospital admissions per day typically follows the Poisson distribution. We can use our knowledge of the Poisson distribution to calculate the probability of a certain number of admissions on any particular day.
- The parameter that describes the Poisson distribution is the **mean**, i.e. the average rate, μ.
- The **mean** equals the **variance** in the Poisson distribution.
- It is a right skewed distribution if the mean is small, but becomes more symmetrical as the mean increases, when it approximates a Normal distribution.

9 Transformations

Why transform?

The observations in our investigation may not comply with the requirements of the intended statistical analysis (Chapter 35).

- A variable may not be Normally distributed, a **distributional** requirement for many different analyses.
- The spread of the observations in each of a number of groups may be different (constant variance is an assumption about a **parameter** in the comparison of means using the unpaired *t*-test and analysis of variance—Chapters 21–22).
- Two variables may not be linearly related (**linearity** is an assumption in many regression analyses—Chapters 27–33 and 42).

It is often helpful to **transform** our data to satisfy the assumptions underlying the proposed statistical techniques.

How do we transform?

We convert our raw data into transformed data by taking the same mathematical transformation of each observation. Suppose we have n observations $(y_1, y_2, \ldots, y_n)$ on a variable, y, and we decide that the log transformation is suitable. We take the log of each observation to produce $(\log y_1, \log y_2, \ldots, \log y_n)$. If we call the transformed variable, z, then $z_i = \log y_i$ for each i $(i = 1, 2, \ldots, n)$, and our transformed data may be written $(z_1, z_2, \ldots, z_n)$.

We check that the transformation has achieved its purpose of producing a data set that satisfies the assumptions of the planned statistical analysis (e.g. by plotting a histogram of the transformed data. See Chapter 35), and proceed to analyse the transformed data $(z_1, z_2, \ldots, z_n)$. We often back-transform any summary measures (such as the mean) to the original scale of measurement; we then rely on the conclusions we draw from hypothesis tests (Chapter 17) on the transformed data.

Typical transformations

The logarithmic transformation, $z = \log y$

When log transforming data, we can choose to take logs either to base 10 ($\log_{10} y$, the 'common' log) or to base e ($\log_e y = \ln y$, the 'natural' or Naperian log) or to any other base, but must be consistent for a particular variable in a data set. Note that we cannot take the log of a negative number or of zero. The back-transformation of a log is called the antilog; the antilog of a Naperian log is the exponential, e.

- If y is skewed to the right, $z = \log y$ is often approximately **Normally distributed** (Fig. 9.1a). Then y has a Lognormal distribution (Chapter 8).
- If there is an exponential relationship between y and another variable, x, so that the resulting curve bends upwards when y (on the vertical axis) is plotted against x (on the horizontal axis), then the relationship between $z = \log y$ and x is approximately **linear** (Fig. 9.1b).
- Suppose we have different groups of observations, each comprising measurements of a continuous variable, y. We may find that the groups that have the higher values of y also have larger variances. In particular, if the coefficient of variation (the standard deviation divided by the mean) of y is constant for all the groups, the log transformation, $z = \log y$, produces groups that have **similar variances** (Fig. 9.1c).

In medicine, the log transformation is frequently used because of its logical interpretation and because many variables have right-skewed distributions.

The square root transformation, $z = \sqrt{y}$

This transformation has properties that are similar to those of the log transformation, although the results after they have been back-transformed are more complicated to interpret. In addition to its **Normalizing** and **linearizing** abilities, it is effective at **stabilizing variance** if the variance increases with increasing values of y, i.e. if the variance divided by the mean is constant. We often apply the square root transformation if y is the count of a rare event occurring in time or space, i.e. it is a Poisson variable (Chapter 8). Remember, we cannot take the square root of a negative number.

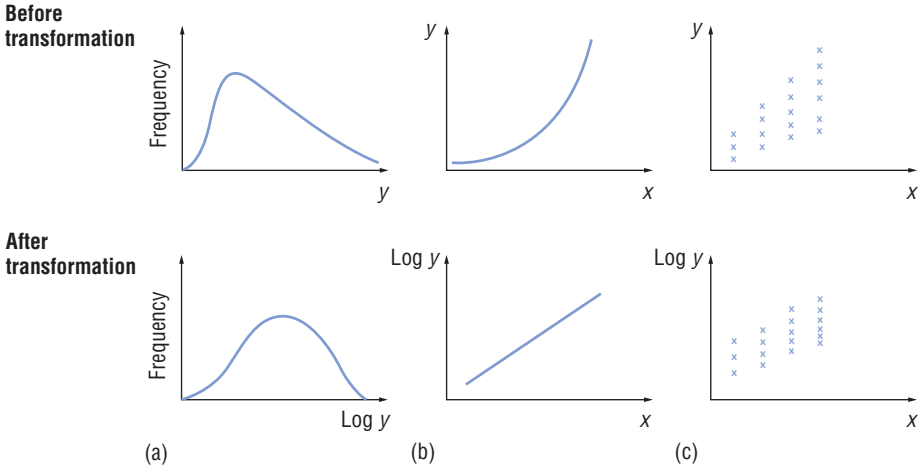


Figure 9.1 The effects of the logarithmic transformation: (a) Normalizing, (b) linearizing, (c) variance stabilizing.

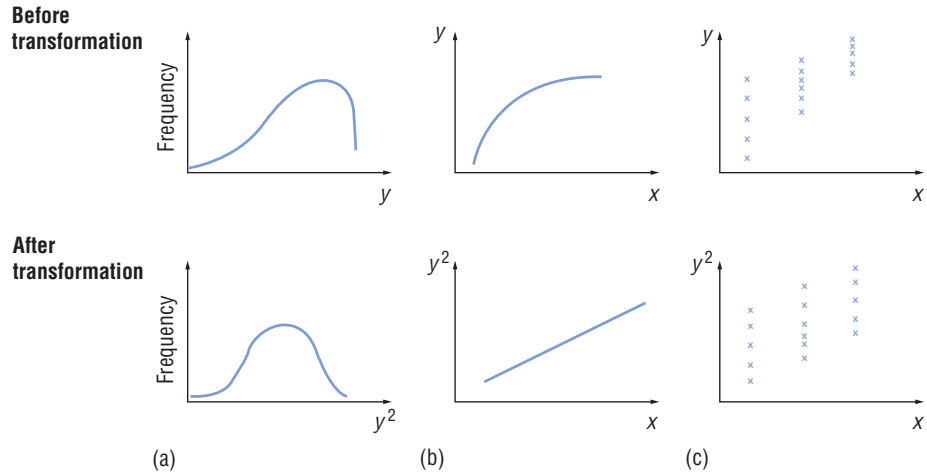


Figure 9.2 The effect of the square transformation: (a) Normalizing, (b) linearizing, (c) variance stabilizing.

The reciprocal transformation, $z = 1/y$

We often apply the reciprocal transformation to survival times unless we are using special techniques for survival analysis (Chapter 41). The reciprocal transformation has properties that are similar to those of the log transformation. In addition to its **Normalizing** and **linearizing** abilities, it is more effective at **stabilizing variance** than the log transformation if the variance increases very markedly with increasing values of y, i.e. if the variance divided by the (mean)^4 is constant. Note that we cannot take the reciprocal of zero.

The square transformation, $z = y^2$

The square transformation achieves the reverse of the log transformation.

- If y is skewed to the left, the distribution of $z = y^2$ is often approximately **Normal** (Fig. 9.2a).
- If the relationship between two variables, x and y, is such that a line curving downwards is produced when we plot y against x, then the relationship between $z = y^2$ and x is approximately **linear** (Fig. 9.2b).
- If the variance of a continuous variable, y, tends to decrease as the value of y increases, then the square transformation, $z = y^2$, **stabilizes the variance** (Fig. 9.2c).

The logit (logistic) transformation, $z = \ln \frac{p}{1-p}$

This is the transformation we apply most often to each proportion, p, in a set of proportions. We cannot take the logit transformation if either $p = 0$ or $p = 1$ because the corresponding logit values are $-\infty$ and $+\infty$. One solution is to take p as $1/(2n)$ instead of 0, and as $\{1 - 1/(2n)\}$ instead of 1, where n is the sample size.

It **linearizes** a sigmoid curve (Fig. 9.3). See Chapter 30 for the use of the logit transformation in regression analysis.

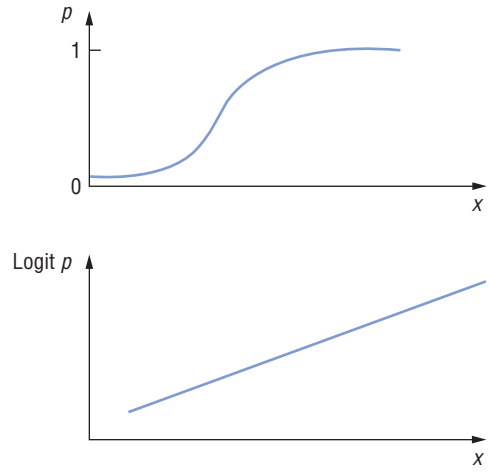


Figure 9.3 The effect of the logit transformation on a sigmoid curve.

10 Sampling and sampling distributions

Why do we sample?

In statistics, a **population** represents the entire group of individuals in whom we are interested. Generally it is costly and labour-intensive to study the entire population and, in some cases, may be impossible because the population may be hypothetical (e.g. patients who may receive a treatment in the future). Therefore we collect data on a **sample** of individuals who we believe are **representative** of this population (i.e. they have similar characteristics to the individuals in the population), and use them to draw conclusions (i.e. make **inferences**) about the population.

When we take a sample of the population, we have to recognize that the information in the sample may not fully reflect what is true in the population. We have introduced **sampling error** by studying only some of the population. In this chapter we show how to use theoretical probability distributions (Chapters 7 and 8) to quantify this error.

Obtaining a representative sample

Ideally, we aim for a **random sample**. A list of all individuals from the population is drawn up (the **sampling frame**), and individuals are selected randomly from this list, i.e. every possible sample of a given size in the population has an equal probability of being chosen. Sometimes, we may have difficulty in constructing this list or the costs involved may be prohibitive, and then we take a **convenience sample**. For example, when studying patients with a particular clinical condition, we may choose a single hospital, and investigate some or all of the patients with the condition in that hospital. Very occasionally, non-random schemes, such as **quota sampling** or **systematic sampling**, may be used. Although the statistical tests described in this book assume that individuals are selected for the sample randomly, the methods are generally reasonable as long as the sample is **representative** of the population.

Point estimates

We are often interested in the value of a **parameter** in the population (Chapter 7), e.g. a mean or a proportion. Parameters are usually denoted by letters of the Greek alphabet. For example, we usually refer to the population mean as μ and the population standard deviation as σ. We estimate the value of the parameter using the data collected from the sample. This estimate is referred to as the **sample statistic** and is a **point estimate** of the parameter (i.e. it takes a single value) as opposed to an **interval estimate** (Chapter 11) which takes a range of values.

Sampling variation

If we take repeated samples of the same size from a population, it is unlikely that the estimates of the population parameter would be exactly the same in each sample. However, our estimates should all be close to the true value of the parameter in the population, and the estimates themselves should be similar to each other. By quantifying the variability of these estimates, we obtain information on the precision of our estimate and can thereby assess the sampling error.

In reality, we usually only take one sample from the population. However, we still make use of our knowledge of the theoretical distribution of sample estimates to draw inferences about the population parameter.

Sampling distribution of the mean

Suppose we are interested in estimating the population mean; we could take many repeated samples of size n from the population, and estimate the mean in each sample. A histogram of the estimates of these means would show their distribution (Fig. 10.1); this is the **sampling distribution of the mean**. We can show that:

- If the sample size is reasonably large, the estimates of the mean follow a **Normal** distribution, whatever the distribution of the original data in the population (this comes from a theorem known as the *Central Limit Theorem*).
- If the sample size is small, the estimates of the mean follow a Normal distribution provided the data in the population follow a Normal distribution.
- The mean of the estimates is an **unbiased** estimate of the true mean in the population, i.e. the mean of the estimates equals the true population mean.
- The variability of the distribution is measured by the standard deviation of the estimates; this is known as the **standard error of the mean** (often denoted by SEM). If we know the population standard deviation (σ), then the standard error of the mean is given by:

$$\text{SEM} = \sigma/\sqrt{n}$$

When we only have one sample, as is customary, our best estimate of the population mean is the sample mean, and because we rarely know the standard deviation in the population, we estimate the standard error of the mean by:

$$\text{SEM} = s/\sqrt{n}$$

where s is the standard deviation of the observations in the sample (Chapter 6). The SEM provides a measure of the precision of our estimate.

Interpreting standard errors

- A **large** standard error indicates that the estimate is ***imprecise***.
- A **small** standard error indicates that the estimate is ***precise***.

The standard error is reduced, i.e. we obtain a more precise estimate, if:

- the size of the sample is increased (Fig. 10.1);
- the data are less variable.

SD or SEM?

Although these two parameters seem to be similar, they are used for different purposes. The standard deviation describes the variation in the data values and should be quoted if you wish to illustrate variability in the data. In contrast, the standard error describes the precision of the sample mean, and should be quoted if you are interested in the mean of a set of data values.

Sampling distribution of the proportion

We may be interested in the proportion of individuals in a population who possess some characteristic. Having taken a sample of size n from the population, our best estimate, p, of the population proportion, π, is given by:

$$p = r/n$$

where r is the number of individuals in the sample with the characteristic. If we were to take repeated samples of size n from our population and plot the estimates of the proportion as a histogram, the resulting **sampling distribution of the proportion** would approximate a Normal distribution with mean value, π. The standard deviation of this distribution of estimated proportions is the **standard error of the proportion**. When we take only a single sample, it is estimated by:

$$\mathrm{SE}(p) = \sqrt{\frac{p(1-p)}{n}}$$

This provides a measure of the precision of our estimate of π; a small standard error indicates a precise estimate.

Example

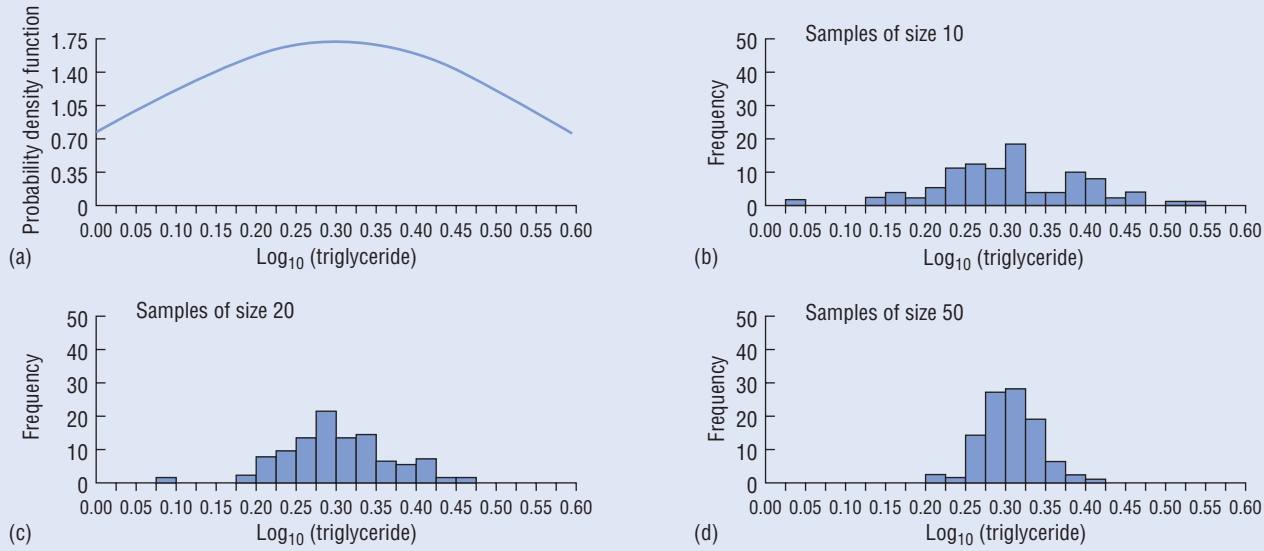


Figure 10.1 (a) Theoretical Normal distribution of $\log_{10}$ (triglyceride levels) with mean = 0.31 $\log_{10}$(mmol/L) and standard deviation = 0.24 $\log_{10}$(mmol/L), and the observed distributions of the means of 100 random samples of size (b)10, (c)20, and (d)50 taken from this theoretical distribution.

11 Confidence intervals

Once we have taken a sample from our population, we obtain a point estimate (Chapter 10) of the parameter of interest, and calculate its standard error to indicate the precision of the estimate. However, to most people the standard error is not, by itself, particularly useful. It is more helpful to incorporate this measure of precision into an **interval estimate** for the population parameter. We do this by making use of our knowledge of the theoretical probability distribution of the sample statistic to calculate a **confidence interval** for the parameter. Generally the confidence interval extends either side of the estimate by some multiple of the standard error; the two values (the **confidence limits**) which define the interval are generally separated by a comma, a dash or the word 'to' and are contained in brackets.

Confidence interval for the mean

Using the Normal distribution

In Chapter 10, we stated that the sample mean follows a Normal distribution if the sample size is large. Therefore we can make use of the properties of the Normal distribution when considering the sample mean. In particular, 95% of the distribution of sample means lies within 1.96 standard deviations (SD) of the population mean. We call this SD the standard error of the mean (SEM), and when we have a single sample, the 95% **confidence interval (CI) for the mean** is:

(Sample mean $-$ (1.96 $\times$ SEM) to Sample mean + (1.96 $\times$ SEM))

If we were to repeat the experiment many times, this range of values would contain the true population mean on 95% of occasions. This range is known as the 95% confidence interval for the mean. We usually interpret this confidence interval as the range of values within which we are 95% confident that the true population mean lies. Although not strictly correct (the population mean is a fixed value and therefore cannot have a probability attached to it), we will interpret the confidence interval in this way as it is conceptually easier to understand.

Using the *t*-distribution

Strictly, we should only use the Normal distribution in the calculation if we know the value of the variance, σ^2, in the population. Furthermore, if the sample size is small, the sample mean only follows a Normal distribution if the underlying population data are Normally distributed. Where the data are not Normally distributed, and/or we do not know the population variance but estimate it by s^2, the sample mean follows a *t*-distribution (Chapter 8). We calculate the 95% confidence interval for the mean as:

(Sample mean $-$ ($t_{0.05}$ $\times$ SEM) to Sample mean + ($t_{0.05}$ $\times$ SEM))

$$\text{i.e. it is} \quad \text{Sample mean} \pm t_{0.05} \times \frac{s}{\sqrt{n}}$$

where $t_{0.05}$ is the **percentage point** (percentile) of the *t*-distribution with $(n-1)$ degrees of freedom which gives a two-tailed probability (Chapter 17) of 0.05 (Appendix A2). This generally provides a slightly wider confidence interval than that using the Normal distribution to allow for the extra uncertainty that we have introduced by estimating the population standard deviation and/or because of the small sample size. When the sample size is large, the difference between the two distributions is negligible. *Therefore, we always use the t-distribution when calculating a confidence interval for the mean even if the sample size is large.*

By convention we usually quote 95% confidence intervals. We could calculate other confidence intervals e.g. a 99% confidence interval for the mean. Instead of multiplying the standard error by the tabulated value of the *t*-distribution corresponding to a two-tailed probability of 0.05, we multiply it by that corresponding to a two-tailed probability of 0.01. The 99% confidence interval is wider than a 95% confidence interval, to reflect our increased confidence that the range includes the true population mean.

Confidence interval for the proportion

The sampling distribution of a proportion follows a Binomial distribution (Chapter 8). However, if the sample size, n, is reasonably large, then the sampling distribution of the proportion is approximately Normal with mean, π. We estimate π by the proportion in the sample, $p = r/n$ (where r is the number of individuals in the sample with the characteristic of interest), and its standard error is estimated by $\sqrt{\frac{p(1-p)}{n}}$ (Chapter 10).

The **95% confidence interval for the proportion** is estimated by:

$$\left(p - \left[1.96 \times \sqrt{\frac{p(1-p)}{n}}\right] \text{ to } p + \left[1.96 \times \sqrt{\frac{p(1-p)}{n}}\right]\right)$$

If the sample size is small (usually when np or $n(1\text{-}p)$ is less than 5) then we have to use the Binomial distribution to calculate exact confidence intervals[1]. Note that if p is expressed as a percentage, we replace $(1-p)$ by $(100-p)$.

Interpretation of confidence intervals

When interpreting a confidence interval we are interested in a number of issues.

- **How wide is it?** A wide interval indicates that the estimate is imprecise; a narrow one indicates a precise estimate. The width of the confidence interval depends on the size of the standard error, which in turn depends on the sample size and, when considering a numerical variable, the variability of the data. Therefore, small studies on variable data give wider confidence intervals than larger studies on less variable data.
- **What clinical implications can be derived from it?** The upper and lower limits provide a means of assessing whether the results are clinically important (see Example).
- **Does it include any values of particular interest?** We can check whether a hypothesized value for the population parameter falls within the confidence interval. If so, then our results are consistent with this hypothesized value. If not, then it is unlikely (for a 95% confidence interval, the chance is at most 5%) that the parameter has this value.

[1] Diem, K. (1970) *Documenta Geigy Scientific Tables*, 7th Edn. Blackwell Publishing: Oxford.

Degrees of freedom

You will come across the term 'degrees of freedom' in statistics. In general they can be calculated as the sample size minus the number of constraints in a particular calculation; these constraints may be the parameters that have to be estimated. As a simple illustration, consider a set of three numbers which add up to a particular total (T). Two of the numbers are 'free' to take any value but the remaining number is fixed by the constraint imposed by T. Therefore the numbers have two degrees of freedom. Similarly, the degrees of freedom of the sample variance, $s^2 = \dfrac{\sum(x-\bar{x})^2}{n-1}$ (Chapter 6), are the sample size minus one, because we have to calculate the sample mean ($\bar{x}$), an estimate of the population mean, in order to evaluate s^2.

Bootstrapping

Bootstrapping is a computer intensive simulation process which we can use to derive a confidence interval for a parameter if we do not want to make assumptions about the sampling distribution of its estimate (e.g. the Normal distribution for the sample mean). From the original sample, we create a large number of random samples (usually at least 1000), each of the same size as the original sample, by sampling with replacement, i.e. by allowing an individual who has been selected to be 'replaced' so that, potentially, this individual can be included more than once in a given sample. Every sample provides an estimate of the parameter, and we use the variability of the distribution of these estimates to obtain a confidence interval for the parameter, for example, by considering relevant percentiles (e.g. the 2.5th and 97.5th percentiles to provide a 95% confidence interval).

Example

Confidence interval for the mean

We are interested in determining the mean age at first birth in women who have bleeding disorders. In a sample of 49 such women (Chapter 2):

Mean age at birth of child, $\bar{x} = 27.01$ years
Standard deviation, $s = 5.1282$ years

$$\text{Standard error, SEM} = \frac{5.1282}{\sqrt{49}} = 0.7326 \text{ years}$$

The variable is approximately Normally distributed but, because the population variance is unknown, we use the t-distribution to calculate the confidence interval. The 95% confidence interval for the mean is:

$$27.01 \pm (2.011 \times 0.7326) = (25.54, 28.48) \text{ years}$$

where 2.011 is the percentage point of the t-distribution with (49 − 1) = 48 degrees of freedom giving a two-tailed probability of 0.05 (Appendix A2).

We are 95% certain that the true mean age at first birth in women with bleeding disorders in the population lies between 25.54 and 28.48 years. This range is fairly narrow, reflecting a precise estimate. In the general population, the mean age at first birth in 1997 was 26.8 years. As 26.8 falls into our confidence interval, there is no evidence that women with bleeding disorders tend to give birth at an older age than other women.

Note that the 99% confidence interval (25.05, 28.97 years), is slightly wider than the 95% CI, reflecting our increased confidence that the population mean lies in the interval.

Confidence interval for the proportion

Of the 64 women included in the study, 27 (42.2%) reported that they experienced bleeding gums at least once a week. This is a relatively high percentage, and may provide a way of identifying undiagnosed women with bleeding disorders in the general population. We calculate a 95% confidence interval for the proportion with bleeding gums in the population.

Sample proportion = 27/64 = 0.422

$$\text{Standard error of proportion} = \sqrt{\frac{0.422(1-0.422)}{64}} = 0.0617$$

$$95\% \text{ confidence interval} = 0.422 \pm (1.96 \times 0.0617)$$
$$= (0.301, 0.543)$$

We are 95% certain that the true percentage of women with bleeding disorders in the population who experience bleeding gums this frequently lies between 30.1% and 54.3%. This is a fairly wide confidence interval, suggesting poor precision; a larger sample size would enable us to obtain a more precise estimate. However, the upper and lower limits of this confidence interval both indicate that a substantial percentage of these women are likely to experience bleeding gums. We would need to obtain an estimate of the frequency of this complaint in the general population before drawing any conclusions about its value for identifying undiagnosed women with bleeding disorders.

12 Study design I

Study design is vitally important as poorly designed studies may give misleading results. Large amounts of data from a poor study will not compensate for problems in its design. In this chapter and in Chapter 13 we discuss some of the main aspects of study design. In Chapters 14–16 we discuss specific types of study: clinical trials, cohort studies and case–control studies.

The aims of any study should be clearly stated at the outset. We may wish to estimate a parameter in the population (such as the risk of some event (Chapter 15)), to consider associations between a particular aetiological factor and an outcome of interest, or to evaluate the effect of an intervention (such as a new treatment). There may be a number of possible designs for any such study. The ultimate choice of design will depend not only on the aims, but on the resources available and ethical considerations (see Table 12.1).

Experimental or observational studies

- **Experimental** studies involve the investigator intervening in some way to affect the outcome. The clinical trial (Chapter 14) is an example of an experimental study in which the investigator introduces some form of treatment. Other examples include animal studies or laboratory studies that are carried out under experimental conditions. Experimental studies provide the most convincing evidence for any hypothesis as it is generally possible to control for factors that may affect the outcome. However, these studies are not always feasible or, if they involve humans or animals, may be unethical.
- **Observational** studies, for example cohort (Chapter 15) or case–control (Chapter 16) studies, are those in which the investigator does nothing to affect the outcome, but simply observes what happens. These studies may provide poorer information than experimental studies because it is often impossible to control for all factors that affect the outcome. However, in some situations, they may be the only types of study that are helpful or possible. **Epidemiological studies**, which assess the relationship between factors of interest and disease in the population, are observational.

Table 12.1 Study designs.

Type of study	Timing	Form	Action in past time	Action in present time (starting point)	Action in future time	Typical uses
Cross-sectional	Cross-sectional	Observational		Collect all information		• Prevalence estimates • Reference ranges and diagnostic tests • Current health status of a group
Repeated cross-sectional	Cross-sectional	Observational		Collect all information	Collect all information; Collect all information	• Changes over time
Cohort (Chapter 15)	Longitudinal (prospective)	Observational		Define cohort and assess risk factors	follow → Observe outcomes	• Prognosis and natural history (what will happen to someone with disease) • Aetiology
Case–control (Chapter 16)	Longitudinal (retrospective)	Observational	Assess risk factors	trace ← Define cases and controls (i.e. outcome)		• Aetiology (particularly for rare diseases)
Experiment	Longitudinal (prospective)	Experimental		Apply intervention	follow → Observe outcomes	• Clinical trial to assess therapy (Chapter 14) • Trial to assess preventative measure, e.g. large scale vaccine trial • Laboratory experiment

Assessing causality in observational studies

Although the most convincing evidence for the causal role of a factor in disease usually comes from experimental studies, information from observational studies may be used provided it meets a number of criteria. The most well known criteria for assessing causation were proposed by Hill[1].

- The cause must precede the effect.
- The association should be plausible, i.e. the results should be biologically sensible.
- There should be consistent results from a number of studies.
- The association between the cause and the effect should be strong.
- There should be a dose–response relationship with the effect, i.e. higher levels of the effect should lead to more severe disease or more rapid disease onset.
- Removing the factor of interest should reduce the risk of disease.

Cross-sectional or longitudinal studies

- **Cross-sectional** studies are carried out at a single point in time. Examples include surveys and censuses of the population. They are particularly suitable for estimating the **point prevalence** of a condition in the population.

$$\text{Point prevalence} = \frac{\text{Number with the disease at a single time point}}{\text{Total number studied at the same time point}}$$

As we do not know when the events occurred prior to the study, we can only say that there is an association between the factor of interest and disease, and not that the factor is likely to have *caused* disease. Furthermore, we cannot estimate the **incidence** of the disease, i.e. the rate of new events in a particular period (Chapter 31). In addition, because cross-sectional studies are only carried out at one point in time, we cannot consider trends over time. However, these studies are generally quick and cheap to perform.

- **Repeated cross-sectional** studies may be carried out at different time points to assess trends over time. However, as these studies involve different groups of individuals at each time point, it can be difficult to assess whether apparent changes over time simply reflect differences in the groups of individuals studied.
- **Longitudinal studies** follow a sample of individuals over time. They are usually **prospective** in that individuals are followed forwards from some point in time (Chapter 15). Sometimes **retrospective** studies, in which individuals are selected and factors that have occurred in their past are identified (Chapter 16), are also perceived as longitudinal. Longitudinal studies generally take longer to carry out than cross-sectional studies, thus requiring more resources, and, if they rely on patient memory or medical records, may be subject to bias (explained at the end of this chapter).

Experimental studies are generally prospective as they consider the impact of an intervention on an outcome that will happen in the future. However, observational studies may be either prospective or retrospective.

Controls

The use of a comparison group, or **control group**, is essential when designing a study and interpreting any research findings. For example, when assessing the causal role of a particular factor for a disease, the risk of disease should be considered both in those who are exposed and in those who are unexposed to the factor of interest (Chapters 15 and 16). See also 'Treatment comparisons' in Chapter 14.

Bias

When there is a systematic difference between the results from a study and the true state of affairs, bias is said to have occurred. Types of bias include:

- **Observer bias** — one observer consistently under- or over-reports a particular variable;
- **Confounding bias** — where a spurious association arises due to a failure to adjust fully for factors related to both the risk factor and outcome (see Chapter 34);
- **Selection bias** — patients selected for inclusion into a study are not representative of the population to which the results will be applied;
- **Information bias** — measurements are incorrectly recorded in a systematic manner; and
- **Publication bias** — a tendency to publish only those papers that report positive or topical results.

Other biases may, for example, be due to **recall** (Chapter 16), **healthy entrant effect** (Chapter 15), **assessment** (Chapter 14) and **allocation** (Chapter 14).

[1] Hill, AB. (1965) The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, **58**, 295.

13 Study design II

Variation

Variation in data may be caused by known factors, measurement 'errors', or may be **unexplainable random variation**. We measure the impact of variation in the data on the estimation of a population parameter by using the standard error (Chapter 10). When the measurement of a variable is subject to considerable variation, estimates relating to that variable will be imprecise, with large standard errors. Clearly, it is desirable to reduce the impact of variation as far as possible, and thereby increase the precision of our estimates. There are various ways in which we can do this.

Replication

Our estimates are more precise if we take replicates (e.g. two or three measurements of a given variable for every individual on each occasion). However, as replicate measurements are not independent, we must take care when analysing these data. A simple approach is to use the mean of each set of replicates in the analysis in place of the original measurements. Alternatively, we can use methods that specifically deal with replicated measurements (see Chapters 41 and 42).

Sample size

The choice of an appropriate size for a study is a crucial aspect of study design. With an increased sample size, the standard error of an estimate will be reduced, leading to increased precision and study power (Chapter 18). Sample size calculations (Chapter 36) should be carried out before starting the study.

Particular study designs

Modifications of simple study designs can lead to more precise estimates. Essentially we are comparing the effect of one or more '**treatments**' on **experimental units**. The experimental unit is the smallest group of 'individuals' which can be regarded as independent for the purposes of analysis, for example, an individual patient, volume of blood or skin patch. If experimental units are assigned randomly (i.e. by chance) to treatments (Chapter 14) and there are no other refinements to the design, then we have a **complete randomized design**. Although this design is straightforward to analyse, it is inefficient if there is substantial variation between the experimental units. In this situation, we can incorporate **blocking** and/or use a **cross-over design** to reduce the impact of this variation.

Blocking

It is often possible to group experimental units that share similar characteristics into a homogeneous **block** or **stratum** (e.g. the blocks may represent different age groups). The variation between units in a block is less than that between units in different blocks. The individuals within each block are randomly assigned to treatments; we compare treatments within each block rather than making an overall comparison between the individuals in different blocks. We can therefore assess the effects of treatment more precisely than if there was no blocking.

Parallel versus cross-over designs (Fig. 13.1)

Generally, we make comparisons between individuals in different groups. For example, most clinical trials (Chapter 14) are **parallel** trials, in which each patient receives one of the two (or occasionally more) treatments that are being compared, i.e. they result in *between-individual* comparisons.

Because there is usually less variation in a measurement within an individual than between different individuals (Chapter 6), in some situations it may be preferable to consider using each individual as his/her own control. These *within-individual* comparisons provide more precise comparisons than those from between-individual designs, and fewer individuals are required for the study to achieve the same level of precision. In a clinical trial setting, the **cross-over** design[1] is an example of a within-individual comparison; if there are two treatments, every individual gets each treatment, one after the other in a random order to eliminate any effect of calendar time. The treatment periods are separated by a **washout period**, which allows any residual effects (**carry-over**) of the previous treatment to dissipate. We analyse the difference in the responses on the two treatments for each individual. This design can only be used when the treatment temporarily alleviates symptoms rather than provides a cure, and the response time is not prolonged.

Factorial experiments

When we are interested in more than one factor, separate studies that assess the effect of varying one factor at a time may be inefficient and costly. **Factorial designs** allow the simultaneous analysis of any number of **factors** of interest. The simplest design, a 2×2 factorial experiment, considers two factors (for example, two different treatments), each at two **levels** (e.g. either active or inactive treatment). As an example, consider the US Physicians Health study[2], designed to assess the importance of aspirin and beta carotene in preventing heart disease and cancer. A 2×2 factorial design was used with the two factors being the different compounds and the two levels of each indicating whether the physician received the active compound or its placebo (see Chapter 14). Table 13.1 shows the possible treatment combinations.

We assess the effect of the level of beta carotene by comparing patients in the left-hand column to those in the right-hand column. Similarly, we assess the effect of the level of aspirin by comparing patients in the top row with those in the bottom row. In addition, we can test whether the two factors are **interactive**, i.e. when the effect of the level of beta carotene is different for the two levels of aspirin.

Table 13.1 Active treatment combinations.

	Beta carotene	
Aspirin	No	Yes
No	None	Beta carotene
Yes	Aspirin	Aspirin + beta carotene

[1] Senn, S. (1993) *Cross-over Trials in Clinical Research*. Wiley, Chichester.
[2] Steering Committee of the Physician's Health Study Research Group. (1989) Final report of the aspirin component of the on-going Physicians Health Study. *New England Journal of Medicine*, **321**, 129–135.

If the effects differ, we then say that there is an **interaction** between the two factors (Chapter 34). In this example, an interaction would suggest that the combination of aspirin and beta carotene together is more (or less) effective than would be expected by simply adding the separate effects of each drug. This design, therefore, provides additional information to two separate studies and is a more efficient use of resources, requiring a smaller sample size to obtain estimates with a given degree of precision.

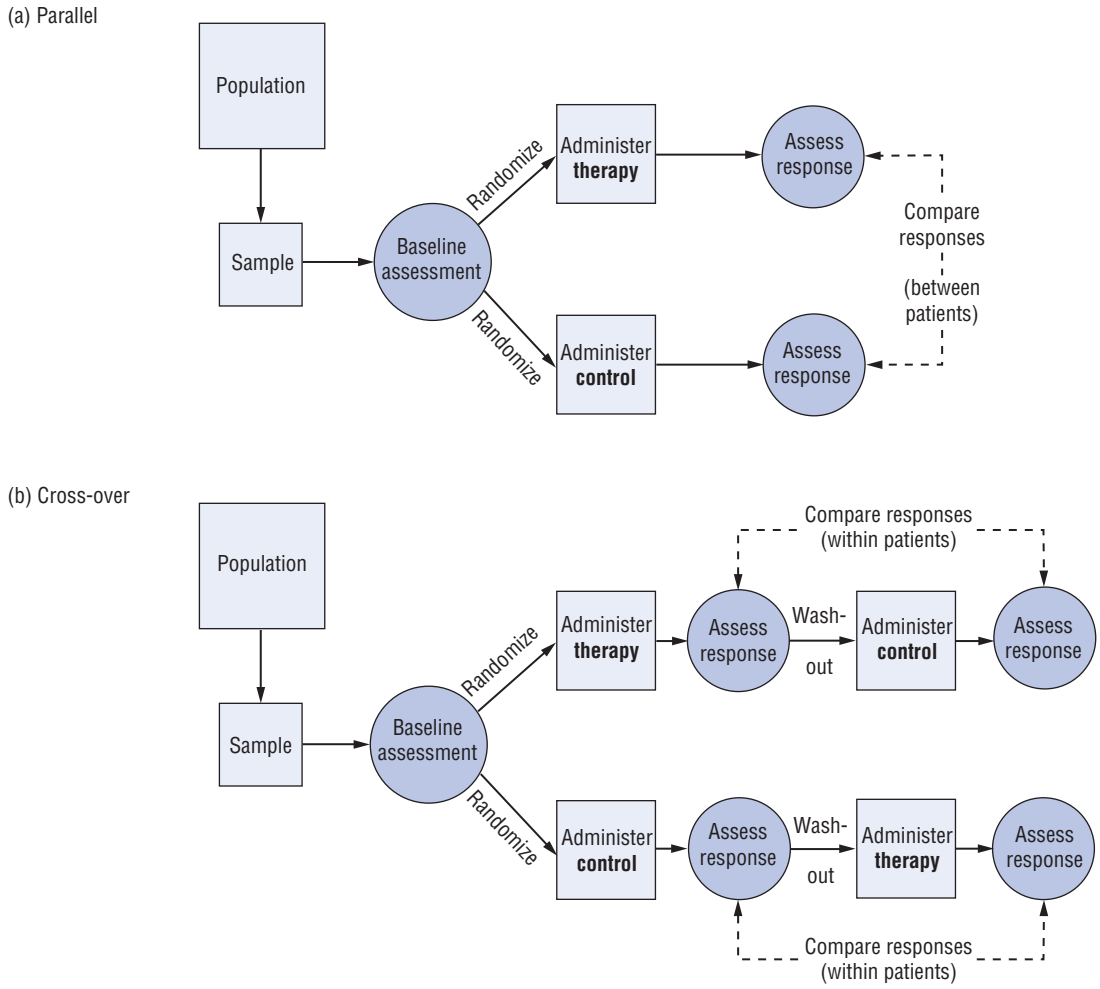


Figure 13.1 (a) Parallel, and (b) Cross-over designs.

14 Clinical trials

A **clinical trial**[1] is any form of planned experimental study designed, in general, to evaluate the effect of a new treatment on a clinical outcome in humans. Clinical trials may either be pre-clinical studies, small clinical studies to investigate effect and safety (**Phase I/II** trials), or full evaluations of the new treatment (**Phase III** trials). In this chapter we discuss the main aspects of Phase III trials, all of which should be reported in any publication (see CONSORT statement checklist in Table 14.1, and Figs 14.1 & 14.2).

Treatment comparisons

Clinical trials are prospective studies in that we are interested in measuring the impact of a treatment given now on a future possible outcome. In general, clinical trials evaluate new interventions (e.g. type or dose of drug, or surgical procedure). Throughout this chapter we assume, for simplicity, that only one *new* treatment is being evaluated in a trial.

An important feature of a clinical trial is that it should be comparative (Chapter 12). Without a **control** treatment, it is impossible to be sure that any response is due solely to the effect of the treatment, and the importance of the new treatment can be over-stated. The control may be the standard treatment (a **positive control**) or, if one does not exist, may be a **negative control**, which can be a **placebo** (a treatment which looks and tastes like the new drug but which does not contain any active compound) or the absence of treatment if ethical considerations permit.

Endpoints

We must decide in advance which outcome most accurately reflects the benefit of the new therapy. This is known as the **primary endpoint** of the study and usually relates to treatment **efficacy**. **Secondary endpoints**, which often relate to toxicity, are of interest and should also be considered at the outset. Generally, all these endpoints are analysed at the end of the study. However, we may wish to carry out some preplanned **interim analyses** (for example, to ensure that no major toxicities have occurred requiring the trial to be stopped). Care should be taken when comparing treatments at these times due to the problems of multiple hypothesis testing (Chapter 18).

Treatment allocation

Once a patient has been formally entered into a clinical trial, s/he is allocated to a treatment group. In general, patients are allocated in a random manner (i.e. based on chance), using a process known as **random allocation** or **randomization**. This is often performed using a computer-generated list of random numbers or by using a table of random numbers (Appendix A12). For example, to allocate patients to two treatments, we might follow a sequence of random numbers, and allocate the patient to treatment A if the number is even (treating zero as even) and to treatment B if it is odd. This process promotes similarity between the treatment groups in terms of baseline characteristics at entry to the trial (i.e. it avoids **allocation bias** and, consequently, **confounding** (Chapters 12 and 34)), maximizing the efficiency of the trial. If a baseline characteristic is not evenly distributed in the treatment groups (evaluated by examining the appropriate summary measures, e.g. the means and standard deviations), the discrepancy must be due to chance if randomization has been used. Therefore, it is inappropriate to perform a formal statistical hypothesis test (e.g. the *t*-test, Chapter 21) to compare the parameters of any baseline characteristic in the treatment groups because the hypothesis test assesses whether the difference between the groups is due to chance.

Trials in which patients are randomized to receive either the new treatment or a control treatment are known as **randomized controlled trials** (often referred to as **RCTs**), and are regarded as optimal.

Further refinements of randomization, including **stratified randomization** (which controls for the effects of important factors), and **blocked randomization** (which ensures roughly equal sized treatment groups) exist. **Systematic allocation**, whereby patients are allocated to treatment groups systematically, possibly by day of visit or date of birth, should be avoided where possible; the clinician may be able to determine the proposed treatment for a particular patient before s/he is entered into the trial, and this may influence his/her decision as to whether to include a patient in the trial. Sometimes we use a process known as **cluster randomization**, whereby we randomly allocate *groups* of individuals (e.g. all people registered at a single general practice) to treatments rather than each individual. We should take care when planning the size of the study and analysing the data in such studies (see also Chapters, 36, 41 and 42)[2].

Blinding or masking

There may be **assessment bias** when patients and/or clinicians are aware of the treatment allocation, particularly if the response is subjective. An awareness of the treatment allocation may influence the recording of signs of improvement or adverse events. Therefore, where possible, all participants (clinicians, patients, assessors) in a trial should be **blinded** or **masked** to the treatment allocation. A trial in which the patient, the treatment team and the assessor are unaware of the treatment allocation is a **double-blind trial**. Trials in which it is impossible to blind the patient may be **single-blind** providing the clinician and/or assessor is blind to the treatment allocation.

Patient issues

As clinical trials involve humans, patient issues are of importance. In particular, any clinical trial must be passed by an **ethical committee** who judge that the trial does not contravene the Declaration of Helsinki. **Informed patient consent** must be obtained from each patient (or from the legal guardian or parent if the patient is a minor) before s/he is entered into a trial.

The protocol

Before any clinical trial is carried out, a written description of all aspects of the trial, known as the **protocol**, should be prepared. This

[1] Pocock, S.J. (1983) *Clinical Trials: A Practical Approach*. Wiley, Chichester.

[2] Kerry, S.M. and Bland, J.M. (1998) Sample size in cluster randomisation. *British Medical Journal*, **316**, 549.

includes information on the aims and objectives of the trial, along with a definition of which patients are to be recruited (**inclusion** and **exclusion criteria**), treatment schedules, data collection and analysis, contingency plans should problems arise, and study personnel. It is important to recruit enough patients into a trial so that the chance of correctly detecting a true treatment effect is sufficiently high. Therefore, before carrying out any clinical trial, the optimal **trial size** should be calculated (Chapter 36).

Protocol deviations are patients who enter the trial but do not fulfil the protocol criteria, e.g. patients who were incorrectly

Table 14.1 Checklist of items from the CONSORT statement (Consolidation of Standards for Reporting Trials) to include when reporting a randomized trial (www.consort-statement.org).

PAPER SECTION and topic	Item	Description	Reported on page #
TITLE & ABSTRACT	1	How participants were allocated to interventions (*e.g.*, 'random allocation', 'randomized', or 'randomly assigned').	
INTRODUCTION Background	2	Scientific background and explanation of rationale.	
METHODS Participants	3	Eligibility criteria for participants and the settings and locations where the data were collected.	
Interventions	4	Precise details of the interventions intended for each group and how and when they were actually administered.	
Objectives	5	Specific objectives and hypotheses.	
Outcomes	6	Clearly defined primary and secondary outcome measures and, when applicable, any methods used to enhance the quality of measurements (*e.g.*, multiple observations, training of assessors).	
Sample size	7	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules.	
Randomization — Sequence generation	8	Method used to generate the random allocation sequence, including details of any restriction (*e.g.*, blocking, stratification).	
Randomization — Allocation concealment	9	Method used to implement the random allocation sequence (*e.g.*, numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned.	
Randomization — Implementation	10	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups.	
Blinding (masking)	11	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. When relevant, how the success of blinding was evaluated.	
Statistical methods	12	Statistical methods used to compare groups for primary outcome(s); Methods for additional analyses, such as subgroup analyses and adjusted analyses.	
RESULTS Participant flow	13	Flow of participants through each stage (a diagram is strongly recommended; see Fig 14.1). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analyzed for the primary outcome. Describe protocol deviations from study as planned, together with reasons.	
Recruitment	14	Dates defining the periods of recruitment and follow-up.	
Baseline data	15	Baseline demographic and clinical characteristics of each group.	
Numbers analyzed	16	Number of participants (denominator) in each group included in each analysis and whether the analysis was by 'intention-to-treat' . State the results in absolute numbers when feasible (*e.g.*, 10/20, not 50%).	
Outcomes and estimation	17	For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (*e.g.*, 95% confidence interval).	
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those pre-specified and those exploratory.	
Adverse events	19	All important adverse events or side effects in each intervention group.	
DISCUSSION Interpretation	20	Interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes.	
Generalizability	21	Generalizability (external validity) of the trial findings.	
Overall evidence	22	General interpretation of the results in the context of current evidence.	

recruited into or who withdrew from the study, and patients who switched treatments. To avoid bias, the study should be analysed on an **intention-to-treat** basis, in which all patients on whom we have information are analysed in the groups to which they were originally allocated, irrespective of whether they followed the treatment regime. Where possible, attempts should be made to collect information on patients who withdraw from the trial. **On-treatment** analyses, in which patients are only included in the analysis if they complete a *full* course of treatment, are not recommended as they often lead to biased treatment comparisons.

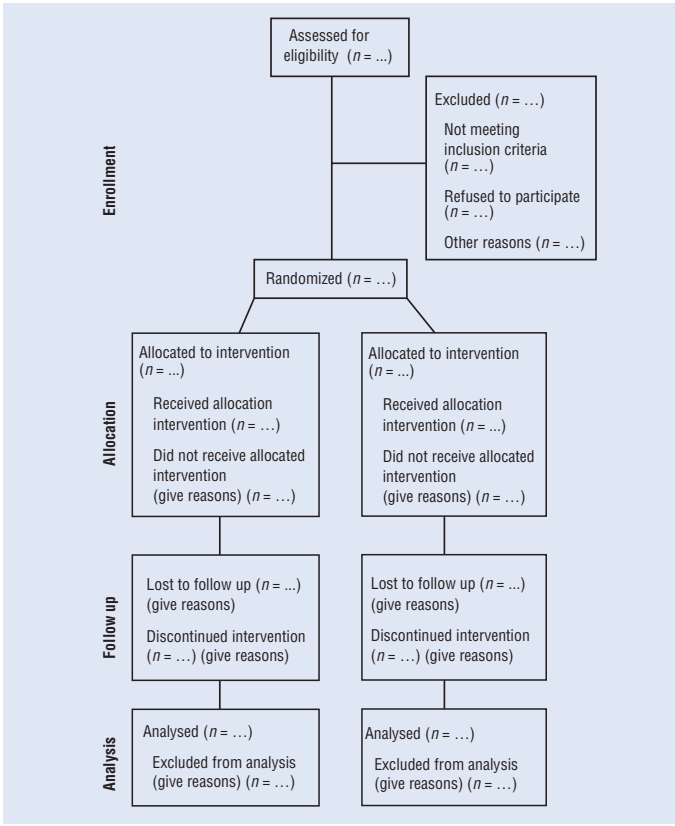


Figure 14.1 The CONSORT statement's trial profile of the Randomized Controlled Trial's progress (www.consort-statement.org).

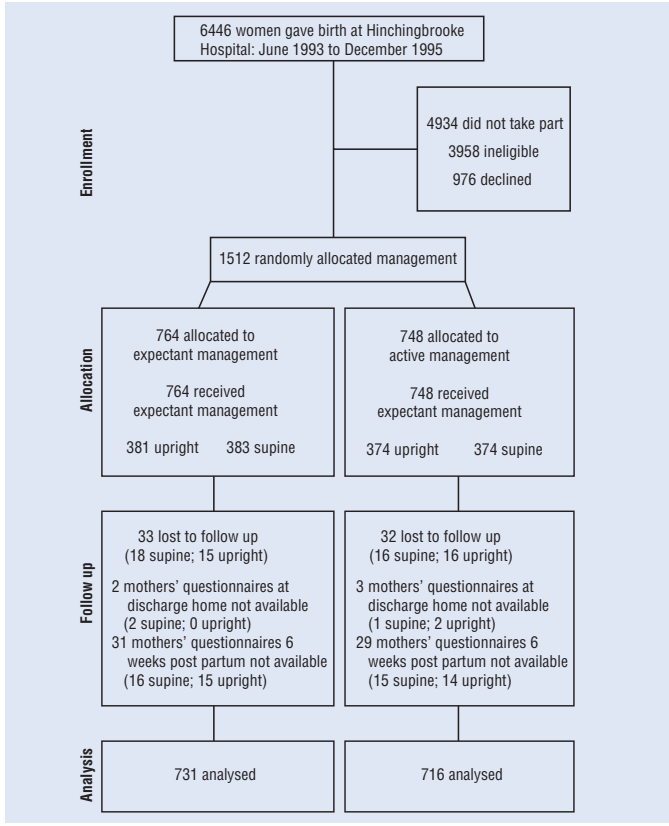


Figure 14.2 Trial profile example (adapted from trial described in Chapter 40 with permission).

15 Cohort studies

A cohort study takes a group of individuals and usually follows them forward in time, the aim being to study whether exposure to a particular aetiological factor will affect the incidence of a disease outcome in the future (Fig. 15.1). If so, the factor is generally known as a **risk factor** for the disease outcome. For example, a number of cohort studies have investigated the relationship between dietary factors and cancer. Although most cohort studies are prospective, **historical** cohorts can be investigated, the information being obtained retrospectively. However, the quality of historical studies is often dependent on medical records and memory, and they may therefore be subject to bias.

Cohort studies can either be fixed or dynamic. If individuals leave a **fixed** cohort, they are not replaced. In **dynamic** cohorts, individuals may drop out of the cohort, and new individuals may join as they become eligible.

Selection of cohort

The cohort should be representative of the population to which the results will be generalized. It is often advantageous if the individuals can be recruited from a similar source, such as a particular occupational group (e.g. civil servants, medical practitioners) as information on mortality and morbidity can be easily obtained from records held at the place of work, and individuals can be re-contacted when necessary. However, such a cohort may not be truly representative of the general population, and may be healthier. Cohorts can also be recruited from GP lists, ensuring that a group of individuals with different health states is included in the study. However, these patients tend to be of similar social backgrounds because they live in the same area.

When trying to assess the aetiological effect of a risk factor, individuals recruited to cohorts should be disease-free at the start of the study. This is to ensure that any exposure to the risk factor occurs before the outcome, thus enabling a causal role for the factor to be postulated. Because individuals are disease-free at the start of the study, we often see a **healthy entrant** effect. Mortality rates in the first period of the study are then often lower than would be expected in the general population. This will be apparent when mortality rates start to increase suddenly a few years into the study.

Follow-up of individuals

When following individuals over time, there is always the problem that they may be **lost to follow-up**. Individuals may move without leaving a forwarding address, or they may decide that they wish to leave the study. The benefits of a cohort study are reduced if a large number of individuals is lost to follow-up. We should thus find ways to minimize these drop-outs, e.g. by maintaining regular contact with the individuals.

Information on outcomes and exposures

It is important to obtain full and accurate information on disease outcomes, e.g. mortality and illness from different causes. This may entail searching through disease registries, mortality statistics, GP and hospital records.

Exposure to the risks of interest may change over the study period. For example, when assessing the relationship between alcohol consumption and heart disease, an individual's typical alcohol consumption is likely to change over time. Therefore it is important to re-interview individuals in the study on repeated occasions to study changes in exposure over time.

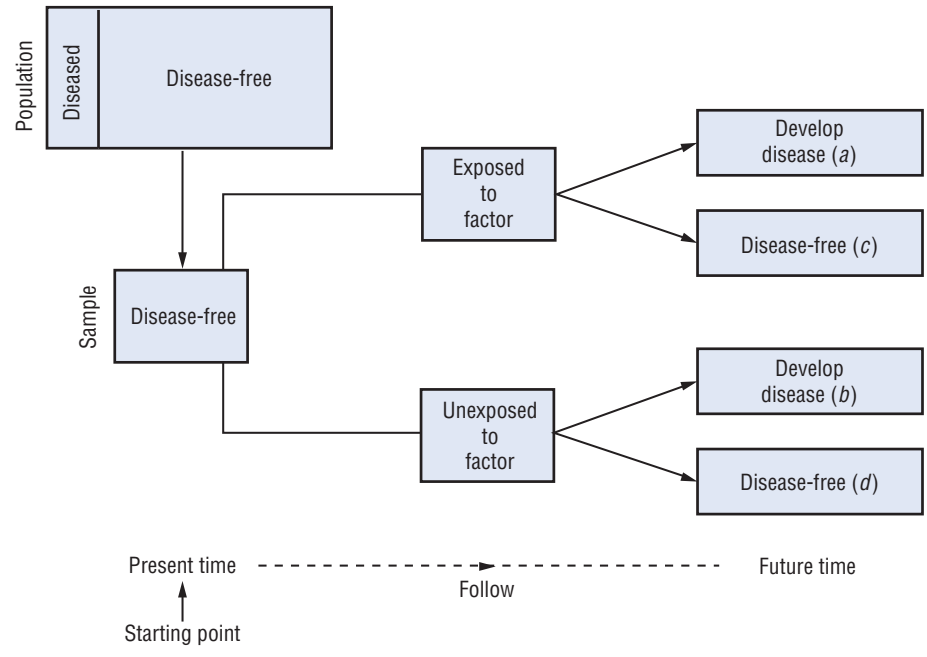


Figure 15.1 Diagrammatic representation of a cohort study (frequencies in parenthesis, see Table 15.1).

Analysis of cohort studies

Table 15.1 contains observed frequencies.

Table 15.1. Observed frequencies (see Fig. 15.1)

	Exposed to factor		
	Yes	No	Total
Disease of interest			
Yes	a	b	$a+b$
No	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

Because patients are followed longitudinally over time, it is possible to estimate the **risk** of developing the disease in the population, by calculating the risk in the sample studied.

$$\text{Estimated risk of disease} = \frac{\text{Number developing disease over study period}}{\text{Total number in the cohort}} = \frac{a+b}{n}$$

The risk of disease in the individuals exposed and unexposed to the factor of interest in the population can be estimated in the same way.

Estimated risk of disease in the exposed group,

$\text{risk}_{\text{exp}} = a/(a+c)$

Estimated risk of disease in the unexposed group,

$\text{risk}_{\text{unexp}} = b/(b+d)$

$$\text{Then, } \textbf{estimated relative risk} = \frac{\text{risk}_{\text{exp}}}{\text{risk}_{\text{un exp}}} = \frac{a/(a+c)}{b/(b+d)}$$

The **relative risk** (RR) indicates the increased (or decreased) risk of disease associated with exposure to the factor of interest. A relative risk of one indicates that the risk is the same in the exposed and unexposed groups. A relative risk greater than one indicates that there is an increased risk in the exposed group compared with the unexposed group; a relative risk less than one indicates a reduction in the risk of disease in the exposed group. For example, a relative risk of 2 would indicate that individuals in the exposed group had twice the risk of disease of those in the unexposed group.

A relative risk should always be interpreted alongside the underlying risk of the disease. Even a large relative risk may have limited clinical implications when the underlying risk of disease is very small.

A confidence interval for the relative risk should be calculated, and we can test whether the relative risk is equal to one. These are easily performed on a computer and therefore we omit details.

Advantages of cohort studies

- The time sequence of events can be assessed.
- They can provide information on a wide range of outcomes.
- It is possible to measure the incidence/risk of disease directly.
- It is possible to collect very detailed information on exposure to a wide range of factors.
- It is possible to study exposure to factors that are rare.
- Exposure can be measured at a number of time points, so that changes in exposure over time can be studied.
- There is reduced **recall** and **selection bias** compared with case–control studies (Chapter 16).

Disadvantages of cohort studies

- In general, cohort studies follow individuals for long periods of time, and are therefore costly to perform.
- Where the outcome of interest is rare, a very large sample size is needed.
- As follow-up increases, there is often increased loss of patients as they migrate or leave the study, leading to biased results.
- As a consequence of the long time-scale, it is often difficult to maintain consistency of measurements and outcomes over time. Furthermore, individuals may modify their behaviour after an initial interview.
- It is possible that disease outcomes and their probabilities, or the aetiology of disease itself, may change over time.

Clinical cohorts

Sometimes we select a cohort of patients with the same clinical condition attending one or more hospitals and follow them (either as in-patients or out-patients) to see how many patients experience a resolution (in the case of a positive outcome of the condition) or some indication of disease progression such as death or relapse. The information we collect on each patient is usually that which is collected as part of routine clinical care. The aims of **clinical cohorts** or **observational databases** may include describing the outcomes of individuals with the condition and assessing the effects of different approaches to treatment (e.g. different drugs or different treatment modalities). In contrast to randomized controlled trials (Chapter 14), which often include a highly selective sample of individuals who are willing to participate in the trial, clinical cohorts often include all patients with the condition at the hospitals in the study. Thus, outcomes from these cohorts are thought to more accurately reflect the outcomes that would be seen in clinical practice. However, as allocation to treatment in these studies is not randomized (Chapter 14), clinical cohorts are particularly prone to confounding bias (Chapters 12 and 34).

Example

The British Regional Heart Study[1] is a large cohort study of 7735 men aged 40–59 years randomly selected from general practices in 24 British towns, with the aim of identifying risk factors for ischaemic heart disease. At recruitment to the study, the men were asked about a number of demographic and lifestyle factors, including information on cigarette smoking habits. Of the 7718 men who provided information on smoking status, 5899 (76.4%) had smoked at some stage during their lives (including those who were current smokers and those who were ex-smokers). Over the subsequent 10 years, 650 of these 7718 men (8.4%) had a myocardial infarction (MI). The results, displayed in the table, show the number (and percentage) of smokers and non-smokers who developed and did not develop a MI over the 10 year period.

	MI in subsequent 10 years		
Smoking status at baseline	Yes	No	Total
Ever smoked	563 (9.5%)	5336 (90.5%)	5899
Never smoked	87 (4.8%)	1732 (95.2%)	1819
Total	650 (8.4%)	7068 (71.6%)	7718

$$\text{The estimated relative risk} = \frac{(563/5899)}{(87/1819)} = 2.00.$$

It can be shown that the 95% confidence interval for the true relative risk is (1.60, 2.49).

We can interpret the relative risk to mean that a middle-aged man who has *ever* smoked is twice as likely to suffer a MI over the next 10 year period as a man who has *never* smoked. Alternatively, the risk of suffering a MI for a man who has ever smoked is 100% greater than that of a man who has never smoked.

[1] Data kindly provided by Dr F.C. Lampe, Ms M. Walker and Dr P. Whincup, Department of Primary Care and Population Sciences, Royal Free and University College Medical School, London, UK.

16 Case–control studies

A case–control study compares the characteristics of a group of patients with a particular disease outcome (the **cases**) to a group of individuals without a disease outcome (the **controls**), to see whether any factors occurred more or less frequently in the cases than the controls (Fig. 16.1). Such retrospective studies do not provide information on the prevalence or incidence of disease but may give clues as to which factors elevate or reduce the risk of disease.

Selection of cases

It is important to define whether **incident** cases (patients who are recruited at the time of diagnosis) or **prevalent** cases (patients who were already diagnosed before entering the study) should be recruited. Prevalent cases may have had time to reflect on their history of exposure to risk factors, especially if the disease is a well-publicized one such as cancer, and may have altered their behaviour after diagnosis. It is important to identify as many cases as possible so that the results carry more weight and the conclusions can be generalized to future populations. To this end, it may be necessary to access hospital lists and disease registries, and to include cases who died during the time period when cases and controls were defined, because their exclusion may lead to a biased sample of cases.

Selection of controls

Controls should be screened at entry to the study to ensure that they do not have the disease of interest. Sometimes there may be more than one control for each case. Where possible, controls should be selected from the same source as cases. Controls are often selected from hospitals. However, as risk factors related to one disease outcome may also be related to other disease outcomes, the selection of hospital-based controls may over-select individuals who have been exposed to the risk factor of interest, and may, therefore, not always be appropriate. It is often acceptable to select controls from the general population, although they may not be as motivated to take part in such a study, and response rates may therefore be poorer in controls than cases. The use of neighbourhood controls may ensure that cases and controls are from similar social backgrounds.

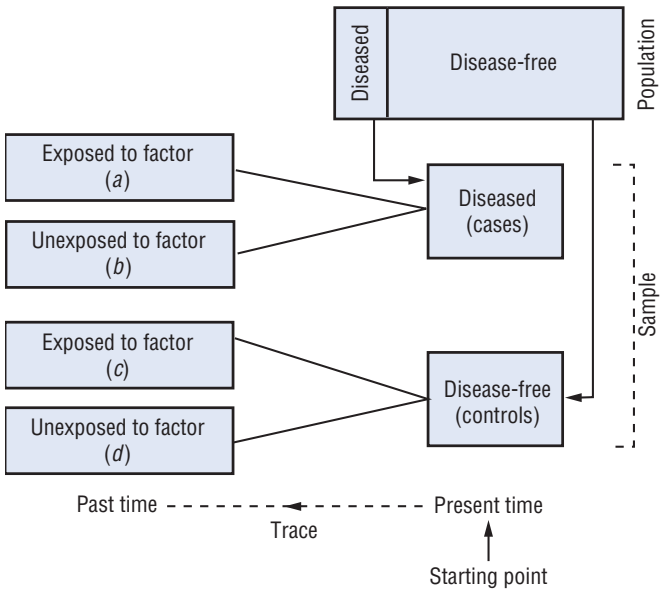


Figure 16.1 Diagrammatic representation of a case–control study.

Matching

Many case–control studies are **matched** in order to select cases and controls who are as similar as possible. In general, it is useful to sex-match individuals (i.e. if the case is male, the control should also be male), and, sometimes, patients will be age-matched. However, it is important not to match on the basis of the risk factor of interest, or on any factor that falls on the causal pathway of the disease (Chapter 34), as this will remove the ability of the study to assess any relationship between the risk factor and the disease. Unfortunately, matching does mean that the effect on disease of the variables that have been used for matching cannot be studied.

Analysis of unmatched case–control studies

Table 16.1 shows observed frequencies. Because patients are selected on the basis of their disease status, it is not possible to estimate the absolute risk of disease. We can calculate the **odds ratio**, which is given by:

$$\text{Odds ratio} = \frac{\text{Odds of being a case in exposed group}}{\text{Odds of being a case in unexposed group}}$$

where, for example, the odds of being a case in the exposed group is equal to:

$$\frac{\text{probability of being a case in the exposed group}}{\text{probability of not being a case in the exposed group}}$$

The odds of being a case in the exposed and unexposed samples are:

$$\text{odds}_{\text{exp}} = \frac{\left(\frac{a}{a+c}\right)}{\left(\frac{c}{a+c}\right)} = \frac{a}{c} \quad \text{odds}_{\text{unexp}} = \frac{\left(\frac{b}{b+d}\right)}{\left(\frac{d}{b+d}\right)} = \frac{b}{d}$$

and therefore the **estimated odds ratio** $= \frac{a/c}{b/d} = \frac{a \times d}{b \times c}$

When a disease is rare, the odds ratio is an estimate of the relative risk, and is interpreted in a similar way, i.e. it gives an indication of the increased (or decreased) odds associated with exposure to the factor of interest. An odds ratio of one indicates that the odds is the same in the exposed and unexposed groups; an odds ratio greater than one indicates that the odds of disease is greater in the exposed group than in the unexposed group, etc. Confidence intervals and hypothesis tests can also be generated for the odds ratio.

Table 16.1 Observed frequencies (see Fig. 16.1).

	Exposed to factor		
	Yes	No	Total
Disease status			
Case	a	b	$a+b$
Control	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

Analysis of matched case–control studies

Where possible, the analysis of matched case–control studies should allow for the fact that cases and controls are linked to each other as a result of the matching. Further details of methods of analysis for matched studies can be found in Chapter 30 (see Conditional Logistic Regression) and in Breslow and Day[1].

[1] Breslow, N.E. & Day, N.E. (1980) *Statistical Methods in Cancer Research. Volume I — The Analysis of Case–control Studies*. International Agency for Cancer Research, Lyon.

Advantages of case–control studies

- They are generally relatively quick, cheap and easy to perform.
- They are particularly suitable for rare diseases.
- A wide range of risk factors can be investigated.
- There is no loss to follow-up.

Disadvantages of case–control studies

- **Recall bias**, when cases have a differential ability to remember certain details about their histories, is a potential problem. For example, a lung cancer patient may well remember the occasional period when s/he smoked, whereas a control may not remember a similar period.
- If the onset of disease preceded exposure to the risk factor, causation cannot be inferred.
- Case–control studies are not suitable when exposures to the risk factor are rare.

Example

A total of 1327 women aged 50–81 years with hip fractures, who lived in a largely urban area in Sweden, were investigated in this unmatched case–control study. They were compared with 3262 controls within the same age range randomly selected from the national register. Interest was centred on determining whether women currently taking postmenopausal hormone replacement therapy (HRT) were less likely to have hip fractures than those not taking it. The results in the table show the number of women who were current users of HRT and those who had never used or formerly used HRT in the case and control groups.

The observed odds ratio = $(40 \times 3023)/(239 \times 1287) = 0.39$.

It can be shown that the 95% confidence interval for the odds ratio is (0.28, 0.56).

Thus the odds of a hip fracture in a postmenopausal woman in this age range in Sweden who was a current user of HRT was 39% of that of a woman who had never used or formerly used HRT, i.e. being a current user of HRT reduced the odds of hip fracture by 61%.

Observed frequencies in fracture study

	Current user of HRT	Never used HRT/former user of HRT	Total
With hip fracture (cases)	40	1287	1327
Without hip fracture (controls)	239	3023	3262
Total	279	4310	4589

Data extracted from: Michaelsson, K., Baron, J.A., Farahmand, B.Y., *et al*. (1998) Hormone replacement therapy and risk of hip fracture: population based case–control study. *British Medical Journal*, **316**, 1858–1863.

17 Hypothesis testing

We often gather sample data in order to assess how much evidence there is against a specific hypothesis about the population. We use a process known as **hypothesis testing** (or **significance testing**) to quantify our belief against a particular hypothesis.

This chapter describes the format of hypothesis testing in general (Box 17.1); details of specific hypothesis tests are given in subsequent chapters. For easy reference, each hypothesis test is contained in a similarly formatted box.

Box 17.1 Hypothesis testing — a general overview

We define five stages when carrying out a hypothesis test:

1 Define the *null* and *alternative hypotheses* under study

2 Collect relevant data from a sample of individuals

3 Calculate the value of the *test statistic* specific to the null hypothesis

4 Compare the value of the test statistic to values from a known probability distribution

5 Interpret the *P-value* and results

Defining the null and alternative hypotheses

We usually test the **null hypothesis** (H_0) which assumes *no effect* (e.g. the difference in means equals zero) in the *population*. For example, if we are interested in comparing smoking rates in men and women in the population, the null hypothesis would be:

H_0: smoking rates are the same in men and women in the population

We then define the **alternative hypothesis** (H_1) which holds if the null hypothesis is not true. The alternative hypothesis relates more directly to the theory we wish to investigate. So, in the example, we might have:

H_1: the smoking rates are different in men and women in the population.

We have not specified any direction for the difference in smoking rates, i.e. we have not stated whether men have higher or lower rates than women in the population. This leads to what is known as a **two-tailed test** because we allow for either eventuality, and is recommended as we are rarely certain, *in advance*, of the direction of any difference, if one exists. In some, very rare, circumstances, we may carry out a **one-tailed test** in which a direction of effect is specified in H_1. This might apply if we are considering a disease from which all untreated individuals die (a new drug cannot make things worse) or if we are conducting a trial of equivalence or non-inferiority (see last section in this chapter).

Obtaining the test statistic

After collecting the data, we substitute values from our sample into a formula, specific to the test we are using, to determine a value for the **test statistic**. This reflects the amount of evidence in the data *against* the null hypothesis — usually, the larger the value, ignoring its sign, the greater the evidence.

Obtaining the *P*-value

All test statistics follow known theoretical probability distributions (Chapters 7 and 8). We relate the value of the test statistic obtained from the sample to the known distribution to obtain the ***P*-value**, the area in both (or occasionally one) tails of the probability distribution. Most computer packages provide the two-tailed *P*-value automatically. **The *P*-value is the probability of obtaining our results, or something more extreme, if the null hypothesis is true.** The null hypothesis relates to the population of interest, rather than the sample. Therefore, the null hypothesis is either true or false and we *cannot* interpret the *P*-value as the probability that the null hypothesis is true.

Using the *P*-value

We must make a decision about how much evidence we require to enable us to decide to reject the null hypothesis in favour of the alternative. The smaller the *P*-value, the greater the evidence against the null hypothesis.

- Conventionally, we consider that if the *P*-value is less than 0.05, there is sufficient evidence to reject the null hypothesis, as there is only a small chance of the results occurring if the null hypothesis were true. We then *reject* the null hypothesis and say that the results are **significant** at the 5% level (Fig. 17.1).
- In contrast, if the *P*-value is equal to or greater than 0.05, we usually conclude that there is insufficient evidence to reject the null hypothesis. We *do not reject* the null hypothesis, and we say that the results are **not significant** at the 5% level (Fig. 17.1). This does not mean that the null hypothesis is true; simply that we do not have enough evidence to reject it.

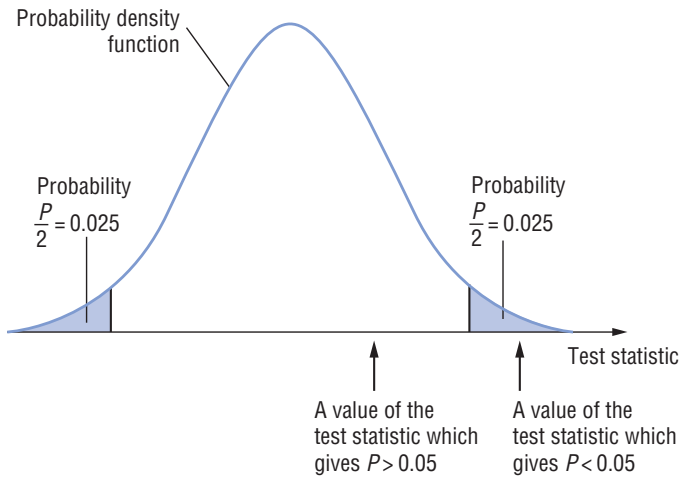


Figure 17.1 Probability distribution of the test statistic showing a two-tailed probability, $P = 0.05$.

The choice of 5% is arbitrary. On 5% of occasions we will incorrectly reject the null hypothesis when it is true. In situations in which the clinical implications of incorrectly rejecting the null hypothesis are severe, we may require stronger evidence before rejecting the null hypothesis (e.g. we may decide to reject the null hypothesis if the *P*-value is less than 0.01 or 0.001). The chosen cut-off for the *P*-value (e.g. 0.05 or 0.01) is called the **significance level** of the test; it must be chosen before the data are collected.

Quoting a result only as significant at a certain cut-off level (e.g. stating only that $P < 0.05$) can be misleading. For example, if $P = 0.04$ we would reject H_0; however, if $P = 0.06$ we would not reject it. Are these really different? Therefore, we recommend quoting the exact *P*-value, often obtained from the computer output.

Non-parametric tests

Hypothesis tests which are based on knowledge of the probability distributions that the data follow are known as **parametric tests.** Often data do not conform to the assumptions that underly these methods (Chapter 35). In these instances we can use **non-parametric tests** (sometimes referred to as **distribution-free** tests, or **rank methods**). These tests generally replace the data with their ranks (i.e. the numbers 1, 2, 3 etc., describing their position in the ordered data set) and make no assumptions about the probability distribution that the data follow.

Non-parametric tests are particularly useful when the sample size is small (so that it is impossible to assess the distribution of the data), and/or when the data are measured on a categorical scale. However, non-parametric tests are generally wasteful of information; consequently they have less power (Chapter 18) to detect a real effect than the equivalent parametric test if all the assumptions underlying the parametric test are satisfied. Furthermore, they are primarily significance tests that often do not provide estimates of the effects of interest; they lead to decisions rather than an appreciation or understanding of the data.

Which test?

Deciding which statistical test to use depends on the design of the study, the type of variable and the distribution that the data being studied follow. The flow chart on the inside front cover will aid your decision.

Hypothesis tests versus confidence intervals

Confidence intervals (Chapter 11) and hypothesis tests are closely linked. The primary aim of a hypothesis test is to make a decision and provide an exact *P*-value. A confidence interval quantifies the effect of interest (e.g. the difference in means), and enables us to assess the clinical implications of the results. However, because it provides a range of plausible values for the true effect, it can also be used to make a decision although an exact *P*-value is not provided. For example, if the hypothesized value for the effect (e.g. zero) lies outside the 95% confidence interval then we believe the hypothesized value is implausible and would reject H_0. In this instance we know that the *P*-value is less than 0.05 but do not know its exact value.

Equivalence and non-inferiority trials

In most randomized controlled trials (Chapter 14) of two or more different treatment strategies, we are usually interested in demonstrating the **superiority** of at least one treatment over the other(s). However, in some situations we may believe that a new treatment (e.g. drug) may be no more effective clinically than an existing treatment but will have other important benefits, perhaps in terms of reduced side effects, pill burden or costs. Then, we may wish to show simply that the efficacy of the new treatment is similar (in an **equivalence** trial) or not *substantially* worse (in a **non-inferiority** trial) than that of the existing treatment.

When carrying out an equivalence or non-inferiority trial, the hypothesis testing procedure used in the usual superiority trial which tests the null hypothesis that the two treatments are the same is irrelevant. This is because (1) a non-significant result does not imply non-inferiority/equivalence, and (2) even if a statistically significant effect is detected, it may be clinically unimportant. Instead, we essentially reverse the null and alternative hypotheses in an equivalence trial, so that the null hypothesis expresses a difference and the alternative hypothesis expresses equivalence.

Rather than calculating test statistics, we generally approach the problem of assessing equivalence and non-inferiority[1] by determining whether the confidence interval for the effect of interest (e.g. the difference in means between two treatment groups) lies wholly or partly within a predefined **equivalence range** (i.e. the range of values, determined by clinical experts, that corresponds to an effect of no clinical importance). If the whole of the confidence interval for the effect of interest lies within the equivalence range, then we conclude that the two treatments are equivalent; in this situation, even if the upper and lower limits of the confidence interval suggest there is benefit of one treatment over the other, it is unlikely to have any clinical importance. In a non-inferiority trial, we want to show that the new treatment is not substantially worse than the standard one (if the new treatment turns out to be better than the standard, this would be an added bonus!). In this situation, if the lower limit of the appropriate confidence interval does not fall below the lower limit of the equivalence range, then we conclude that the new treatment is not inferior.

Unless otherwise specified, the hypothesis tests in subsequent chapters are tests of superiority. Note that the methods for determining sample size described in Chapter 36 do not apply to equivalence or non-inferiority trials. The sample size required for an equivalence or non-inferiority trial[2] is generally greater than that of the comparable superiority trial if all factors that affect sample size (e.g. significance level, power) are the same.

[1] John, B., Jarvis, P., Lewis, J.A. and Ebbutt, A.F. (1996). Trials to assess equivalence: the importance of rigorous methods. *British Medical Journal* **313**; 36–39.

[2] Julious, S.A. (2004). Tutorial in Biostatistics: Sample sizes for clinical trials with Normal data. *Statistics in Medicine* **23**: 1921–1986.

18 Errors in hypothesis testing

Making a decision

Most hypothesis tests in medical statistics compare groups of people who are exposed to a variety of experiences. We may, for example, be interested in comparing the effectiveness of two forms of treatment for reducing 5 year mortality from breast cancer. For a given outcome (e.g. death), we call the *comparison of interest* (e.g. the difference in 5 year mortality rates) the **effect** of interest or, if relevant, the **treatment effect**. We express the null hypothesis in terms of no effect (e.g. the 5 year mortality from breast cancer is the same in two treatment groups); the two-sided alternative hypothesis is that the effect is not zero. We perform a hypothesis test that enables us to decide whether we have enough evidence to reject the null hypothesis (Chapter 17). We can make one of two decisions; either we reject the null hypothesis, or we do not reject it.

Making the wrong decision

Although we hope we will draw the correct conclusion about the null hypothesis, we have to recognize that, because we only have a sample of information, we may make the wrong decision when we reject/do not reject the null hypothesis. The possible mistakes we can make are shown in Table 18.1.

- **Type I error**: *we reject the null hypothesis when it is true*, and conclude that there is an effect when, in reality, there is none. The maximum chance (probability) of making a Type I error is denoted by α (alpha). This is the significance level of the test (Chapter 17); we reject the null hypothesis if our P-value is less than the significance level, i.e. if $P < \alpha$.

We must decide on the value of α before we collect our data; we usually assign a conventional value of 0.05 to it, although we might choose a more restrictive value such as 0.01 or a less restrictive value such as 0.10. Our chance of making a Type I error will never exceed our chosen significance level, say $\alpha = 0.05$, because we will only reject the null hypothesis if $P < 0.05$. If we find that $P > 0.05$, we will not reject the null hypothesis, and, consequently, do not make a Type I error.

- **Type II error**: *we do not reject the null hypothesis when it is false*, and conclude that there is no effect when one really exists. The chance of making a Type II error is denoted by β (beta); its compliment, $(1 - \beta)$, is the **power** of the test. The power, therefore, is the *probability of rejecting the null hypothesis when it is false*; i.e. it is the chance (usually expressed as a percentage) of detecting, as statistically significant, a real treatment effect of a given size.

Ideally, we should like the power of our test to be 100%; we must recognize, however, that this is impossible because there is always a chance, albeit slim, that we could make a Type II error. Fortunately, however, we know which factors affect power, and thus we can control the power of a test by giving consideration to them.

Power and related factors

It is essential that we know the power of a proposed test at the planning stage of our investigation. Clearly, we should only embark on a study if we believe that it has a 'good' chance of detecting a clinically relevant effect, if one exists (by 'good' we mean that the power should be at least 80%). It is ethically irresponsible, and wasteful of time and resources, to undertake a clinical trial that has, say, only a 40% chance of detecting a real treatment effect.

A number of factors have a direct bearing on power for a given test.

- The **sample size**: power increases with increasing sample size. This means that a large sample has a greater ability than a small sample to detect a clinically important effect if it exists. When the sample size is very small, the test may have an inadequate power to detect a particular effect. We explain how to choose sample size, with power considerations, in Chapter 36. The methods can also be used to evaluate the power of the test for a specified sample size.
- The **variability of the observations**: power increases as the variability of the observations decreases (Fig. 18.1).
- The **effect of interest**: the power of the test is greater for larger effects. A hypothesis test thus has a greater chance of detecting a large real effect than a small one.
- The **significance level**: the power is greater if the significance level is larger (this is equivalent to the probability of the Type I error (α) increasing as the probability of the Type II error (β) decreases). So, we are more likely to detect a real effect if we decide at the planning stage that we will regard our P-value as significant if it is less than 0.05 rather than less than 0.01. We can see this relationship between power and the significance level in Fig. 18.2.

Note that an inspection of the confidence interval (Chapter 11) for the effect of interest gives an indication of whether the power of the test was adequate. A wide confidence interval results from a small sample and/or data with substantial variability, and is a suggestion of low power.

Multiple hypothesis testing

Often, we want to carry out a number of significance tests on a data set, e.g. when it comprises many variables or there are more than two treatments. The Type I error rate increases dramatically as the number of comparisons increases, leading to spurious conclusions. Therefore, we should only perform a small number of tests, chosen to relate to the primary aims of the study and specified *a priori*. It is possible to use some form of *post-hoc* adjustment to the P-value to take account of the number of tests performed (Chapter 22). For example, the **Bonferroni** approach (often regarded as rather conservative) multiplies each P-value by the number of tests carried out; any decisions about significance are then based on this adjusted P-value.

Table 18.1 The consequences of hypothesis testing.

	Reject H_0	Do not reject H_0
H_0 true	Type I error	No error
H_0 false	No error	Type II error

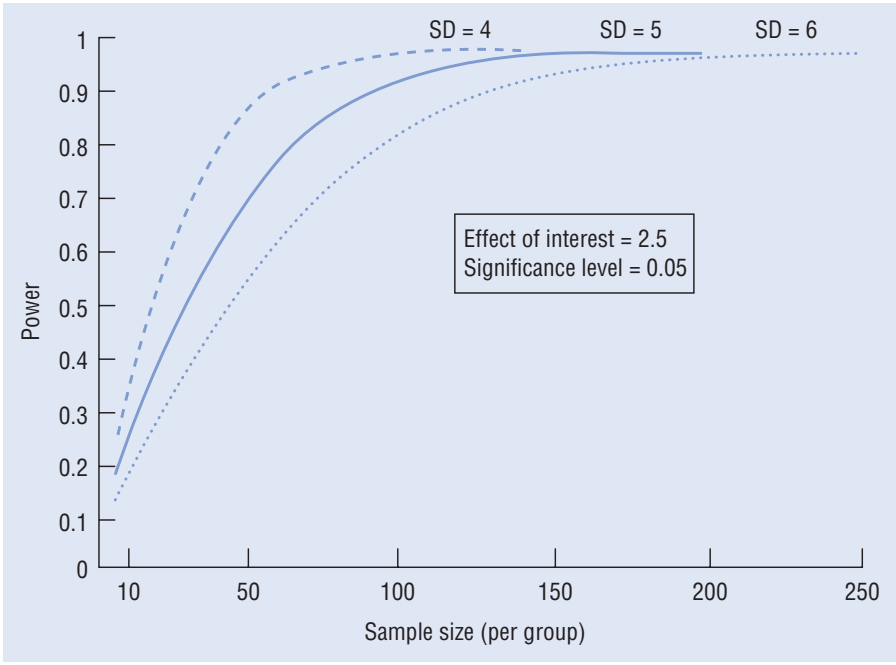


Figure 18.1 Power curves showing the relationship between power and the sample size in each of two groups for the comparison of two means using the unpaired *t*-test (Chapter 21). Each power curve relates to a two-sided test for which the significance level is 0.05, and the effect of interest (e.g. the difference between the treatment means) is 2.5. The assumed equal standard deviation of the measurements in the two groups is different for each power curve (see Example, Chapter 36).

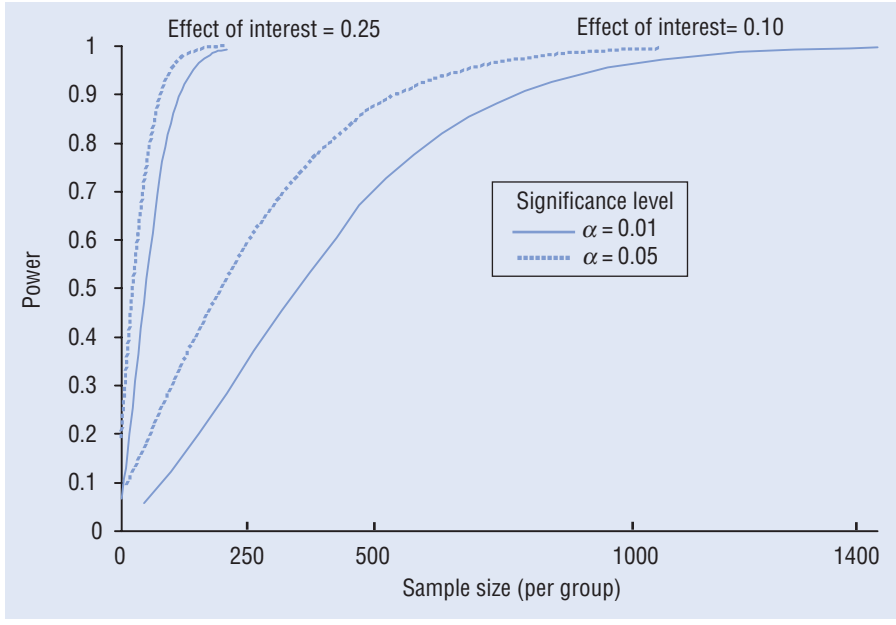


Figure 18.2 Power curves showing the relationship between power and the sample size in each of two groups for the comparison of two proportions using the Chi-squared test (Chapter 24). Curves are drawn when the effect of interest (e.g. the difference in the proportions with the characteristic of interest in the two treatment groups) is either 0.25 (i.e. 0.65 – 0.40) or 0.10 (i.e. 0.50 – 0.40); the significance level of the two-sided test is either 0.05 or 0.01 (see Example, Chapter 36).

19 Numerical data: a single group

The problem

We have a sample from a single group of individuals and one numerical or ordinal variable of interest. We are interested in whether the average of this variable takes a particular value. For example, we may have a sample of patients with a specific medical condition. We have been monitoring triglyceride levels in the blood of healthy individuals and know that they have a geometric mean of 1.74 mmol/L. We wish to know whether the average level in our patients is the same as this value.

The one-sample *t*-test

Assumptions

In the population, the variable is Normally distributed with a given (usually unknown) variance. In addition, we have taken a reasonable sample size so that we can check the assumption of Normality (Chapter 35).

Rationale

We are interested in whether the mean, μ, of the variable in the population of interest differs from some hypothesized value, μ_1. We use a test statistic that is based on the difference between the sample mean, $\bar{x}$, and μ_1. Assuming that we do not know the population variance, then this test statistic, often referred to as t, follows the t-distribution. If we do know the population variance, or the sample size is very large, then an alternative test (often called a z-test), based on the Normal distribution, may be used. However, in these situations, results from both tests are virtually identical.

Additional notation

Our sample is of size n and the estimated standard deviation is s.

1 Define the null and alternative hypotheses under study

H_0: the mean in the population, μ, equals μ_1

H_1: the mean in the population does not equal μ_1.

2 Collect relevant data from a sample of individuals

continued

3 Calculate the value of the test statistic specific to H_0

$$t = \frac{(\bar{x} - \mu_1)}{s/\sqrt{n}}$$

which follows the t-distribution with $(n-1)$ degrees of freedom.

4 Compare the value of the test statistic to values from a known probability distribution

Refer t to Appendix A2.

5 Interpret the *P*-value and results

Interpret the P-value and calculate a confidence interval for the true mean in the population (Chapter 11).

The 95% confidence interval is given by:

$$\bar{x} \pm t_{0.05} \times (s/\sqrt{n})$$

where $t_{0.05}$ is the percentage point of the t-distribution with $(n-1)$ degrees of freedom which gives a two-tailed probability of 0.05.

Interpretation of the confidence interval

The 95% confidence interval provides a range of values in which we are 95% certain that the true population mean lies. If the 95% confidence interval does not include the hypothesized value for the mean, μ_1, we reject the null hypothesis at the 5% level. If, however, the confidence interval includes μ_1, then we fail to reject the null hypothesis at that level.

If the assumptions are not satisfied

We may be concerned that the variable does not follow a Normal distribution in the population. Whereas the t-test is relatively **robust** (Chapter 35) to some degree of non-Normality, extreme skewness may be a concern. We can either transform the data, so that the variable is Normally distributed (Chapter 9), or use a non-parametric test such as the sign test or Wilcoxon signed ranks test (Chapter 20).

The sign test

Rationale

The sign test is a simple test based on the median of the distribution. We have some hypothesized value, λ, for the median in the population. If our sample comes from this population, then approximately half of the values in our sample should be greater than λ and half should be less than λ (after excluding any values which equal λ). The sign test considers the number of values in our sample that are greater (or less) than λ.

The sign test is a simple test; we can use a more powerful test, the Wilcoxon signed ranks test (Chapter 20), which takes into account the ranks of the data as well as their signs when carrying out such an analysis.

1 Define the null and alternative hypotheses under study

H_0: the median in the population equals λ

H_1: the median in the population does not equal λ.

2 Collect relevant data from a sample of individuals

3 Calculate the value of the test statistic specific to H_0

Ignore all values that are equal to λ, leaving n' values. Count the values that are greater than λ. Similarly, count the values that are less than λ. (In practice this will often involve calculating the difference between each value in the sample and λ, and noting its sign.) Consider r, the smaller of these two counts.

- If $n' \leq 10$, the test statistic is r
- If $n' > 10$, calculate $z = \dfrac{\left|r - \frac{n'}{2}\right| - \frac{1}{2}}{\frac{\sqrt{n'}}{2}}$

where $n'/2$ is the number of values above (or below) the median that we would expect if the null hypothesis were true. The vertical bars indicate that we take the absolute (i.e. the positive) value of the number inside the bars. The distribution of z is approximately Normal. The subtraction of $\frac{1}{2}$ in the formula for z is a **continuity correction**, which we have to include to allow for the fact that we are relating a discrete value (r) to a continuous distribution (the Normal distribution).

4 Compare the value of the test statistic to values from a known probability distribution

- If $n' \leq 10$, refer r to Appendix A6
- If $n' > 10$, refer z to Appendix A1.

5 Interpret the *P*-value and results

Interpret the P-value and calculate a confidence interval for the median—some statistical packages provide this automatically; if not, we can rank the values in order of size and refer to Appendix A7 to identify the ranks of the values that are to be used to define the limits of the confidence interval. In general, confidence intervals for the median will be wider than those for the mean.

Example

There is some evidence that high blood triglyceride levels are associated with heart disease. As part of a large cohort study on heart disease, triglyceride levels were available in 232 men who developed heart disease over the 5 years after recruitment. We are interested in whether the average triglyceride level in the population of men from which this sample is chosen is the same as that in the general population. A **one-sample *t*-test** was performed to investigate this. Triglyceride levels are skewed to the right (Fig. 8.3a); log triglyceride levels are approximately Normally distributed (Fig. 8.3b), so we performed our analysis on the log values. In the men in the general population, the mean of the log values equals 0.24 $\log_{10}$ (mmol/L) equivalent to a geometric mean of 1.74 mmol/L.

1 H_0: the mean $\log_{10}$ (triglyceride level) in the population of men who develop heart disease equals 0.24 log (mmol/L)

H_1: the mean $\log_{10}$ (triglyceride level) in the population of men who develop heart disease does not equal 0.24 log (mmol/L).

2 Sample size, $n = 232$

Mean of log values, $\bar{x} = 0.31$ log (mmol/L)

Standard deviation of log values, $s = 0.23$ log (mmol/L).

3 Test statistic, $t = \dfrac{0.31 - 0.24}{0.23/\sqrt{232}} = 4.64$

4 We refer t to Appendix A2 with 231 degrees of freedom: $P < 0.001$

5 There is strong evidence to reject the null hypothesis that the geometric mean triglyceride level in the population of men who develop heart disease equals 1.74 mmol/L. The geometric mean triglyceride level in the population of men who develop heart disease is estimated as antilog (0.31) = $10^{0.31}$, which equals 2.04 mmol/L. The 95% confidence interval for the geometric mean triglyceride level ranges from 1.90 to 2.19 mmol/L (i.e. antilog $[0.31 \pm 1.96 \times 0.23/\sqrt{232}]$). Therefore, in this population of patients, the geometric mean triglyceride level is significantly higher than that in the general population.

continued

We can use the **sign test** to carry out a similar analysis on the untransformed triglyceride levels as this does not make any distributional assumptions. We assume that the median and geometric mean triglyceride level in the male population are similar.

1 H_0: the median triglyceride level in the population of men who develop heart disease equals 1.74 mmol/L.

H_1: the median triglyceride level in the population of men who develop heart disease does not equal 1.74 mmol/L.

2 In this data set, the median value equals 1.94 mmol/L.

3 We investigate the differences between each value and 1.74. There are 231 non-zero differences, of which 135 are positive and 96 are negative. Therefore, $r = 96$. As the number of non-zero differences is greater than 10, we calculate:

$$z = \frac{\left|96 - \frac{231}{2}\right| - \frac{1}{2}}{\frac{\sqrt{231}}{2}} = 2.50$$

4 We refer z to Appendix A1: $P = 0.012$.

5 There is evidence to reject the null hypothesis that the median triglyceride level in the population of men who develop heart disease equals 1.74 mmol/L. The formula in Appendix A7 indicates that the 95% confidence interval for the population median is given by the 101st and 132nd ranked values; these are 1.77 and 2.16 mmol/L. Therefore, in this population of patients, the median triglyceride level is significantly higher than that in the general population.

Data kindly provided by Dr F.C. Lampe, Ms M. Walker and Dr P. Whincup, Department of Primary Care and Population Sciences, Royal Free and University College Medical School, London, UK.

20 Numerical data: two related groups

The problem

We have two samples that are related to each other and one numerical or ordinal variable of interest.

- The variable may be measured on each individual in two circumstances. For example, in a cross-over trial (Chapter 13), each patient has two measurements on the variable, one while taking active treatment and one while taking placebo.
- The individuals in each sample may be different, but are linked to each other in some way. For example, patients in one group may be individually matched to patients in the other group in a case–control study (Chapter 16).

Such data are known as paired data. It is important to take account of the dependence between the two samples when analysing the data, otherwise the advantages of pairing (Chapter 13) are lost. We do this by considering the differences in the values for each pair, thereby reducing our two samples to a single sample of differences.

The paired *t*-test

Assumption

In the population of interest, the individual differences are Normally distributed with a given (usually unknown) variance. We have a reasonable sample size so that we can check the assumption of Normality.

Rationale

If the two sets of measurements were the same, then we would expect the mean of the differences between each pair of measurements to be zero in the population of interest. Therefore, our test statistic simplifies to a one-sample *t*-test (Chapter 19) on the differences, where the hypothesized value for the mean difference in the population is zero.

Additional notation

Because of the paired nature of the data, our two samples must be of the same size, n. We have n differences, with sample mean, $\bar{x}$, and estimated standard deviation s_d.

1 Define the null and alternative hypotheses under study

H_0: the mean difference in the population equals zero

H_1: the mean difference in the population does not equal zero.

2 Collect relevant data from two related samples

3 Calculate the value of the test statistic specific to H_0

$$t = \frac{(\bar{d} - 0)}{\mathrm{SE}(\bar{d})} = \frac{\bar{d}}{s_d/\sqrt{n}}$$

which follows the *t*-distribution with $(n-1)$ degrees of freedom.

4 Compare the value of the test statistic to values from a known probability distribution

Refer *t* to Appendix A2.

continued

5 Interpret the *P*-value and results

Interpret the *P*-value and calculate a confidence interval for the true mean difference in the population. The 95% confidence interval is given by

$$\bar{d} \pm t_{0.05} \times (s_d/\sqrt{n})$$

where $t_{0.05}$ is the percentage point of the *t*-distribution with $(n-1)$ degrees of freedom which gives a two-tailed probability of 0.05.

If the assumption is not satisfied

If the differences do not follow a Normal distribution, the assumption underlying the *t*-test is not satisfied. We can either transform the data (Chapter 9), or use a non-parametric test such as the sign test (Chapter 19) or Wilcoxon signed ranks test to assess whether the differences are centred around zero.

The Wilcoxon signed ranks test

Rationale

In Chapter 19, we explained how to use the **sign test** on a single sample of numerical measurements to test the null hypothesis that the population median equals a particular value. We can also use the sign test when we have **paired** observations, the pair representing matched individuals (e.g. in a case–control study, Chapter 16) or measurements made on the same individual in different circumstances (as in a cross-over trial of two treatments, A and B, Chapter 13). For each pair, we evaluate the **difference** in the measurements. The sign test can be used to assess whether the median difference in the population equals zero by considering the differences in the sample and observing how many are greater (or less) than zero. However, the sign test does not incorporate information on the sizes of these differences.

The **Wilcoxon signed ranks test** takes account not only of the signs of the differences but also their magnitude, and therefore is a more powerful test (Chapter 18). The individual difference is calculated for each pair of results. Ignoring zero differences, these are then classed as being either positive or negative. In addition, the differences are placed in order of size, ignoring their signs, and are **ranked** accordingly. The smallest difference thus gets the value 1, the second smallest gets the value 2, etc. up to the largest difference, which is assigned the value n', if there are n' non-zero differences. If two or more of the differences are the same, they each receive the mean of the ranks these values would have received if they had not been tied. Under the null hypothesis of no difference, the sums of the ranks relating to the positive and negative differences should be the same (see following box).

1 Define the null and alternative hypotheses under study

H_0: the median difference in the population equals zero

H_1: the median difference in the population does not equal zero.

2 Collect relevant data from two related samples

3 Calculate the value of the test statistic specific to H_0

Calculate the difference for each pair of results. Ignoring their signs, rank all n' non-zero differences, assigning the value 1 to the smallest difference and the value n' to the largest. Sum the ranks of the positive (T_+) and negative differences (T_-).

- If $n' \leq 25$, the test statistic, T, takes the value T_+ or T_-, whichever is smaller
- If $n' > 25$, calculate the test statistic z, where:

$$z = \frac{\left|T - \frac{n'(n'+1)}{4}\right| - \frac{1}{2}}{\sqrt{\frac{n'(n'+1)(2n'+1)}{24}}}$$

z follows a Normal distribution (its value has to be adjusted if there are many tied values[1]).

4 Compare the value of the test statistic to values from a known probability distribution

- If $n' \leq 25$, refer T to Appendix A8
- If $n' > 25$, refer z to Appendix A1.

5 Interpret the *P*-value and results

Interpret the P-value and calculate a confidence interval for the median difference (Chapter 19) in the entire sample.

[1] Siegel, S. & Castellan, N.J. (1988) *Nonparametric Statistics for the Behavioural Sciences*, 2nd edn, McGraw-Hill, New York.

Examples

Ninety-six new recruits, all men aged between 16 and 20 years, had their teeth examined when they enlisted in the Royal Air Force. After receiving the necessary treatment to make their teeth dentally fit, they were examined one year later. A complete mouth, excluding wisdom teeth, has 28 teeth and, in this study, every tooth had four sites of periodontal interest; each recruit had a minimum of 84 and a maximum of 112 measurable sites on each occasion. It was of interest to examine the effect of treatment on pocket depth, a measure of gum disease (greater pocket depth indicates worse disease). Pocket depth was evaluated for each recruit as the mean pocket depth over the measurable sites in his mouth.

As the difference in pocket depth was approximately Normally distributed in this sample of recruits, a **paired *t*-test** was performed to determine whether the average pocket depth was the same before and after treatment. Full computer output is shown in Appendix C.

1 H_0: the mean difference in a man's average pocket depth before and after treatment in the population of recruits equals zero

H_1: the mean difference in a man's average pocket depth before and after treatment in the population of recruits does not equal zero.

2 Sample size, $n = 96$. Mean difference of average pocket depth, $\bar{x} = 0.1486$ mm. Standard deviation of differences, $s_d = 0.5601$ mm.

3 Test statistic, $t = \dfrac{0.1486}{0.5601/\sqrt{96}} = 2.60$

4 We refer t to Appendix A2 with $(96 - 1) = 95$ degrees of freedom: $0.01 < P < 0.05$ (computer output gives $P = 0.011$).

5 We have evidence to reject the null hypothesis, and can infer that a recruit's average pocket depth was reduced after treatment. The 95% confidence interval for the true mean difference in average pocket depth is 0.035 to 0.262 mm (i.e. $0.1486 \pm 1.95 \times 0.5601/\sqrt{96}$). Of course, we have to be careful here if we want to conclude that it is the treatment that has reduced average pocket depth, as we have no control group of recruits who did not receive treatment. The improvement may be a consequence of time or a change in dental hygiene habits, and may not be due to the treatment received.

continued

The data in the following table show the percentage of measurable sites for which there was loss of attachment at each assessment in each of 14 of these recruits who were sent to a particular air force base. Loss of attachment is an indication of gum disease that may be more advanced than that assessed by pocket depth. As the differences in the percentages were not Normally distributed, we performed a **Wilcoxon signed ranks test** to investigate whether treatment had any effect on loss of attachment.

1 H_0: the median of the differences (before and after treatment) in the percentages of sites with loss of attachment equals zero in the population of recruits

H_1: the median of the differences in the percentages of sites with loss of attachment does not equal zero in the population.

2 The percentages of measurable sites with loss of attachment, before and after treatment, for each recruit are shown in the table below.

3 There is one zero difference; of the remaining $n' = 13$ differences, three are positive and 10 are negative. The sum of the ranks of the positive differences, $T_+ = 3 + 5 + 13 = 21$.

4 As $n' < 25$, we refer T_+ to Appendix A8: $P > 0.05$ (computer output gives $P = 0.09$).

5 There is insufficient evidence to reject the null hypothesis of no change in the percentage of sites with loss of attachment. The median difference in the percentage of sites with loss of attachment is −3.1% (i.e. the mean of −2.5% and −3.6%), a negative median difference indicating that, on average, the percentage of sites with loss of attachment is greater *after* treatment, although this difference is not significant. Appendix A7 shows that the approximate 95% confidence interval for the median difference in the population is given by the 3rd and the 12th ranked differences (including the zero difference); these are −12.8% and 0.9%. Although the result of the test is not significant, the lower limit indicates that the percentage of sites with loss of attachment could be as much as 12.8% more after the recruit received treatment!

Recruit	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Before (%)	65.5	75.0	87.2	97.1	100.0	92.6	82.3	90.0	93.0	100.0	91.7	97.7	79.0	95.4
After (%)	100.0	10.0	100.0	97.1	99.1	100.0	91.6	94.6	95.5	97.3	92.3	98.0	100.0	99.0
Difference (%)	−34.5	65.0	−12.8	0.0	0.9	−7.4	−9.3	−4.6	−2.5	2.7	−0.6	−0.3	−21.0	−3.6
Rank	12	13	10	—	3	8	9	7	4	5	2	1	11	6

Duffy, S. (1997) Results of a three year longitudinal study of early periodontitis in a group of British male adolescents. MSc Dissertation, University of London, Eastman Dental Institute for Oral Health Care Sciences.

21 Numerical data: two unrelated groups

The problem

We have samples from two independent (unrelated) groups of individuals and one numerical or ordinal variable of interest. We are interested in whether the mean or distribution of the variable is the same in the two groups. For example, we may wish to compare the weights in two groups of children, each child being randomly allocated to receive either a dietary supplement or placebo.

The unpaired (two-sample) *t*-test

Assumptions

In the population, the variable is Normally distributed in each group and the variances of the two groups are the same. In addition, we have reasonable sample sizes so that we can check the assumptions of Normality and equal variances.

Rationale

We consider the difference in the means of the two groups. Under the null hypothesis that the population means in the two groups are the same, this difference will equal zero. Therefore, we use a test statistic that is based on the difference in the two sample means, and on the value of the difference in population means under the null hypothesis (i.e. zero). This test statistic, often referred to as t, follows the t-distribution.

Notation

Our two samples are of size n_1 and n_2. Their means are $\bar{x}_1$ and $\bar{x}_2$; their standard deviations are s_1 and s_2.

1 Define the null and alternative hypotheses under study

H_0: the population means in the two groups are equal

H_1: the population means in the two groups are not equal.

2 Collect relevant data from two samples of individuals

3 Calculate the value of the test statistic specific to H_0

If s is an estimate of the pooled standard deviation of the two groups,

$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

then the test statistic is given by t where:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which follows the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

continued

4 Compare the value of the test statistic to values from a known probability distribution

Refer t to Appendix A2. When the sample sizes in the two groups are large, the t-distribution approximates a Normal distribution, and then we reject the null hypothesis at the 5% level if the absolute value (i.e. ignoring the sign) of t is greater than 1.96.

5 Interpret the *P*-value and results

Interpret the P-value and calculate a confidence interval for the difference in the two means. The 95% confidence interval, assuming equal variances, is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0.05} \times SE(\bar{x}_1 - \bar{x}_2)$$

where $t_{0.05}$ is the percentage point of the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom which gives a two-tailed probability of 0.05.

Interpretation of the confidence interval

The upper and lower limits of the confidence interval can be used to assess whether the difference between the two mean values is clinically important. For example, if the upper and/or lower limit is close to zero, this indicates that the true difference may be very small and clinically meaningless, even if the test is statistically significant.

If the assumptions are not satisfied

When the sample sizes are reasonably large, the t-test is fairly robust (Chapter 35) to departures from Normality. However, it is less robust to unequal variances. There is a modification of the unpaired t-test that allows for unequal variances, and results from it are often provided in computer output. However, if you are concerned that the assumptions are not satisfied, then you either transform the data (Chapter 9) to achieve approximate Normality and/or equal variances, or use a non-parametric test such as the Wilcoxon rank sum test.

The Wilcoxon rank sum (two-sample) test

Rationale

The **Wilcoxon rank sum test** makes no distributional assumptions and is the non-parametric equivalent to the unpaired t-test. The test is based on the sum of the ranks of the values in each of the two groups; these should be comparable after allowing for differences in sample size if the groups have similar distributions. An equivalent test, known as the **Mann–Whitney *U* test**, gives identical results although it is slightly more complicated to carry out by hand.

1 Define the null and alternative hypotheses under study

H_0: the two groups have the same distribution in the population

H_1: the two groups have different distributions in the population.

2 Collect relevant data from two samples of individuals

3 Calculate the value of the test statistic specific to H_0

All observations are ranked as if they were from a single sample. Tied observations are given the mean of the ranks the values would have received if they had not been tied. The sum of the ranks, T, is then calculated in the group with the smaller sample size.

- If the sample size in each group is 15 or less, T is the test statistic
- If at least one of the groups has a sample size of more than 15, calculate the test statistic

$$z = \frac{(T - \mu_T)}{\sigma_T}$$

which follows a Normal distribution, where

$$\mu_T = \frac{n_S(n_S + n_L + 1)}{2} \quad \sigma_T = \sqrt{n_L \mu_T / 6}$$

and n_S and n_L are the sample sizes of the smaller and larger groups, respectively. z must be adjusted if there are many tied values[1].

4 Compare the value of the test statistic to values from a known probability distribution

- If the sample size in each group is 15 or less, refer T to Appendix A9
- If at least one of the groups has a sample size of more than 15, refer z to Appendix A1.

5 Interpret the *P*-value and results

Interpret the P-value and obtain a confidence interval for the difference in the two medians. This is time-consuming to calculate by hand so details have not been included; some statistical packages will provide the CI. If this confidence interval is not included in your package, you can quote a confidence interval for the median in each of the two groups.

[1] Siegel, S. & Castellan, N.J. (1988) *Nonparametric Statistics for the Behavioural Sciences*, 2nd edn. McGraw-Hill, New York.

Example 1

In order to determine the effect of regular prophylactic inhaled corticosteroids on wheezing episodes associated with viral infection in school age children, a randomized, double-blind controlled trial was carried out comparing inhaled beclomethasone dipropionate with placebo. In this investigation, the primary endpoint was the mean forced expiratory volume (FEV1) over a 6 month period. After checking the assumptions of Normality and constant variance (see Fig. 4.2), we performed an **unpaired *t*-test** to compare the means in the two groups. Full computer output is shown in Appendix C.

1 H_0: the mean FEV1 in the population of school age children is the same in the two treatment groups

H_1: the mean FEV1 in the population of school age children is not the same in the two treatment groups.

2 Treated group: sample size, $n_1 = 50$; mean, $\bar{x}_1 = 1.64$ litres, standard deviation, $s_1 = 0.29$ litres

Placebo group: sample size, $n_2 = 48$; mean, $\bar{x}_2 = 1.54$ litres; standard deviation, $s_2 = 0.25$ litres.

3 Pooled standard deviation,

$$s = \sqrt{\frac{(49 \times 0.29^2) + (47 \times 0.25^2)}{(50 + 48 - 2)}} = 0.2670 \text{ litres.}$$

$$\text{Test statistic, } t = \frac{1.64 - 1.54}{0.2670 \times \sqrt{\frac{1}{50} + \frac{1}{48}}} = 1.9145$$

4 We refer t to Appendix A2 with $50 + 48 - 2 = 96$ degrees of freedom. Because Appendix A2 is restricted to certain degrees of freedom, we have to **interpolate** (estimate the required value that lies between two known values). We therefore interpolate between the values relating to 50 and 100 degrees of freedom. Hence, $P > 0.05$ (computer output gives $P = 0.06$).

5 We have insufficient evidence to reject the null hypothesis at the 5% level. However, as the P-value is only just greater than 0.05, there may be an indication that the two population means are different. The estimated difference between the two means is $1.64 - 1.54 = 0.10$ litres. The 95% confidence interval for the true difference in the two means ranges from –0.006 to 0.206 litres

$$\left[= 0.10 \pm \left(1.96 \times 0.2670 \times \sqrt{\frac{1}{50} + \frac{1}{48}} \right) \right].$$

Data kindly provided by Dr I. Doull, Cystic Fibrosis/Respiratory Unit, Department of Child Health, University Hospital of Wales, Cardiff, UK and Dr F.C. Lampe, Department of Primary Care and Population Sciences, Royal Free and University College Medical School, London, UK.

Example 2

In order to study whether the mechanisms involved in fatal soybean asthma are different from that of normal fatal asthma, the number of CD3+ T cells in the submucosa, a measure of the body's immune system, was compared in seven cases of fatal soybean dust-induced asthma and 10 fatal asthma cases. Because of the small sample sizes and obviously skewed data, we performed a **Wilcoxon rank sum test** to compare the distributions.

1 H_0: the distributions of CD3+ T-cell numbers in the two groups in the population are the same

H_1: the distributions of CD3+ T-cell numbers in the two groups in the population are not the same.

2 Soy-bean group: sample size, n_S = 7, CD3+ T-cell levels (cells/mm^2) were 34.45, 0.00, 1.36, 0.00, 1.43, 0.00, 4.01

Asthma group: sample size, n_L = 10, CD3+ T-cell levels (cells/mm^2) were 74.17, 13.75, 37.50, 1225.51, 99.99, 3.76, 58.33, 73.63, 4.32, 154.86.

The ranked data are shown in the table below.

3 Sum of the ranks in the soy-bean group = 2 + 2 + 2 + 4 + 5 + 7 + 10 = 32

Sum of the ranks in the asthma group = 6 + 8 + 9 + 11 + 12 + 13 + 14 + 15 + 16 + 17 = 121.

4 Because there are 10 or less values in each group, we obtain the P-value from Appendix A9: $P < 0.01$ (computer output gives $P = 0.002$).

5 There is evidence to reject the null hypothesis that the distributions of CD3+ T-cell levels are the same in the two groups. The median number of CD3+ T cells in the soybean and fatal asthma groups are 1.36 (95% confidence interval 0 to 34.45) and (58.33 + 73.63)/2 = 65.98 (95% confidence interval 4.32 to 154.86) cells/mm^2, respectively. We thus believe that CD3+ T cells are reduced in fatal soybean asthma, suggesting a different mechanism from that described for most asthma deaths.

Soy-bean	0.00	0.00	0.00	1.36	1.43		4.01			34.45							
Asthma						3.76		4.32	13.75		37.50	58.33	73.63	74.17	99.99	154.86	1225.51
Rank	2	2	2	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Data kindly provided by Dr M. Synek, Coldeast Hospital, Sarisbury, and Dr F.C. Lampe, Department of Primary Care and Population Sciences, Royal Free and University College Medical School, London UK.

22 Numerical data: more than two groups

The problem

We have samples from a number of independent groups. We have a single numerical or ordinal variable and are interested in whether the average value of the variable varies in the different groups, e.g. whether the average platelet count varies in groups of women with different ethnic backgrounds. Although we could perform tests to compare the averages in each pair of groups, the high Type I error rate, resulting from the large number of comparisons, means that we may draw incorrect conclusions (Chapter 18). Therefore, we carry out a single **global** test to determine whether the averages differ in any groups.

One-way analysis of variance

Assumptions

The groups are defined by the *levels* of a single factor (e.g. different ethnic backgrounds). In the population of interest, the variable is Normally distributed in each group and the variance in each group is the same. We have a reasonable sample size so that we can check these assumptions.

Rationale

The one-way analysis of variance separates the total variability in the data into that which can be attributed to differences between the individuals from the different groups (the **between-group variation**), and to the random variation between the individuals within each group (the **within-group variation**, sometimes called **unexplained** or **residual** variation). These components of variation are measured using variances, hence the name **analysis of variance** (**ANOVA**). Under the null hypothesis that the group means are the same, the between-group variance will be similar to the within-group variance. If, however, there are differences between the groups, then the between-group variance will be larger than the within-group variance. The test is based on the ratio of these two variances.

Notation

We have k independent samples, each derived from a different group. The sample sizes, means and standard deviations in each group are n_i, $\bar{x}_i$ and s_i, respectively ($i = 1, 2, \ldots, k$). The total sample size is $n = n_1 + n_2 + \ldots + n_k$.

1 Define the null and alternative hypotheses under study

H_0: all group means in the population are equal

H_1: at least one group mean in the population differs from the others.

2 Collect relevant data from samples of individuals

3 Calculate the value of the test statistic specific to H_0

The test statistic for ANOVA is a ratio, F, of the between-group variance to the within-group variance. This F-statistic follows the F-distribution (Chapter 8) with $(k-1, n-1)$ degrees of freedom in the numerator and denominator, respectively.

The calculations involved in ANOVA are complex and are not shown here. Most computer packages will output the values directly in an ANOVA table, which usually includes the F-ratio and P-value (see Example 1).

4 Compare the value of the test statistic to values from a known probability distribution

Refer the F-ratio to Appendix A5. Because the between-group variation is greater than or equal to the within-group variation, we look at the one-sided P-values.

5 Interpret the *P*-value and results

If we obtain a significant result at this initial stage, we may consider performing specific pairwise ***post-hoc*** comparisons. We can use one of a number of special tests devised for this purpose (e.g. **Duncan's**, **Scheffé's**) or we can use the unpaired t-test (Chapter 21) adjusted for multiple hypothesis testing (Chapter 18). We can also calculate a confidence interval for each individual group mean (Chapter 11). Note that we use a pooled estimate of the variance of the values from *all* groups when calculating confidence intervals and performing t-tests. Most packages refer to this estimate of the variance as the **residual variance** or **residual mean square**; it is found in the ANOVA table.

Although the two tests appear to be different, the unpaired t-test and ANOVA give equivalent results when there are only two groups of individuals.

If the assumptions are not satisfied

Although ANOVA is relatively robust (Chapter 35) to moderate departures from Normality, it is not robust to unequal variances. Therefore, before carrying out the analysis, we check for Normality, and test whether the variances are similar in the groups either by eyeballing them, or by using **Levene's** test or **Bartlett's** test (Chapter 35). If the assumptions are not satisfied, we can either transform the data (Chapter 9) or use the non-parametric equivalent of one-way ANOVA, the **Kruskal–Wallis** test.

The Kruskal–Wallis test

Rationale

This non-parametric test is an extension of the Wilcoxon rank sum test (Chapter 21). Under the null hypothesis of no differences in the distributions between the groups, the sums of the ranks in each of the k groups should be comparable after allowing for any differences in sample size.

1 Define the null and alternative hypotheses under study

H_0: each group has the same distribution of values in the population

H_1: each group does not have the same distribution of values in the population.

2 Collect relevant data from samples of individuals

3 Calculate the value of the test statistic specific to H_0

Rank all n values and calculate the sum of the ranks in each of the groups: these sums are $R_1, \ldots R_k$. The test statistic (which should be modified if there are many tied values[1]) is given by:

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$$

which follows a Chi-squared distribution with $(k-1)$ *df*

4 Compare the value of the test statistic to values from a known probability distribution

Refer H to Appendix A3.

5 Interpret the *P*-value and results

Interpret the P-value and, if significant, perform two-sample non-parametric tests, adjusting for multiple testing. Calculate a confidence interval for the median in each group.

[1] Siegel, S. and Castellan, N.J. (1988) *Nonparametric Statistics for the Behavioral Sciences*, 2nd edn. McGraw-Hill, New York.

We use one-way ANOVA or its non-parametric equivalent when the groups relate to a single factor and are independent. We can use other forms of ANOVA when the study design is more complex[2].

[2] Mickey, R.M., Dunn, O.J. and Clark, V.A. (2004) *Applied Statistics: Analysis of Variance and Regression.* 3rd Edn. Wiley, Chichester.

Example 1

A total of 150 women of different ethnic backgrounds were included in a cross-sectional study of factors related to blood clotting. We compared mean platelet levels in the four groups using a **one-way ANOVA**. It was reasonable to assume Normality and constant variance, as shown in the computer output (Appendix C).

1 H_0: there are no differences in the mean platelet levels in the four groups in the population

H_1: at least one group mean platelet level differs from the others in the population.

2 The following table summarizes the data in each group.

Group	Sample size n (%)	Mean ($\times 10^9$) $\bar{x}$	Standard deviation ($\times 10^9$), s	95% CI for mean (using pooled standard deviation—see point 3)
Caucasian	90 (60.0)	268.1	77.08	252.7 to 283.5
Afro-Caribbean	21 (14.0)	254.3	67.50	220.9 to 287.7
Mediterranean	19 (12.7)	281.1	71.09	245.7 to 316.5
Other	20 (13.3)	273.3	63.42	238.9 to 307.7

continued

3 The following ANOVA table is extracted from the computer output:

Source	Sum of squares	*df*	Mean square	*F*-ratio	*P*-value
Between ethnic group	7711.967	3	2570.656	0.477	0.6990
Within ethnic group	787289.533	146	5392.394		

Pooled standard deviation = $\sqrt{5392.394} \times 10^9 = 73.43 \times 10^9$.

4 The ANOVA table gives $P = 0.70$. (We could have referred F to Appendix A5 with (3, 146) degrees of freedom to determine the P-value.)

5 There is insufficient evidence to reject the null hypothesis that the mean levels in the four groups in the population are the same.

Data kindly provided by Dr R.A. Kadir, University Department of Obstetrics and Gynaecology, and Professor C.A. Lee, Haemophilia Centre and Haemostasis Unit, Royal Free Hospital, London, UK.

Example 2

Quality-of-life scores, measured using the SF-36 questionnaire, were obtained in three groups of individuals: those with severe haemophilia, those with mild/moderate haemophilia, and normal controls. Each group comprised a sample of 20 individuals. Scores on the physical functioning scale (PFS), which can take values from 0 to 100, were compared in the three groups. As visual inspection of Fig. 22.1 showed that the data were not Normally distributed, we performed a **Kruskal–Wallis** test.

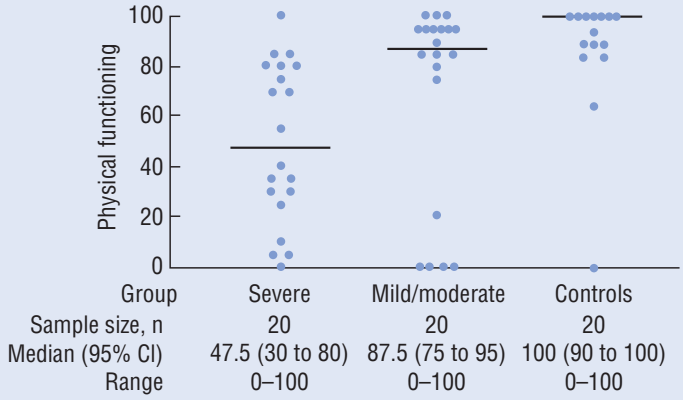


Figure 22.1 Dot plot showing physical functioning scores (from the SF-36 questionnaire) in individuals with severe and mild/moderate haemophilia and in normal controls. The horizontal bars are the medians.

1 H_0: each group has the same distribution of PFS scores in the population

H_1: at least one of the groups has a different distribution of PFS scores in the population.

2 The data are shown in Fig. 22.1.

3 Sum of ranks in severe haemophilia group = 372
Sum of ranks in mild/moderate haemophilia group = 599
Sum of ranks in normal control group = 859.

$$H = \frac{12}{60(60+1)}\left(\frac{372^2}{20} + \frac{599^2}{20} + \frac{859^2}{20}\right) - 3(60+1) = 19.47$$

4 We refer H to Appendix A3: $P < 0.001$.

5 There is substantial evidence to reject the null hypothesis that the distribution of PFS scores is the same in the three groups. Pairwise comparisons were carried out using Wilcoxon rank sum tests, adjusting the P-values for the number of tests performed using the Bonferroni correction (Chapter 18). The individuals with severe and mild/moderate haemophilia both had significantly lower PFS scores than the controls ($P = 0.0003$ and $P = 0.03$, respectively) but the distributions of the scores in the haemophilia groups were not significantly different from each other ($P = 0.09$).

Data kindly provided by Dr A. Miners, Department of Primary Care and Population Sciences, Royal Free and University College Medical School, London, UK, and Dr C. Jenkinson, Health Services Research Unit, University of Oxford, Oxford, UK.

23 Categorical data: a single proportion

The problem

We have a single sample of n individuals; each individual either 'possesses' a characteristic of interest (e.g. is male, is pregnant, has died) or does not possess that characteristic (e.g. is female, is not pregnant, is still alive). A useful summary of the data is provided by the **proportion** of individuals with the characteristic. We are interested in determining whether the true proportion in the population of interest takes a particular value.

The test of a single proportion

Assumptions

Our sample of individuals is selected from the population of interest. Each individual either has or does not have the particular characteristic.

Notation

r individuals in our sample of size n have the characteristic. The estimated proportion with the characteristic is $p = r/n$. The proportion of individuals with the characteristic in the population is π. We are interested in determining whether π takes a particular value, π_1.

Rationale

The number of individuals with the characteristic follows the Binomial distribution (Chapter 8), but this can be approximated by the Normal distribution, providing np and $n(1-p)$ are each greater than 5. Then p is approximately Normally distributed with an estimated mean $= p$ and an estimated standard deviation $= \sqrt{\frac{p(1-p)}{n}}$. Therefore, our test statistic, which is based on p, also follows the Normal distribution.

1 Define the null and alternative hypotheses under study

H_0: the population proportion, π, is equal to a particular value, π_1

H_1: the population proportion, π, is not equal to π_1.

2 Collect relevant data from a sample of individuals

continued

3 Calculate the value of the test statistic specific to H_0

$$z = \frac{|p - \pi_1| - \frac{1}{2n}}{\sqrt{\frac{p(1-p)}{n}}}$$

which follows a Normal distribution.

The $1/2n$ in the numerator is a **continuity correction**: it is included to make an allowance for the fact that we are approximating the discrete Binomial distribution by the continuous Normal distribution.

4 Compare the value of the test statistic to values from a known probability distribution

Refer z to Appendix A1.

5 Interpret the *P*-value and results

Interpret the P-value and calculate a confidence interval for the true population proportion, π. The 95% confidence interval for π is:

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

We can use this confidence interval to assess the clinical or biological importance of the results. A wide confidence interval is an indication that our estimate has poor precision.

The sign test applied to a proportion

Rationale

The sign test (Chapter 19) may be used if the response of interest can be expressed as a **preference** (e.g. in a cross-over trial, patients may have a preference for either treatment A or treatment B). If there is no preference overall, then we would expect the proportion preferring A, say, to equal 1/2. We use the sign test to assess whether this is so.

Although this formulation of the problem and its test statistic appear to be different from those of Chapter 19, both approaches to the sign test produce the same result.

1 Define the null and alternative hypotheses under study

H_0: the proportion, π, of preferences for A in the population is equal to 1/2

H_1: the proportion of preferences for A in the population is not equal to 1/2.

2 Collect relevant data from a sample of individuals

3 Calculate the value of the test statistic specific to H_0

Ignore any individuals who have no preference and reduce the sample size from n to n' accordingly. Then $p = r/n'$, where r is the number of preferences for A.

- If $n' \leq 10$, count r, the number of preferences for A
- If $n' > 10$, calculate the test statistic:

$$z' = \frac{\left|p - \frac{1}{2}\right| - \frac{1}{2n'}}{\sqrt{\frac{p(1-p)}{n'}}}$$

z' follows the Normal distribution. Note that this formula is based on the test statistic, z, used in the previous box to test the null hypothesis that the population proportion equals π_1; here we replace n by n', and π_1 by 1/2.

4 Compare the value of the test statistic to values from a known probability distribution

- If $n' \leq 10$, refer r to Appendix A6
- If $n' > 10$, refer z' to Appendix A1.

5 Interpret the *P*-value and results

Interpret the P-value and calculate a confidence interval for the proportion of preferences for A in the entire sample of size n.

Example 1

Human herpes-virus 8 (HHV-8) has been linked to Kaposi's sarcoma, primary effusion lymphoma and certain types of multicentric Castleman's disease. It has been suggested that HHV-8 can be transmitted sexually. In order to assess the relationships between sexual behaviour and HHV-8 infection, the prevalence of antibodies to HHV-8 was determined in a group of 271 homo/bisexual men attending a London sexually transmitted disease clinic. In the blood donor population in the UK, the seroprevalence of HHV-8 has been documented to be 2.7%. Initially, the seroprevalence from this study was compared to 2.7% using a **single proportion** test.

1 H_0: the seroprevalence of HHV-8 in the population of homo/bisexual men equals 2.7%

H_1: the seroprevalence of HHV-8 in the population of homo/bisexual men does not equal 2.7%.

2 Sample size, $n = 271$; number who are seropositive to HHV8, $r = 50$

Seroprevalence, $p = 50/271 = 0.185$ (i.e. 18.5%).

3 Test statistic is $z = \dfrac{|0.185 - 0.027| - \dfrac{1}{2 \times 271}}{\sqrt{\dfrac{0.185(1-0.185)}{271}}} = 6.62$

4 We refer z to Appendix A1: $P < 0.0001$.

5 There is substantial evidence that the seroprevalence of HHV-8 in homo/bisexual men attending sexually transmitted disease clinics in the UK is higher than that in the blood donor population. The 95% confidence interval for the seroprevalence of HHV-8 in the population of homo/bisexual men is 13.9% to 23.1%, calculated as

$$\left\{0.185 \pm 1.96 \times \sqrt{\frac{0.185 \times (1-0.185)}{271}}\right\} \times 100\%.$$

Data kindly provided by Drs N.A. Smith, D. Barlow, and B.S. Peters, Department of Genitourinary Medicine, Guy's and St Thomas' NHS Trust, London, and Dr J. Best, Department of Virology, Guy's, King's College and St Thomas's School of Medicine, King's College, London, UK.

Example 2

In a double-blind cross-over study, 36 adults with perennial allergic rhinitis were treated with subcutaneous injections of either inhalant allergens or placebo, each treatment being given daily for a defined period. The patients were asked whether they preferred the active drug or the placebo. The **sign test** was performed to investigate whether the proportions preferring the two preparations were the same.

1 H_0: the proportion preferring the active preparation in the population equals 0.5

H_1: the proportion preferring the active preparation in the population does not equal 0.5.

2 Of the 36 adults, 27 expressed a preference; 21 preferred the active preparation. Of those with a preference, the proportion preferring the active preparation, $p = 21/27 = 0.778$.

3 Test statistic is $$z' = \frac{|0.778 - 0.5| - \frac{1}{2 \times 27}}{\sqrt{\frac{0.778(1-0.778)}{27}}} = 3.24$$

4 We refer z' to Appendix A1: $P = 0.001$

5 There is substantial evidence to reject the null hypothesis that the two preparations are preferred equally in the population. The 95% confidence interval for the true proportion is 0.62 to 0.94, calculated as

$$0.778 \pm 1.96 \times \sqrt{\frac{0.778 \times (1-0.778)}{27}}.$$

Therefore, at the very least, we believe that almost two-thirds of individuals in the population prefer the active preparation.

Data adapted from: Radcliffe, M.J., Lampe, F.C., Brostoff, J. (1996) Allergen-specific low-dose immunotherapy in perennial allergic rhinitis: a double-blind placebo-controlled crossover study. *Journal of Investigational Allergology and Clinical Immunology*, **6**, 242–247.

24 Categorical data: two proportions

The problems

- We have two independent groups of individuals (e.g. homosexual men with and without a history of gonorrhoea). We want to know if the proportions of individuals with a characteristic (e.g. infected with human herpesvirus-8, HHV-8) are the same in the two groups.
- We have two related groups, e.g. individuals may be matched, or measured twice in different circumstances (say, before and after treatment). We want to know if the proportions with a characteristic (e.g. raised test result) are the same in the two groups.

Independent groups: the Chi-squared test

Terminology

The data are obtained, initially, as **frequencies**, i.e. the numbers with and without the characteristic in each sample. A table in which the entries are frequencies is called a **contingency table**; when this table has two rows and two columns it is called a **2 × 2 table**. Table 24.1 shows the **observed** frequencies in the four cells corresponding to each row/column combination, the four **marginal totals** (the frequency in a specific row or column, e.g. $a + b$), and the **overall total**, n. We can calculate (see Rationale) the frequency that we would expect in each of the four cells of the table if H_0 were true (the **expected frequencies**).

Assumptions

We have samples of sizes n_1 and n_2 from two independent groups of individuals. We are interested in whether the proportions of individuals who possess the characteristic are the same in the two groups. Each individual is represented only once in the study. The rows (and columns) of the table are **mutually exclusive**, implying that each individual can belong in only one row and only one column. The usual, albeit conservative, approach requires that the expected frequency in each of the four cells is at least five.

Rationale

If the proportions with the characteristic in the two groups are equal, we can estimate the overall proportion of individuals with the characteristic by $p = (a + b)/n$; we **expect** $n_1 \times p$ of them to be in Group 1 and $n_2 \times p$ to be in Group 2. We evaluate expected numbers without the characteristic similarly. Therefore, *each expected frequency is the product of the two relevant marginal totals divided by the overall total*. A large discrepancy between the observed (O) and the corresponding expected (E) frequencies is an indication that the proportions in the two groups differ. The test statistic is based on this discrepancy.

Table 24.1 Observed frequencies.

Characteristic	Group 1	Group 2	Total
Present	a	b	$a+b$
Absent	c	d	$c+d$
Total	$n_1 = a+c$	$n_2 = b+d$	$n = a+b+c+d$
Proportion with characteristic	$p_1 = \frac{a}{n_1}$	$p_2 = \frac{b}{n_2}$	$p = \frac{a+b}{n}$

1 Define the null and alternative hypotheses under study

H_0: the proportions of individuals with the characteristic are equal in the two groups in the population

H_1: these population proportions are not equal.

2 Collect relevant data from samples of individuals

3 Calculate the value of the test statistic specific to H_0

$$\chi^2 = \sum \frac{\left(|O-E| - \frac{1}{2}\right)^2}{E}$$

where O and E are the observed and expected frequencies, respectively, in each of the four cells of the table. The vertical lines around O–E indicate that we ignore its sign. The 1/2 in the numerator is the continuity correction (Chapter 19). The test statistic follows the Chi-squared distribution with 1 degree of freedom.

4 Compare the value of the test statistic to values from a known probability distribution

Refer χ^2 to Appendix A3.

5 Interpret the *P*-value and results

Interpret the P-value and calculate the confidence interval for the difference in the true population proportions. The 95% confidence interval is given by:

$$(p_1 - p_2) \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

If the assumptions are not satisfied

If $E < 5$ in any one cell, we use **Fisher's exact test** to obtain a P-value that does not rely on the approximation to the Chi-squared distribution. This is best left to a computer program as the calculations are tedious to perform by hand.

Related groups: McNemar's test

Assumptions

The two groups are related or dependent, e.g. each individual may be measured in two different circumstances. Every individual is classified according to whether the characteristic is present in both circumstances, one circumstance only, or in neither (Table 24.2).

Table 24.2 Observed frequencies of pairs in which the characteristic is present or absent.

	Circumstance 1		
	Present	Absent	Total no. of pairs
Circumstance 2			
Present	w	x	$w+x$
Absent	y	z	$y+z$
Total	$w+y$	$x+z$	$m = w+x+y+z$

Rationale

The observed proportions with the characteristic in the two circumstances are $(w+y)/m$ and $(w+x)/m$. They will differ if x and y differ. Therefore, to compare the proportions with the characteristic, we ignore those individuals who agree in the two circumstances, and concentrate on the discordant pairs, x and y.

1 Define the null and alternative hypotheses under study

H_0: the proportions with the characteristic are equal in the two groups in the population

H_1: these population proportions are not equal.

2 Collect relevant data from two samples

3 Calculate the value of the test statistic specific to H_0

$$\chi^2 = \frac{(|x-y|-1)^2}{x+y}$$

which follows the Chi-squared distribution with 1 degree of freedom. The 1 in the numerator is a continuity correction (Chapter 19).

4 Compare the value of the test statistic with values from a known probability distribution

Refer χ^2 to Appendix A3.

5 Interpret the *P*-value and results

Interpret the P-value and calculate the confidence interval for the difference in the true population proportions. The approximate 95% CI is:

$$\frac{x-y}{m} \pm \frac{1.96}{m}\sqrt{x+y-\frac{(x-y)^2}{m}}$$

Example 1

In order to assess the relationship between sexual risk factors and HHV-8 infection (study described in Chapter 23), the prevalence of seropositivity to HHV-8 was compared in homo/bisexual men who had a previous history of gonorrhoea, and those who had not previously had gonorrhoea, using the **Chi-squared test**. A typical computer output is shown in Appendix C.

1 H_0: the seroprevalence of HHV-8 is the same in those with and without a history of gonorrhoea in the population

H_1: the seroprevalence is not the same in the two groups in the population.

2 The observed frequencies are shown in the following contingency table: 14/43 (32.6%) and 36/228 (15.8%) of those with and without a previous history of gonorrhoea are seropositive for HHV-8, respectively.

3 The expected frequencies are shown in the four cells of the contingency table.

The test statistic is

$$\chi^2 = \left\{\frac{(|14-7.93|-\frac{1}{2})^2}{7.93} + \frac{(|36-42.07|-\frac{1}{2})^2}{42.07} + \frac{(|29-35.07|-\frac{1}{2})^2}{35.07} + \frac{(|192-185.93|-\frac{1}{2})^2}{185.93}\right\} = 5.70$$

4 We refer χ^2 to Appendix A3 with 1 *df*: $0.01 < P < 0.05$ (computer output gives $P = 0.017$).

5 There is evidence of a real difference in the seroprevalence in the two groups in the population. We estimate this difference as 32.6% − 15.8% = 16.8%. The 95% CI for the true difference in the two percentages is 2.0% to 31.6%
i.e. $16.8 \pm 1.96 \times \sqrt{(\{32.6 \times 67.4\}/43 + \{15.8 \times 84.2\}/228)}$.

	Previous history of gonorrhoea				
	Yes		No		
HHV-8	Observed	Expected	Observed	Expected	Total observed
Seropositive	14	$(43 \times 50/271) = 7.93$	36	$(228 \times 50/271) = 42.07$	50
Seronegative	29	$(43 \times 221/271) = 35.07$	192	$(228 \times 221/271) = 185.93$	221
Total	43		228		271

Example 2

In order to compare two methods of establishing the cavity status (present or absent) of teeth, a dentist assessed the condition of 100 first permanent molar teeth that had either tiny or no cavities using radiographic techniques. These results were compared with those obtained using the more objective approach of visually assessing a section of each tooth. The percentages of teeth detected as having cavities by the two methods of assessment were compared using **McNemar's test**.

1 H_0: the two methods of assessment identify the same percentage of teeth with cavities in the population

H_1: these percentages are not equal.

2 The frequencies for the matched pairs are displayed in the table:

	Radiographic diagnosis		
Diagnosis on section	Cavities absent	Cavities present	Total
Cavities absent	45	4	49
Cavities present	17	34	51
Total	62	38	100

3 Test statistic, $\chi^2 = \dfrac{(|17-4|-1)^2}{17+4} = 6.86$

4 We refer χ^2 to Appendix A3 with 1 degree of freedom: $0.001 < P < 0.01$ (computer output gives $P = 0.009$).

5 There is substantial evidence to reject the null hypothesis that the same percentage of teeth are detected as having cavities using the two methods of assessment. The radiographic method has a tendency to fail to detect cavities. We estimate the difference in percentages of teeth detected as having cavities as 51% − 38% = 13%. An approximate confidence interval for the true difference in the percentages is given by 4.4% to 21.6%

$$\left(\text{i.e. } \left\{\frac{|17-4|}{100} \pm \frac{1.96}{100} \times \sqrt{(17+4) - \frac{(17-4)^2}{100}}\right\} \times 100\%\right).$$

Adapted from Ketley, C.E. & Holt, R.D. (1993) Visual and radiographic diagnosis of occlusal caries in first permanent molars and in second primary molars. *British Dental Journal*, **174**, 364–370.

25 Categorical data: more than two categories

Chi-squared test: large contingency tables

The problem

Individuals can be classified by two factors. For example, one factor may represent disease severity (mild, moderate or severe) and the other factor may represent blood group (A, B, O, AB). We are interested in whether the two factors are associated. Are individuals of a particular blood group likely to be more severely ill?

Assumptions

The data may be presented in an $r \times c$ contingency table with r rows and c columns (Table 25.1). The entries in the table are **frequencies**; each cell contains the number of individuals in a particular row and a particular column. Every individual is represented once, and can only belong in one row and in one column, i.e. the categories of each factor are mutually exclusive. At least 80% of the expected frequencies are greater than or equal to 5.

Rationale

The null hypothesis is that there is no association between the two factors. Note that if there are only two rows and two columns, then this test of no association is the same as that of two proportions (Chapter 24). We calculate the frequency that we expect in each cell of the contingency table if the null hypothesis is true. As explained in Chapter 24, the expected frequency in a particular cell is the product of the relevant row total and relevant column total, divided by the overall total. We calculate a test statistic that focuses on the discrepancy between the observed and expected frequencies in every cell of the table. If the overall discrepancy is large, then it is unlikely the null hypothesis is true.

1 Define the null and alternative hypotheses under study

H_0: there is no association between the categories of one factor and the categories of the other factor in the population

H_1: the two factors are associated in the population.

2 Collect relevant data from a sample of individuals

3 Calculate the value of the test statistic specific to H_0

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O and E are the observed and expected frequencies in each cell of the table. The test statistic follows the Chi-squared distribution with degrees of freedom equal to $(r-1) \times (c-1)$.

Because the approximation to the Chi-squared distribution is reasonable if the degrees of freedom are greater than one, we do not need to include a continuity correction (as we did in Chapter 24).

4 Compare the value of the test statistic to values from a known probability distribution

Refer χ^2 to Appendix A3.

5 Interpret the *P*-value and results

If the assumptions are not satisfied

If more than 20% of the expected frequencies are less than 5, we try to combine, appropriately (i.e. so that it makes scientific sense), two or more rows and/or two or more columns of the contingency table. We then recalculate the expected frequencies of this reduced table, and carry on reducing the table, if necessary, to ensure that the $E \geq 5$ condition is satisfied. If we have reduced our table to a 2×2 table so that it can be reduced no further and we still have small expected frequencies, we use Fisher's exact test (Chapter 24) to evaluate the exact P-value. Some computer packages will compute Fisher's exact P-values for larger contingency tables.

Chi-squared test for trend

The problem

Sometimes we investigate relationships in categorical data when one of the two factors has only two categories (e.g. the presence or absence of a characteristic) and the second factor can be categorized into k, say, mutually exclusive categories that are ordered in some sense. For example, one factor might be whether or not an individual responds to treatment, and the ordered categories of the other factor may represent four different age (in years) categories 65–69, 70–74, 75–79 and ≥80. We can then assess whether there is a trend in the proportions with the characteristic over the categories of the second factor. For example, we may wish to know if the proportion responding to treatment tends to increase (say) with increasing age.

Table 25.1 Observed frequencies in an $r \times c$ table.

Row categories	Col 1	Col 2	Col 3	...	Col c	Total
Row 1	f_{11}	f_{12}	f_{13}	...	f_{1c}	R_1
Row 2	f_{21}	f_{22}	f_{23}	...	f_{2c}	R_2
Row 3	f_{31}	f_{32}	f_{33}	...	f_{3c}	R_3
...	...	...	...	...	...	...
...	...	...	...	...	...	...
Row r	f_{r1}	f_{r2}	f_{r3}	...	f_{rc}	R_r
Total	C_1	C_2	C_3	...	C_c	n

Table 25.2 Observed frequencies and assigned scores in a $2 \times k$ table.

Characteristic	Col 1	Col 2	Col 3	...	Col k	Total
Present	f_{11}	f_{12}	f_{13}	...	f_{1k}	R_1
Absent	f_{21}	f_{22}	f_{23}	...	f_{2k}	R_2
Total	C_1	C_2	C_3	...	C_k	n
Score	w_1	w_2	w_3	...	w_k	

1 Define the null and alternative hypotheses under study

H_0: there is no trend in the proportions with the characteristic in the population

H_1: there is a trend in the proportions in the population.

2 Collect relevant data from a sample of individuals

We estimate the proportion with the characteristic in each of the k categories. We assign a score to each of the column categories (Table 25.2). Typically these are the successive values, 1, 2, 3, . . . , k, but, depending on how we have classified the column factor, they could be numbers that in some way suggest the relative values of the ordered categories (e.g. the midpoint of the age range defining each category) or the trend we wish to investigate (e.g. linear or quadratic). The use of any equally spaced numbers (e.g. 1, 2, 3, . . . , k) allows us to investigate a *linear trend*.

3 Calculate the value of the test statistic specific to H_0

$$\chi^2 = \frac{\left(\sum w_i f_{1i} - R_1 \sum \frac{w_i C_i}{n}\right)^2}{\frac{R_1}{n}\left(1 - \frac{R_1}{n}\right)\left(\sum C_i w_i^2 - n\left(\sum \frac{w_i C_i}{n}\right)^2\right)}$$

using the notation of Table 25.2, and where the sums extend over all the k categories. The test statistic follows the Chi-squared distribution with 1 degree of freedom.

4 Compare the value of the test statistic to values from a known probability distribution

Refer χ^2 to Appendix A3.

5 Interpret the *P*-value and results

Interpret the P-value and calculate a confidence interval for each of the k proportions (Chapter 11).

Example

A cross-sectional survey was carried out among the elderly population living in Southampton, with the objective of measuring the frequency of cardiovascular disease. A total of 259 individuals, ranging between 65 and 95 years of age, were interviewed. Individuals were grouped into four age groups (65–69, 70–74, 75–79 and 80+ years) at the time of interview. We used the **Chi-squared test** to determine whether the prevalence of chest pain differed in the four age groups.

1 H_0: there is no association between age and chest pain in the population

H_1: there is an association between age and chest pain in the population.

2 The observed frequencies (%) and expected frequencies are shown in the following table.

3 Test statistic, $\chi^2 = \left[\frac{(15-9.7)^2}{9.7} + \ldots + \frac{(41-39.1)^2}{39.1}\right]$

$= 4.839$

4 We refer χ^2 to Appendix A3 with 3 degrees of freedom: $P > 0.10$ (computer output gives $P = 0.18$).

5 There is insufficient evidence to reject the null hypothesis of no association between chest pain and age in the population of elderly people. The estimated proportions (95% confidence intervals) with chest pain for the four successive age groups, starting with the youngest, are: 0.20 (0.11, 0.29), 0.12 (0.04, 0.19), 0.10 (0.02, 0.17) and 0.09 (0.02, 0.21).

	Age (years)				
Chest pain	65–69	70–74	75–79	80+	Total
Yes					
Observed	15 (20.3%)	9 (11.5%)	6 (9.7%)	4 (8.9%)	34
Expected	9.7	10.2	8.1	5.9	
No					
Observed	59 (79.7%)	69 (88.5%)	56 (90.3%)	41 (91.1%)	225
Expected	64.3	67.8	53.9	39.1	
Total	74	78	62	45	259

continued

As the four age groups in this study are ordered, it is also possible to analyse these data using a **Chi-squared test for trend**, which takes into account the ordering of the groups. We may obtain a significant result from this test, even though the general test of association gave a non-significant result. We assign the scores of 1, 2, 3 and 4 to each of the four age groups, respectively, and because of their even spacing, can test for a linear trend.

1 H_0: there is no linear association between age and chest pain in the population

H_1: there is a linear association between age and chest pain in the population.

2 The data are displayed in the previous table. We assign scores of 1, 2, 3 and 4 to the four age groups, respectively.

3 Test statistic is χ^2.

4 We refer χ^2 to Appendix A3 with 1 degree of freedom: $0.05 < P < 0.10$ (computer output gives $P = 0.052$).

5 There is insufficient evidence to reject the null hypothesis of no linear association between chest pain and age in the population of elderly people. However, the P-value is very close to 0.05 and there is a suggestion that the proportion of elderly people with chest pain decreases with increasing age.

$$\chi^2 = \frac{\left\{[(1\times 15)+\ldots+(4\times 4)]-34\times\left[\left(\frac{1\times 74}{259}\right)+\ldots+\left(\frac{4\times 45}{259}\right)\right]\right\}^2}{\frac{34}{259}\times\left(1-\frac{34}{259}\right)\times\left\{[(74\times 1^2)+\ldots+(45\times 4^2)]-259\times\left[\left(\frac{1\times 74}{259}\right)+\ldots+\left(\frac{4\times 45}{259}\right)\right]^2\right\}} = 3.79$$

Adapted from: Dewhurst, G., Wood, D.A., Walker, F., *et al*. (1991) A population survey of cardiovascular disease in elderly people: design, methods and prevalence results. *Age and Ageing* **20**, 353–360.

26 Correlation

Introduction

Correlation analysis is concerned with measuring the degree of association between two variables, x and y. Initially, we assume that both x and y are **numerical**, e.g. height and weight.

Suppose we have a pair of values, (x, y), measured on each of the n individuals in our sample. We can mark the point corresponding to each individual's pair of values on a two-dimensional **scatter diagram** (Chapter 4). Conventionally, we put the x variable on the horizontal axis, and the y variable on the vertical axis in this diagram. Plotting the points for all n individuals, we obtain a scatter of points that may suggest a relationship between the two variables.

Pearson correlation coefficient

We say that we have a **linear relationship** between x and y if a straight line drawn through the midst of the points provides the most appropriate approximation to the observed relationship. We measure how close the observations are to the straight line that best describes their linear relationship by calculating the **Pearson product moment correlation coefficient**, usually simply called the **correlation coefficient**. Its true value in the *population*, ρ (the Greek letter, rho), is estimated in the *sample* by r, where

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

which is usually obtained from computer output.

Properties

- r ranges from -1 to $+1$.
- Its **sign** indicates whether one variable increases as the other variable increases (positive r) or whether one variable decreases as the other increases (negative r) (see Fig. 26.1).
- Its **magnitude** indicates how close the points are to the straight line. In particular if $r = +1$ or -1, then there is perfect correlation with all the points lying on the line (this is most unusual, in practice); if $r = 0$, then there is no **linear** correlation (although there may be a non-linear relationship). The closer r is to the extremes, the greater the degree of linear association (Fig. 26.1).
- It is dimensionless, i.e. it has no units of measurement.
- Its value is valid only within the range of values of x and y in the sample. Its absolute value (ignoring sign) tends to increase as the range of values of x and/or y increases and therefore you cannot infer that it will have the same value when considering values of x or y that are more extreme than the sample values.
- x and y can be interchanged without affecting the value of r.
- A correlation between x and y does not necessarily imply a 'cause and effect' relationship.
- r^2 represents the proportion of the variability of y that can be attributed to its linear relationship with x (Chapter 28).

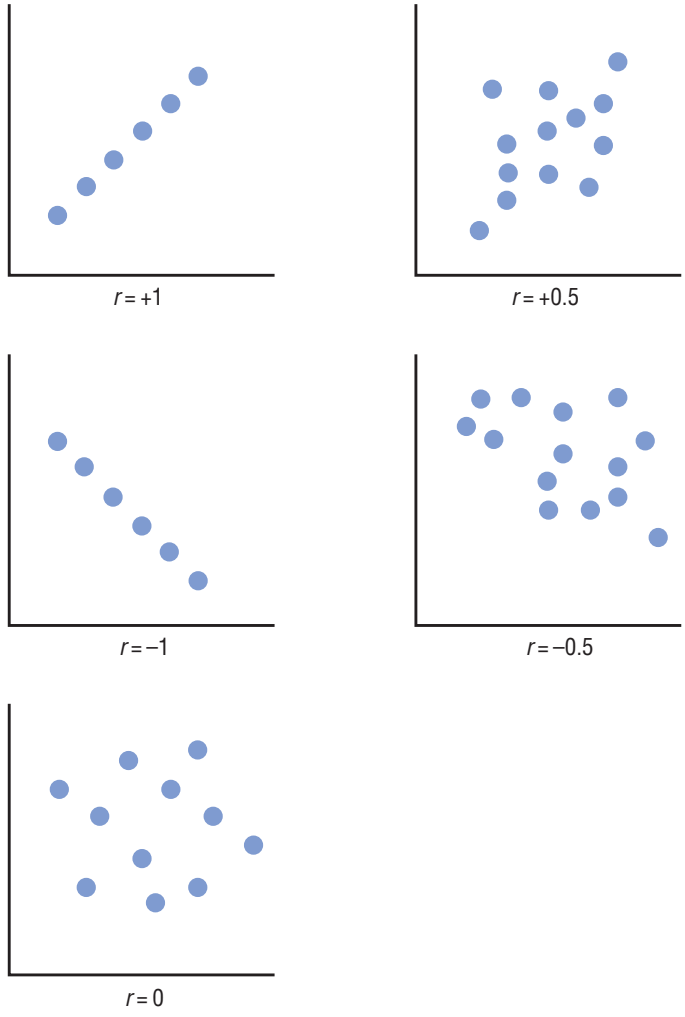


Figure 26.1 Five diagrams indicating values of r in different situations.

When not to calculate *r*

It may be misleading to calculate *r* when:

- there is a non-linear relationship between the two variables (Fig. 26.2a), e.g. a quadratic relationship (Chapter 33);
- the data include more than one observation on each individual;
- one or more outliers are present (Fig. 26.2b);
- the data comprise subgroups of individuals for which the mean levels of the observations on at least one of the variables are different (Fig. 26.2c);

Hypothesis test for the Pearson correlation coefficient

We want to know if there is any linear correlation between two numerical variables. Our sample consists of n independent pairs of values of x and y. We assume that at least one of the two variables is Normally distributed.

1 Define the null and alternative hypotheses under study

H_0: $\rho = 0$

H_1: $\rho \neq 0$

2 Collect relevant data from a sample of individuals

3 Calculate the value of the test statistic specific to H_0

Calculate r.

- If $n \leq 150$, r is the test statistic
- If $n > 150$, calculate $T = \sqrt{\frac{(n-2)}{(1-r^2)}}$

which follows a t-distribution with $n-2$ degrees of freedom.

4 Compare the value of the test statistic to values from a known probability distribution

- If $n \leq 150$, refer r to Appendix A10
- If $n > 150$, refer T to Appendix A2.

5 Interpret the *P*-value and results

Calculate a confidence interval for ρ. Provided *both variables are approximately Normally distributed*, the approximate 95% confidence interval for ρ is:

$$\left(\frac{e^{2z_1}-1}{e^{2z_1}+1}, \frac{e^{2z_2}-1}{e^{2z_2}+1}\right)$$

where $z_1 = z - \frac{1.96}{\sqrt{n-3}}$, $z_2 = z + \frac{1.96}{\sqrt{n-3}}$,

and $z = 0.5\log_e\left[\frac{(1+r)}{(1-r)}\right]$.

Note that, if the sample size is large, H_0 may be rejected even if r is quite close to zero. Alternatively, even if r is large, H_0 may not be rejected if the sample size is small. For this reason, it is particularly helpful to calculate r^2, the proportion of the total variance of one variable explained by its linear relationship with the other. For example, if $r = 0.40$ then $P < 0.05$ for a sample size of 25, but the relationship is only explaining 16% ($= 0.40^2 \times 100$) of the variability of one variable.

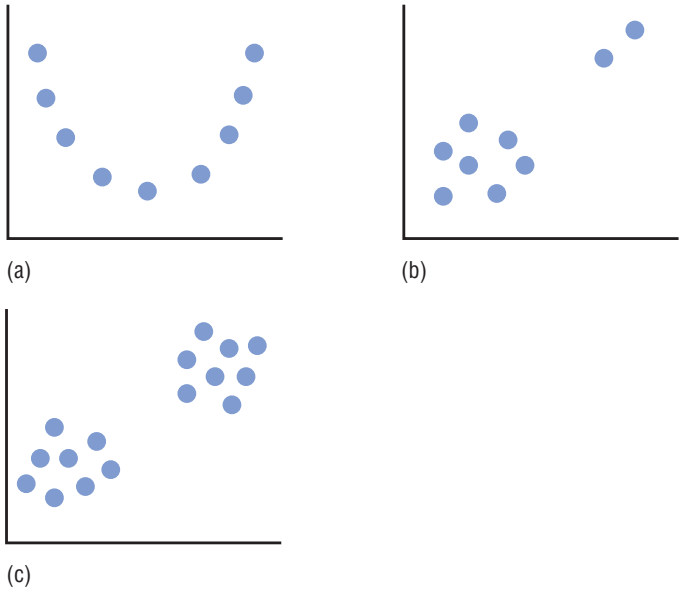


Figure 26.2 Diagrams showing when it is inappropriate to calculate the correlation coefficient. (a) Relationship not linear, $r = 0$. (b) In the presence of outlier(s). (c) Data comprise subgroups.

Spearman's rank correlation coefficient

We calculate **Spearman's rank correlation coefficient**, the non-parametric equivalent to Pearson's correlation coefficient, if one or more of the following points is true:

- at least one of the variables, x or y, is measured on an ordinal scale;
- neither x nor y is Normally distributed;
- the sample size is small;
- we require a measure of the association between two variables when their relationship is non-linear.

Calculation

To estimate the population value of Spearman's rank correlation coefficient, ρ_s, by its sample value, r_s:

1 Arrange the values of x in increasing order, starting with the smallest value, and assign successive *ranks* (the numbers 1, 2, 3, . . . , n) to them. Tied values receive the mean of the ranks these values would have received had there been no ties.

2 Assign ranks to the values of y in a similar manner.

3 r_s is the Pearson correlation coefficient between the *ranks* of x and y.

Properties and hypothesis tests

These are the same as for Pearson's correlation coefficient, replacing r by r_s, except that:

- r_s provides a measure of association (not necessarily linear) between x and y;
- when testing the null hypothesis that $\rho_s = 0$, refer to Appendix A11 if the sample size is less than or equal to 10;
- we do not calculate r_s^2 (it does not represent the proportion of the total variation in one variable that can be attributed to its linear relationship with the other).

Example

As part of a study to investigate the factors associated with changes in blood pressure in children, information was collected on demographic and lifestyle factors, and clinical and anthropometric measures in 4245 children aged from 5 to 7 years. The relationship between height (cm) and systolic blood pressure (mmHg) in a sample of 100 of these children is shown in the scatter diagram (Fig. 28.1); there is a tendency for taller children in the sample to have higher blood pressures. **Pearson's correlation coefficient** between these two variables was investigated. Appendix C contains a computer output from the analysis.

1 H_0: the population value of the Pearson correlation coefficient, ρ, is zero

H_1: the population value of the Pearson correlation coefficient is not zero.

2 We can show (Fig. 37.1) that the sample values of both height and systolic blood pressure are approximately Normally distributed.

3 We calculate r as 0.33. This is the test statistic since $n \leq 150$.

4 We refer r to Appendix A10 with a sample size of 100: $P < 0.001$.

5 There is strong evidence to reject the null hypothesis; we conclude that there is a linear relationship between systolic blood pressure and height in the population of such children. However, $r^2 = 0.33 \times 0.33 = 0.11$. Therefore, despite the highly significant result, the relationship between height and systolic blood pressure explains only a small percentage, 11%, of the variation in systolic blood pressure.

In order to determine the 95% confidence interval for the true correlation coefficient, we calculate:

$$z = 0.5\ln\left(\frac{1.33}{0.67}\right) = 0.3428$$

$$z_1 = 0.3428 - \frac{1.96}{9.849} = 0.1438$$

$$z_2 = 0.3428 + \frac{1.96}{9.849} = 0.5418$$

Thus the confidence interval ranges from

$$\frac{(e^{2\times 0.1438} - 1)}{(e^{2\times 0.1438} + 1)} \text{ to } \frac{(e^{2\times 0.5418} - 1)}{(e^{2\times 0.5418} + 1)}, \text{ i.e. from } \frac{0.33}{2.33} \text{ to } \frac{1.96}{3.96}.$$

We are thus 95% certain that ρ lies between 0.14 and 0.49.

As we might expect, given that each variable is Normally distributed, **Spearman's rank correlation coefficient** between these variables gave a comparable estimate of 0.32. To test H_0: $\rho_s = 0$, we refer this value to Appendix A10 and again find $P < 0.001$.

Data kindly provided by Ms O. Papacosta and Dr P. Whincup, Department of Primary Care and Population Sciences, Royal Free and University College Medical School, London, UK.

27 The theory of linear regression

What is linear regression?

To investigate the relationship between two numerical variables, x and y, we measure the values of x and y on each of the n individuals in our sample. We plot the points on a **scatter diagram** (Chapters 4 and 26), and say that we have a **linear** relationship if the data approximate a straight line. If we believe y is dependent on x, with a change in y being attributed to a change in x, rather than the other way round, we can determine the **linear regression line** (the **regression of y on x**) that best describes the straight line relationship between the two variables. In general, we describe the regression as **univariable** because we are concerned with only one x variable in the analysis; this contrasts with *multivariable* regression which involves two or more x's (see Chapters 29–31).

The regression line

The mathematical equation which estimates the **simple linear regression** line is:

$$Y = a + bx$$

- x is called the **independent**, **predictor** or **explanatory** variable;
- for a given value of x, Y is the value of y (called the **dependent**, **outcome** or **response** variable), which lies on the estimated line. It is an estimate of the value we *expect* for y (i.e. its mean) if we know the value of x, and is called the **fitted** value of y;
- a is the **intercept** of the estimated line; it is the value of Y when $x = 0$ (Fig. 27.1);
- b is the **slope** or **gradient** of the estimated line; it represents the amount by which Y increases on average if we increase x by one unit (Fig. 27.1).

a and b are called the **regression coefficients** of the estimated line, although this term is often reserved only for b. We show how to evaluate these coefficients in Chapter 28. Simple linear regression can be extended to include more than one explanatory variable; in this case, it is known as **multiple linear regression** (Chapter 29).

Method of least squares

We perform regression analysis using a sample of observations. a and b are the sample estimates of the true parameters, α and β, which define the linear regression line in the population. a and b are determined by the **method of least squares** (often called ordinary least squares, OLS) in such a way that the 'fit' of the line $Y = a + b\,x$ to the points in the scatter diagram is optimal. We assess this by considering the **residuals** (the vertical distance of each point from the line, i.e. **residual = observed y − fitted Y**, Fig. 27.2). The **line of best fit** is chosen so that the sum of the *squared* residuals is a *minimum*.

Assumptions

1 There is a linear relationship between x and y

2 The observations in the sample are independent. The observations are independent if there is no more than one pair of observations on each individual.

3 For each value of x, there is a distribution of values of y in the population; this distribution is Normal. The mean of this distribution of y values lies on the true regression line (Fig. 27.3).

4 The variability of the distribution of the y values in the population is the same for all values of x, i.e. the variance, σ^2, is constant (Fig. 27.3).

5 The x variable can be measured without error. Note that we do not make any assumptions about the distribution of the x variable.

Many of the assumptions which underlie regression analysis relate to the distribution of the y population for a specified value of x, but they may be framed in terms of the residuals. It is easier to check the assumptions (Chapter 28) by studying the residuals than the values of y.

Analysis of variance table

Description

Usually the computer output in a regression analysis contains an **analysis of variance table**. In analysis of variance (Chapter 22), the total variation of the variable of interest, in this case 'y', is partitioned into its component parts. Because of the linear relationship of y on x, we expect y to vary as x varies; we call this the variation which is **due to** or **explained by the regression**. The remaining variability is called the **residual error** or **unexplained** variation. The residual variation should be as small as possible; if so, most of the variation in y will be explained by the regression, and the points will lie close to or on the line; i.e. the line is a **good fit**.

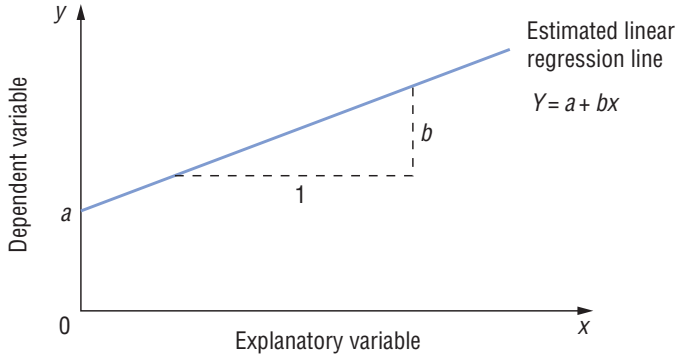


Figure 27.1 Estimated linear regression line showing the intercept, a, and the slope, b (the mean increase in Y for a unit increase in x).

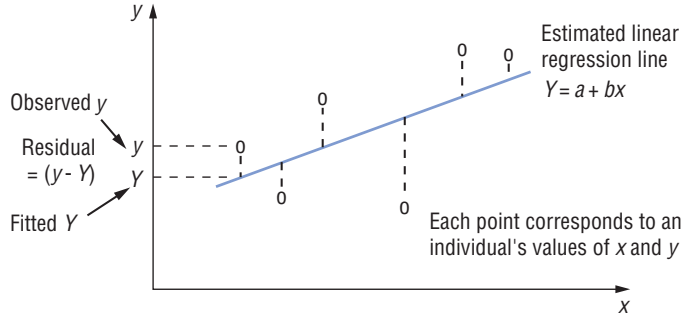


Figure 27.2 Estimated linear regression line showing the residual (vertical dotted line) for each point.

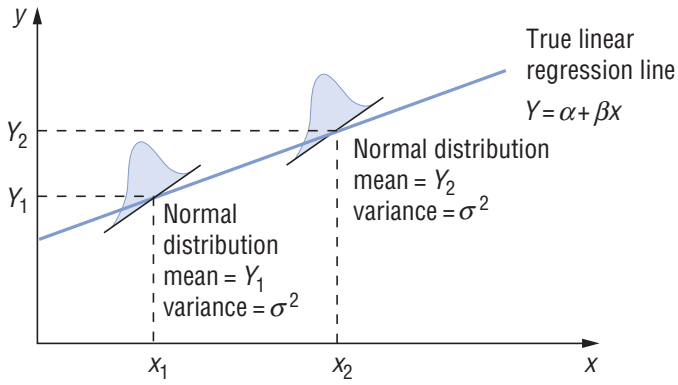


Figure 27.3 Illustration of assumptions made in linear regression.

Purposes

The analysis of variance table enables us to do the following.

1 Assess how well the line fits the data points. From the information provided in the table, we can calculate the proportion of the total variation in y that is explained by the regression. This proportion, usually expressed as a percentage and denoted by R^2 (in simple linear regression it is r^2, the square of the correlation coefficient; Chapter 26), allows us to assess subjectively the **goodness of fit** of the regression equation.

2 Test the **null hypothesis** that the true slope of the line, β, is zero; a significant result indicates that there is evidence of a linear relationship between x and y.

3 Obtain an estimate of the **residual variance**. We need this for testing hypotheses about the slope or the intercept, and for calculating confidence intervals for these parameters and for predicted values of y.

We provide details of the more common procedures in Chapter 28.

Regression to the mean

The statistical use of the word 'regression' derives from a phenomenon known as **regression to the mean**, attributed to Sir Francis Galton in 1889. He demonstrated that although tall fathers tend to have tall sons, the average height of the sons is less than that of their tall fathers. The average height of the sons has 'regressed' or 'gone back' towards the mean height of all the fathers in the population. So, on average, tall fathers have shorter (but still tall) sons and short fathers have taller (but still short) sons.

We observe regression to the mean in **screening** (Chapter 38) and in **clinical trials** (Chapter 14), when a subgroup of patients may be selected for treatment because their levels of a certain variable, say cholesterol, are extremely high (or low). If the measurement is repeated some time later, the average value for the second reading for the subgroup is usually less than that of the first reading, tending towards (i.e. regressing to) the average of the age- and sex-matched population, irrespective of any treatment they may have received. Patients recruited into a clinical trial on the basis of a high cholesterol level on their first examination are thus likely to show a drop in cholesterol levels on average at their second examination, even if they remain untreated during this period.

28 Performing a linear regression analysis

The linear regression line

After selecting a sample of size n from our population, and drawing a **scatter diagram** to confirm that the data approximate a straight line, we estimate the **regression of y on x** as:

$$Y = a + bx$$

where Y is the estimated fitted or predicted value of y, a is the estimated intercept, and b is the estimated slope that represents the average change in Y for a unit change in x (Chapter 27).

Drawing the line

To draw the line $Y = a + bx$ on the scatter diagram, we choose three values of x (i.e. x_1, x_2 and x_3) along its range. We substitute x_1 in the equation to obtain the corresponding value of Y, namely $Y_1 = a + bx_1$; Y_1 is our estimated **fitted** value for x_1 which corresponds to the **observed** value, y_1. We repeat the procedure for x_2 and x_3 to obtain the corresponding values of Y_2 and Y_3. We plot these points on the scatter diagram and join them to produce a straight line.

Checking the assumptions

For each observed value of x, the **residual** is the observed y minus the corresponding fitted Y. Each residual may be either positive or negative. We can use the residuals to check the following assumptions underlying linear regression.

1 There is a linear relationship between x and y: *Either* plot y against x (the data should approximate a straight line), *or* plot the residuals against x (we should observe a random scatter of points rather than any systematic pattern).

2 The observations are independent: the observations are independent if there is no more than one pair of observations on each individual.

3 The residuals are Normally distributed with a mean of zero: Draw a histogram, stem-and-leaf plot, box-and-whisker plot (Chapter 4) or Normal plot (Chapter 35) of the residuals and 'eyeball' the result.

4 The residuals have the same variability (constant variance) for all the fitted values of y: Plot the residuals against the fitted values, Y, of y; we should observe a random scatter of points. If the scatter of residuals progressively increases or decreases as Y increases, then this assumption is not satisfied.

5 The x variable can be measured without error.

Failure to satisfy the assumptions

If the linearity, Normality and/or constant variance assumptions are in doubt, we may be able to transform x or y (Chapter 9), and calculate a new regression line for which these assumptions are satisfied. It is not always possible to find a satisfactory transformation. The linearity and independence assumptions are the most important. If you are dubious about the Normality and/or constant variance assumptions, you may proceed, but the P-values in your hypothesis tests, and the estimates of the standard errors, may be affected. Note that the x variable is rarely measured without any error; provided the error is small, this is usually acceptable because the effect on the conclusions is minimal.

Outliers and influential points

- An **influential** observation will, if omitted, alter one or more of the parameter estimates (i.e. the slope or the intercept) in the model. Formal methods of detection are discussed briefly in Chapter 29. If these methods are not available, you may have to rely on intuition.
- An **outlier** (an observation that is inconsistent with most of the values in the data set (Chapter 3)) may or may not be an influential point, and can often be detected by looking at the scatter diagram or the residual plots (see also Chapter 29). For both outliers and influential points, we fit the model with and without the suspect individual's data and note the effect on the estimate(s). Do not discard outliers or influential points routinely because their omission may affect your conclusions. Always investigate the reasons for their presence and report them.

Assessing goodness of fit

We can judge how well the line fits the data by calculating $\boldsymbol{R^2}$ (usually expressed as a percentage), which is equal to the square of the correlation coefficient (Chapters 26 and 27). This represents the percentage of the variability of y that can be **explained** by its relationship with x. Its compliment, $(100 - R^2)$, represents the percentage of the variation in y that is **unexplained** by the relationship. There is no formal test to assess R^2; we have to relay on subjective judgement to evaluate the fit of the regression line.

Investigating the slope

If the slope of the line is zero, there is no linear relationship between x and y: changing x has no effect on y. There are two approaches, with identical results, to **testing the null hypothesis that the true slope, β, is zero**.

- *Examine the F-ratio* (equal to the ratio of the 'explained' to the 'unexplained' mean squares) in the analysis of variance table. It follows the F-distribution and has $(1, n - 2)$ degrees of freedom in the numerator and denominator, respectively.
- *Calculate the test statistic* $= \dfrac{b}{SE(b)}$ which follows the t-distribution on $n - 2$ degrees of freedom, where SE(b) is the standard error of b.

In either case, a significant result, usually if $P < 0.05$, leads to rejection of the null hypothesis.

We calculate the **95% confidence interval** for β as $b \pm t_{0.05}\,\mathrm{SE}(b)$, where $t_{0.05}$ is the percentage point of the t-distribution with $n - 2$ degrees of freedom which gives a two-tailed probability of 0.05. It is the interval that contains the true slope with 95% certainty. For large samples, say $n \geq 100$, we can approximate $t_{0.05}$ by 1.96.

Regression analysis is rarely performed by hand; computer output from most statistical packages will provide all of this information.

Using the line for prediction

We can use the regression line for predicting values of y for values of x within the observed range (never extrapolate beyond these limits). We predict the mean value of y for individuals who have a certain value of x by substituting that value of x into the equation of the line. So, if $x = x_0$, we predict y as $Y_0 = a + bx_0$. We use this estimated predicted value, and its standard error, to evaluate the confidence interval for the true mean value of y in the population. Repeating this procedure for various values of x allows us to construct confidence limits for the line. This is a band or region that contains the true line with, say, 95% certainty. Similarly, we can calculate a wider region within which we expect most (usually 95%) of the *observations* to lie.

Useful formulae for hand calculations

$$\bar{x} = \sum x/n \quad \text{and} \quad \bar{y} = \sum y/n$$

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$s^2_{\text{res}} = \frac{\sum (y-Y)^2}{(n-2)}, \text{ the estimated residual variance}$$

$$SE(b) = \frac{s_{\text{res}}}{\sqrt{\sum (x-\bar{x})^2}}$$

Example

The relationship between height (measured in cm) and systolic blood pressure (SBP, measured in mmHg) in the 100 children described in Chapter 26 is shown in Fig. 28.1. We performed a **simple linear regression analysis** of systolic blood pressure on height. Assumptions underlying this analysis are verified in Figs 28.2 to 28.4. A typical computer output is shown in Appendix C. There is a significant linear relationship between height and systolic blood pressure, as can be seen by the significant F-ratio in the analysis of variance table in Appendix C ($F = 12.03$ with 1 and 98 degrees of freedom in the numerator and denominator, respectively, $P = 0.0008$). The R^2 of the model is 10.9%. Only approximately a tenth of the variability in the systolic blood pressure can thus be explained by the model; that is, by differences in the heights of the children. The computer output provides the information shown in the table.

Variable	Parameter estimate	Standard Error	Test statistic	P-value
Intercept	46.2817	16.7845	2.7574	0.0070
Height	0.4842	0.1396	3.4684	0.0008

The parameter estimate for 'Intercept' corresponds to a, and that for 'Height' corresponds to b (the slope of the regression line). So, the equation of the estimated regression line is:

$$\text{SBP} = 46.28 + 0.48 \times \text{height}$$

In this example, the intercept is of no interest in its own right (it relates to the predicted blood pressure for a child who has a height of zero centimetres—clearly out of the range of values seen in the study). However, we can interpret the slope coefficient; in these children, systolic blood pressure is predicted to increase by 0.48 mmHg, on average, for each centimetre increase in height.

$P = 0.0008$ for the hypothesis test for height (i.e. H_0: true slope equals zero) is identical to that obtained from the analysis of variance table in Appendix C, as expected.

Since the sample size is large (it is 100), we can approximate $t_{0.05}$ by 1.96 and calculate the 95% confidence interval for the true slope as:

$$b \pm 1.96 \times SE(b) = 0.48 \pm (1.96 \times 0.14)$$

Therefore, the 95% confidence interval for the slope ranges from 0.21 to 0.75 mmHg per cm increase in height. This confidence interval does not include zero, confirming the finding that the slope is significantly different from zero.

We can use the regression equation to predict the systolic blood pressure we expect a child of a given height to have. For example, a child who is 115 cm tall has an estimated predicted systolic blood pressure of $46.28 + (0.48 \times 115) = 101.48$ mmHg; a child who is 130 cm tall has an estimated predicted systolic blood pressure of $46.28 + (0.48 \times 130) = 108.68$ mmHg.

continued

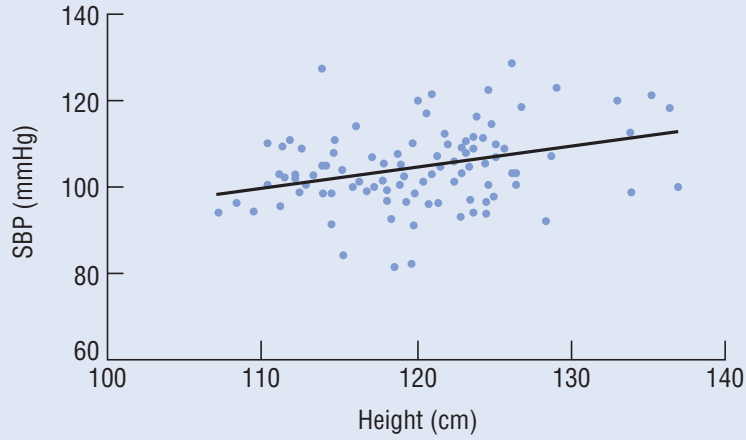


Figure 28.1 Scatter plot showing relationship between systolic blood pressure (SBP) and height. The estimated regression line, SBP = 46.28 + 0.48 × height, is marked on the scatter plot.

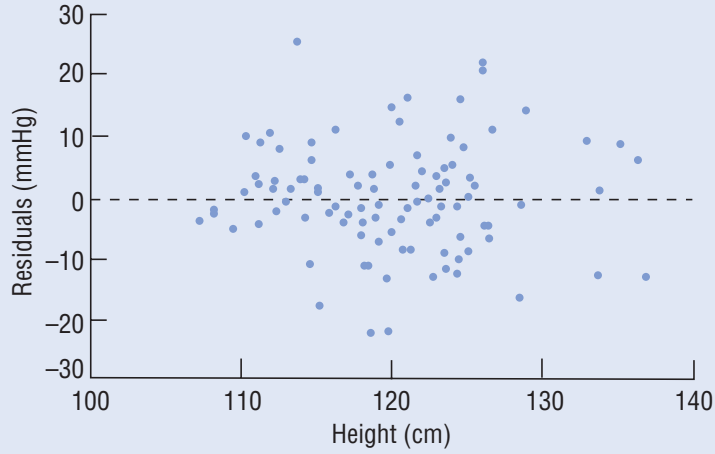


Figure 28.2 No relationship is apparent in this diagram, indicating that a linear relationship between height and systolic blood pressure is appropriate.

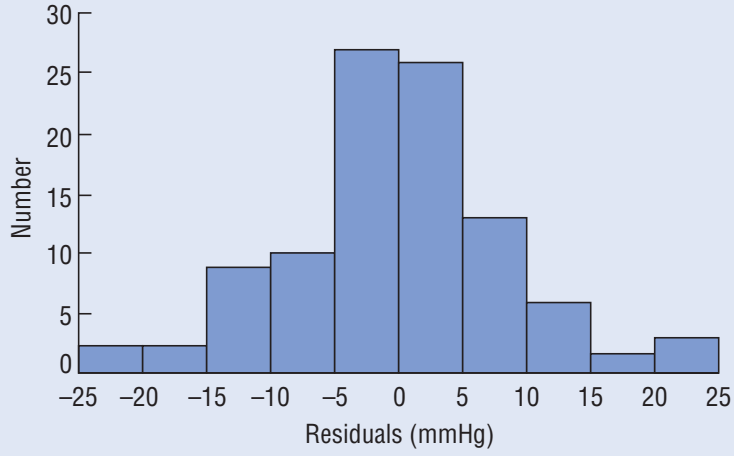


Figure 28.3 The distribution of the residuals is approximately Normal.

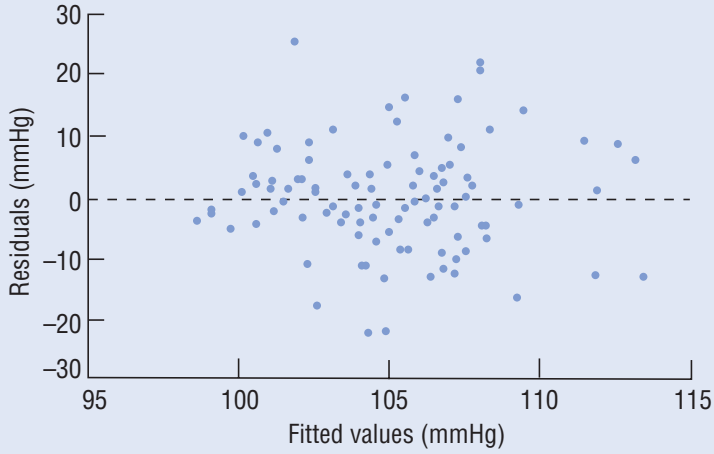


Figure 28.4 There is no tendency for the residuals to increase or decrease systematically with the fitted values. Hence the constant variance assumption is satisfied.

29 Multiple linear regression

What is it?

We may be interested in the effect of several explanatory variables, $x_1, x_2, \ldots, x_k$, on a response variable, y. If we believe that these x's may be inter-related, we should not look, in isolation, at the effect on y of changing the value of a single x, but should simultaneously take into account the values of the other x's. For example, as there is a strong relationship between a child's height and weight, we may want to know whether the relationship between height and systolic blood pressure (Chapter 28) is changed when we take the child's weight into account. Multiple linear regression allows us to investigate the joint effect of these explanatory variables on y; it is an example of a **multivariable** analysis where we relate a single outcome variable to two or more explanatory variables simultaneously. Note that, although the explanatory variables are sometimes called independent variables, this is a misnomer because they may be related.

We take a sample of n individuals, and measure the value of each of the variables on every individual. The multiple linear regression equation which estimates the relationships in the population is:

$$Y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

- x_i is the ith explanatory variable or **covariate** ($i = 1, 2, 3, \ldots, k$);
- Y is the estimated predicted, expected, mean or fitted value of y, which corresponds to a particular set of values of $x_1, x_2, \ldots, x_k$;
- a is a constant term, the estimated intercept; it is the value of Y when all the x's are zero;
- $b_1, b_2, \ldots, b_k$ are the estimated **partial regression coefficients**; b_1 represents the amount by which Y increases on average if we increase x_1 by one unit but keep all the other x's constant (i.e. **adjust** or **control** for them). If there is a relationship between x_1 and the other x's, b_1 differs from the estimate of the regression coefficient obtained by regressing y on only x_1, because the latter approach does not adjust for the other variables. b_1 represents the effect of x_1 on y that is **independent** of the other x's.

Invariably, you will perform a multiple linear regression analysis on the computer, and so we omit the formulae for these estimated parameters.

Why do it?

To be able to:

- identify explanatory variables that are associated with the dependent variable in order to promote understanding of the underlying process;
- determine the extent to which one or more of the explanatory variables is/are linearly related to the dependent variable, after adjusting for other variables that may be related to it; and,
- possibly, predict the value of the dependent variable as accurately as possible from the explanatory variables.

Assumptions

The assumptions in multiple linear regression are the same (if we replace 'x' by 'each of the x's') as those in simple linear regression (Chapter 27), and are checked in the same way. Failure to satisfy the linearity or independence assumptions is particularly important. We can transform (Chapter 9) the y variable and/or some or all of the x variables if the assumptions are in doubt, and then repeat the analysis (including checking the assumptions) on the transformed data.

Categorical explanatory variables

We can perform a multiple linear regression analysis using ***categorical*** explanatory variables. In particular, if we have a **binary** variable, x_1 (e.g. male = 0, female = 1), and we increase x_1 by one unit, we are 'changing' from males to females. b_1 thus represents the difference in the estimated mean values of y between females and males, after adjusting for the other x's.

If we have a **nominal** explanatory variable (Chapter 1) that has more than two categories of response, we have to create a number of **dummy** or **indicator** variables[1]. In general, for a nominal variable with k categories, we create $k - 1$ binary dummy variables. We choose one of the categories to represent our **reference category**, and each dummy variable allows us to compare one of the remaining $k - 1$ categories of the variable with the reference category. For example, we may be interested in comparing mean systolic blood pressure levels in individuals living in four countries in Europe (the Netherlands, UK, Spain and France). Suppose we choose our reference category to be the Netherlands. We generate one binary variable to identify those living in the UK; this variable takes the value 1 if the individual lives in the UK and 0 otherwise. We then generate binary variables to identify those living in Spain and France in a similar way. By default, those living in the Netherlands can then be identified since these individuals will have the value 0 for each of the three binary variables. In a multiple linear regression analysis, the regression coefficient for each of the other three countries represents the amount by which Y (systolic blood pressure) differs, on average, among those living in the relevant country *compared to* those living in the Netherlands. The intercept provides an estimate of the mean systolic blood pressure for those living in the Netherlands (when all of the other explanatory variables take the value zero). Some computer packages will create dummy variables automatically once you have specified that the variable is categorical.

If we have an **ordinal** explanatory variable and its three or more categories can be assigned values on a meaningful linear scale (e.g. social classes 1–5), then we can either use these values directly in the multiple linear regression equation (see also Chapter 33), or generate a series of dummy variables as for a nominal variable (but this does not make use of the ordering of the categories).

Analysis of covariance

An extension of analysis of variance (ANOVA, Chapter 22) is the **analysis of covariance**, in which we compare the response of interest between groups of individuals (e.g. two or more treatment groups) when other variables measured on each individual are taken into account. Such data can be analysed using multiple linear regression techniques by creating one or more dummy binary variables to differentiate between the groups. So, if we wish to compare the mean values of y in two treatment groups, while controlling for the effect of variables, $x_2, x_3, \ldots, x_k$ (e.g. age, weight, . . .), we

[1] Armitage, P., Berry, G. and Matthews, J.N.S. (2001) *Statistical Methods in Medical Research*, 4th edn. Blackwell Science (UK).

create a binary variable, x_1, to represent 'treatment' (e.g. $x_1 = 0$ for treatment A, $x_1 = 1$ for treatment B). In the multiple linear regression equation, b_1 is the estimated difference in the mean responses on y between treatments B and A, adjusting for the other x's.

Analysis of covariance is the preferred analysis for a randomized controlled trial comparing treatments when each individual in the study has a baseline and post-treatment follow-up measurement. In this instance the response variable, y, is the follow-up measurement and two of the explanatory variables in the regression model are a binary variable representing treatment, x_1, and the individual's baseline level at the start of the study, x_2. This approach is generally better (i.e. has a greater power—see Chapter 36) than using either the change from baseline or the percentage change from follow-up as the response variable.

Choice of explanatory variables

As a rule of thumb, we should not perform a multiple linear regression analysis if the number of variables is greater than the number of individuals divided by 10. Most computer packages have automatic procedures for selecting variables, e.g. stepwise selection (Chapter 33). These are particularly useful when many of the explanatory variables are related. A particular problem arises when **collinearity** is present, i.e. when pairs of explanatory variables are extremely highly correlated (Chapter 34).

Analysis

Most computer output contains the following items.

1 An assessment of goodness of fit

The **adjusted R^2** represents the proportion (often expressed as a percentage) of the variability of y which can be explained by its relationship with the x's. R^2 is adjusted so that models with different numbers of explanatory variables can be compared. If it has a low value (judged subjectively), the model is a poor fit. Goodness of fit is particularly important when we use the multiple linear regression equation for prediction.

2 The F-test in the ANOVA table

This tests the null hypothesis that all the partial regression coefficients in the population, $\beta_1, \beta_2, \ldots, \beta_k$, are zero. A significant result indicates that there is a linear relationship between y and at least one of the x's.

3 The t-test of each partial regression coefficient, β_i ($i = 1, 2, \ldots, k$)

Each t-test relates to one explanatory variable, and is relevant if we want to determine whether that explanatory variable affects the response variable, while controlling for the effects of the other covariates. To test H_0: $\beta_i = 0$, we calculate the test statistic $= \dfrac{b_i}{SE(b_i)}$, which follows the t-distribution with (n − *number of explanatory variables* − 1) degrees of freedom. Computer output includes the values of each b_i, SE(b_i) and the related test statistic with its P-value. Sometimes the 95% confidence interval for β_i is included; if not, it can be calculated as $b_i \pm t_{0.05}\,\mathrm{SE}(b_i)$.

Outliers and influential points

As discussed briefly in Chapter 28, an **outlier** (an observation that is inconsistent with most of the values in the data set (Chapter 3)) may or may not be **influential** (i.e. affect the parameter estimate(s) of the model if omitted). An outlier and/or influential observation may have one or both of the following:

- A large **residual** (a residual is the difference between the predicted and observed values of the outcome variable, y, for that individual's value(s) of the explanatory variable).
- High **leverage** when the individual's value of x (or set of x's) is a long way from the mean value of x (or set of x's). High leverage values may be taken as those greater than $2(k + 1)/n$ where k is the number of explanatory variables in the model and n is the number of individuals in the study.

Various methods are available for investigating model **sensitivity**—the extent to which estimates are affected by subsets of the data. We can determine suspect influential observations by, for example, (1) investigating those individuals having large residuals, high leverage and/or values of **Cook's distance** (an overall measure of influence incorporating both residual and leverage values) greater than one, or (2) examining special diagnostic plots in which influential points may become apparent.

Example

In Chapter 28, we studied the relationship between systolic blood pressure and height in 100 children. It is known that height and weight are positively correlated. We therefore performed a **multiple linear regression analysis** to investigate the effects of height (cm), weight (kg) and sex (0 = boy, 1 = girl) on systolic blood pressure (mmHg) in these children. Assumptions underlying this analysis are verified in Figs 29.1 to 29.4.

A typical output from a computer analysis of these data is contained in Appendix C. The analysis of variance table indicates that at least one of the explanatory variables is related to systolic blood pressure ($F = 14.95$ with 3 and 96 degrees of freedom in the numerator and denominator, respectively, $P = 0.0001$). The adjusted R^2 value of 0.2972 indicates that 29.7% of the variability in systolic blood pressure can be explained by the model—that is, by differences in the height, weight *and* sex of the children. Thus this provides a much better fit to the data than the simple linear regression in Chapter 28 in which $R^2 = 0.11$. Typical computer output contains the information in the following table about the explanatory variables in the model:

Variable	Parameter estimate	Standard error	95% CI for parameter	Test statistic	P-value
Intercept	79.4395	17.1182	(45.89 to 112.99)	4.6406	0.0001
Height	–0.0310	0.1717	(–0.37 to 0.31)	–0.1807	0.8570
Weight	1.1795	0.2614	(0.67 to 1.69)	4.5123	0.0001
Sex	4.2295	1.6105	(1.07 to 7.39)	2.6261	0.0101

continued

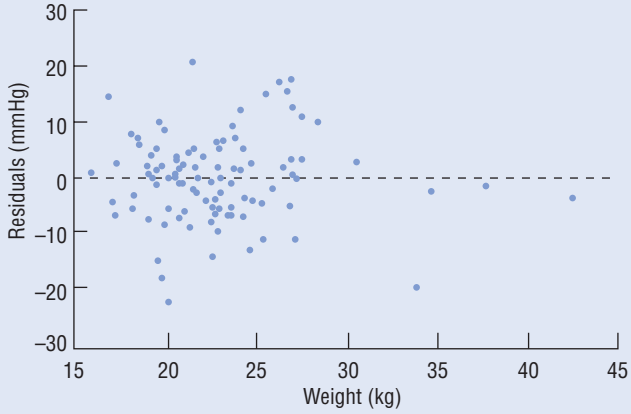


Figure 29.1 There is no systematic pattern to the residuals when plotted against weight. (Note that, similarly to Fig. 28.2, a plot of the residuals from this model against height also shows no systematic pattern).

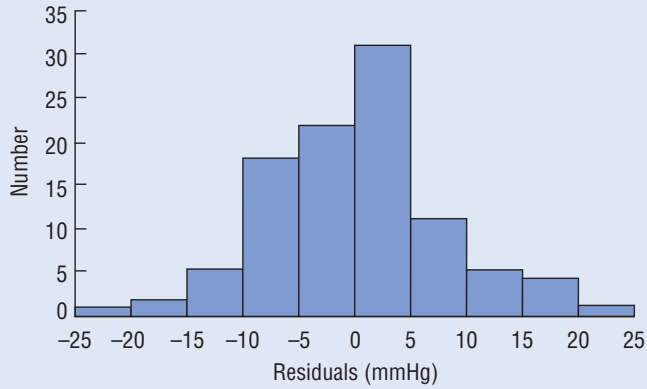


Figure 29.2 The distribution of the residuals is approximately Normal and the variance is slightly less than that from the simple regression model (Chapter 28), reflecting the improved fit of the multiple linear regression model over the simple model.

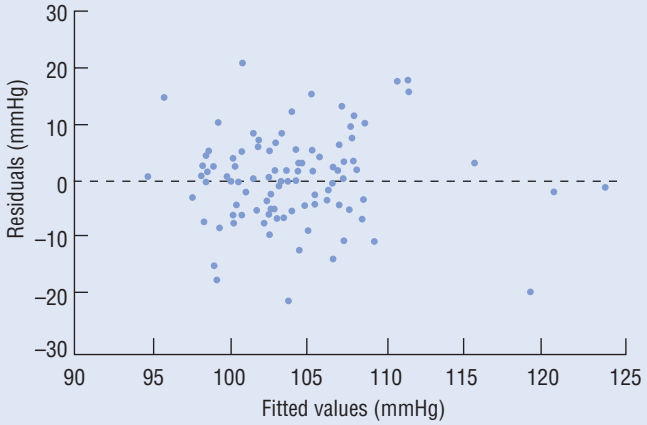


Figure 29.3 As with the univariable model, there is no tendency for the residuals to increase or decrease systematically with fitted values. Hence the constant variance assumption is satisfied.

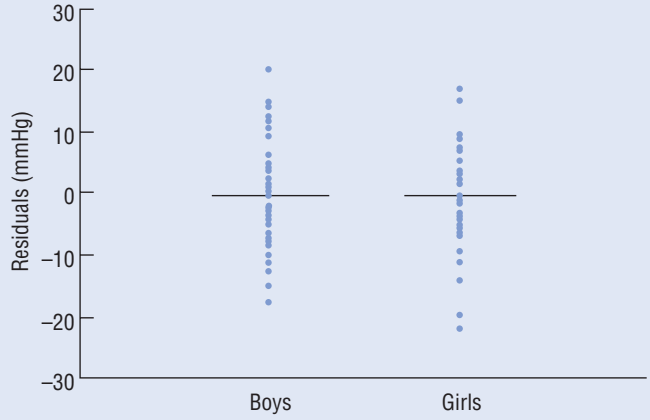


Figure 29.4 The distribution of the residuals is similar in boys and girls, suggesting that the model fits equally well in the two groups.

The multiple linear regression equation is estimated by:

$$\text{SBP} = 79.44 - (0.03 \times \text{height}) + (1.18 \times \text{weight}) + (4.23 \times \text{sex})$$

The relationship between weight and systolic blood pressure is highly significant ($P < 0.0001$), with a one kilogram increase in weight being associated with a mean increase of 1.18 mmHg in systolic blood pressure, after adjusting for height and sex. However, after adjusting for the weight and sex of the child, the relationship between height and systolic blood pressure becomes non-significant ($P = 0.86$). This suggests that the significant relationship between height and systolic blood pressure in the simple regression analysis reflects the fact that taller children tend to be heavier than shorter children. There is a significant relationship ($P = 0.01$) between sex and systolic blood pressure; systolic blood pressure in girls tends to be 4.23 mmHg higher, on average, than that of boys, even after taking account of possible differences in height and weight. Hence, both weight and sex are independent predictors of a child's systolic blood pressure.

We can calculate the systolic blood pressures we would expect for children of given heights and weights. If the first child mentioned in Chapter 28 who is 115 cm tall is a girl and weighs 37 kg, she now has an estimated predicted systolic blood pressure of $79.44 - (0.03 \times 115) + (1.18 \times 37) + (4.23 \times 1) = 123.88$ mmHg (higher than the 101.48 mmHg predicted in Chapter 28); if the second child who is 130 cm tall is a boy and weighs 30 kg, he now has an estimated predicted systolic blood pressure of $79.44 - (0.03 \times 130) + (1.18 \times 30) + (4.23 \times 0) = 110.94$ mmHg (higher than the 108.68 mmHg predicted in Chapter 28).

30 Binary outcomes and logistic regression

Reasoning

Logistic regression is very similar to linear regression; we use it when we have a **binary outcome** of interest (e.g. the presence/absence of a symptom, or an individual who does/does not have a disease) and a number of explanatory variables. From the logistic regression equation, we can determine which explanatory variables influence the outcome and, using an individual's values of the explanatory variables, evaluate the probability that s/he will have a particular outcome.

We start by creating a binary variable to represent the two outcomes (e.g. 'has disease' = 1, 'does not have disease' = 0). However, we cannot use this as the dependent variable in a linear regression analysis since the Normality assumption is violated, and we cannot interpret predicted values that are not equal to zero or one. So, instead, we take the probability, p, that an individual is classified into the highest coded category (i.e., has disease) as the dependent variable, and, to overcome mathematical difficulties, use the logistic or logit transformation (Chapter 9) of it in the regression equation. The logit of this probability is the natural logarithm (i.e. to base e) of the odds of 'disease', i.e.

$$logit(p) = \ln \frac{p}{1-p}$$

The logistic regression equation

An iterative process, called maximum likelihood (Chapter 32), rather than ordinary least squares regression (so we cannot use linear regression software), produces, from the sample data, an estimated **logistic regression equation** of the form:

$$logit(p) = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

- x_i is the ith explanatory variable ($i = 1,2,3, \ldots ,k$);
- p is the estimated value of the true probability that an individual with a particular set of values for $x_1, \ldots, x_k$ has the disease. p corresponds to the proportion with the disease; it has an underlying Binomial distribution (Chapter 8);
- a is the estimated constant term;
- $b_1, b_2, \ldots, b_k$ are the estimated **logistic regression coefficients**.

The exponential of a particular coefficient, for example, e^{b_1}, is an estimate of the **odds ratio** (Chapter 16). For a particular value of x_1, it is the estimated odds of disease for $(x_1 + 1)$ relative to the estimated odds of disease for x_1, while adjusting for all other x's in the equation. If the odds ratio is equal to one (unity), then these two odds are the same. A value of the odds ratio above one indicates an increased odds of having the disease, and a value below one indicates a decreased odds of having the disease, as x_1 increases by one unit. When the disease is rare, the odds ratio can be interpreted as a **relative risk**.

We can manipulate the logistic regression equation to estimate the probability that an individual has the disease. For each individual, with a set of covariate values for $x_1, \ldots, x_k$, we calculate

$$z = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

Then, the probability that the individual has the disease is estimated as:

$$p = \frac{e^z}{1+e^z}$$

As the logistic regression model is fitted on a log scale, the effects of the x_i's are *multiplicative* on the odds of disease. This means that their combined effect is the product of their separate effects (see Example). This is unlike the situation in linear regression where the effects of the x_i's on the dependent variable are additive.

Computer output

For each explanatory variable

Comprehensive computer output for a logistic regression analysis includes, for each explanatory variable, the estimated logistic regression coefficient with standard error, the estimated odds ratio (i.e. the exponential of the coefficient) with a confidence interval for its true value, and a Wald test statistic (testing the null hypothesis that the relevant logistic regression coefficient is zero which is equivalent to testing the hypothesis that the odds ratio of 'disease' associated with this variable is unity) and associated P-value. We use this information to determine whether each variable is related to the outcome of interest (e.g. disease), and to quantify the extent to which this is so. Automatic selection procedures (Chapter 33) can be used, as in multiple linear regression, to select the best combination of explanatory variables. A rule of thumb for the maximum number of explanatory variables to include in the model is that there should be at least 10 times as many responses in each of the two categories defining the outcome (e.g. presence/absence of a symptom) as there are variables[1].

To assess the adequacy of the model

Usually, interest is centred on examining the explanatory variables and their effect on the outcome. This information is routinely available in all advanced statistical computer packages. However, there are inconsistencies between the packages in the way in which the adequacy of the model is assessed, and in the way it is described. Your computer output may contain the following (in one guise or another).

- A quantity called **–2 log likelihood, likelihood ratio statistic (LRS)** or **deviance**: it has an approximately Chi-squared distribution, and indicates how poorly the model fits with all the explanatory variables in the model (a significant result indicates poor prediction—Chapter 32).
- The **model Chi-square** or the **Chi-square for covariates**: this tests the null hypothesis that all the regression coefficients in the model are zero (Chapter 32). A significant result suggests that at least one covariate is significantly associated with the dependent variable.

[1] Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* **49**: 1373–9.

• The percentages of individuals correctly predicted as 'diseased' or 'disease-free' by the model. This information may be in a **classification table**.
• A **histogram**: this has the predicted probabilities along the horizontal axis, and uses symbols (such as 1 and 0) to designate the group ('diseased' or 'disease-free') to which an individual belongs. A good model will separate the symbols into two groups which show little or no overlap.
• **Indices of predictive efficiency**: these are not routinely available in every computer package but may include the **false positive** and **false negative** proportions and the **sensitivity** and **specificity** of the model (Chapter 38). Our advice is to refer to more advanced texts for further information[2].

Comparing the odds ratio and the relative risk

Although the odds ratio is often taken as an estimate of the relative risk, it will only give a similar value if the outcome is rare. Where the outcome is not rare, the odds ratio will be greater than the relative risk if the relative risk is greater than one, and it will be less than the relative risk otherwise. Although the odds ratio is less easily interpreted than the relative risk, it does have attractive statistical properties and thus is usually preferred (and must be used in a case–control study when the relative risk cannot be estimated directly (Chapter 16)).

Multinomial and ordinal logistic regression

Multinomial (also called **polychotomous**) and **ordinal** logistic regression are extensions of logistic regression; we use them when we have a **categorical dependent variable** with more than two categories. When the dependent variable is *nominal* (Chapter 1) (e.g. the patient has one of three back disorders; lumbar disc hernia, chronic low-back pain, or acute low-back pain) we use **multinomial logistic regression**. When the dependent variable is *ordinal* or ranked (e.g. mild, moderate or severe pain) we use **ordinal** logistic regression. These methods are complex and so you should refer to more advanced texts[3] and/or seek specialist advice if you want to use them. As a simple alternative, we can combine the categories in some appropriate way to create a new binary outcome variable, and then perform the usual two-category logistic regression analysis (recognizing that this approach may be wasteful of information). The decision on how to combine the categories should be made in advance, before looking at the data, in order to avoid bias.

Conditional logistic regression

We can use **conditional logistic regression** when we have *matched* individuals (as in a **matched case–control study** (Chapter 16)) and we wish to adjust for possible confounding factors. Analysis of a matched case–control study using ordinary logistic regression or the methods described in Chapter 16 is inefficient and lacks power because neither acknowledges that cases and controls are linked to each other. Conditional logistic regression allows us to compare cases to controls in the same matched 'set' (i.e. each pair in the case of one-to-one matching). In this situation, the 'outcome' is defined by the patient being a case (usually coded 1) or a control (usually coded 0). Whilst advanced statistical packages may sometimes allow you to perform conditional logistic regression directly, it may be necessary to use the **Cox proportional hazards regression model** (Chapter 44).

[2] Menard S. (1995). Applied logistic regression analysis. In: *Sage University Paper Series on Quantitative Applications in the Social Sciences*, Series no. 07-106. Sage University Press, Thousand Oaks, California.

[3] Ananth, C.V. and Kleinbaum, D.G. (1997). Regression methods for ordinal responses: a review of methods and applications. *International Journal of Epidemiology* **27**: 1323–33.

Example

In a study of the relationship between human herpesvirus type 8 (HHV-8) infection (described in Chapter 23) and sexual behaviour, 271 homo/bisexual men were asked questions relating to their past histories of a number of sexually transmitted diseases (gonorrhoea, syphilis, herpes simplex type 2 [HSV-2] and HIV). In Chapter 24 we showed that men who had a history of gonorrhoea had a higher seroprevalence of HHV-8 than those without a previous history of gonorrhoea. A multivariable **logistic regression** analysis was performed to investigate whether this effect was simply a reflection of the relationships between HHV-8 and the other infections and/or the man's age. The explanatory variables were the presence of each of the four infections, each coded as '0' if the patient had no history of the particular infection or '1' if he had a history of that infection, and the patient's age in years.

A typical computer output is displayed in Appendix C. It shows that the Chi-square for covariates equals 24.598 on 5 degrees of freedom ($P = 0.0002$), indicating that at least one of the covariates is significantly associated with HHV-8 serostatus. The following table summarizes the information about each variable in the model.

Variable	Parameter estimate	Standard error	Wald Chi-square	*P*-value	Estimated odds ratio	95% CI for odds ratio
Intercept	–2.2242	0.6512	11.6670	0.0006	—	—
Gonorrhoea	0.5093	0.4363	1.3626	0.2431	1.664	(0.71–3.91)
Syphilis	1.1924	0.7111	2.8122	0.0935	3.295	(0.82–13.28)
HSV-2 positivity	0.7910	0.3871	4.1753	0.0410	2.206	(1.03–4.71)
HIV	1.6357	0.6028	7.3625	0.0067	5.133	(1.57–16.73)
Age	0.0062	0.0204	0.0911	0.7628	1.006	(0.97–1.05)

continued

These results indicate that HSV-2 positivity ($P = 0.04$) and HIV status ($P = 0.007$) are independently associated with HHV-8 infection; individuals who are HSV-2 seropositive have 2.21 times (= exp[0.7910]) the odds of being HHV-8 seropositive as those who are HSV-2 seronegative, after adjusting for the other infections. In other words, the odds of HHV-8 seropositivity in these individuals is increased by 121%. The upper limit of the confidence interval for this odds ratio shows that this increased odds could be as much as 371%. HSV-2 infection is a well-documented marker of sexual activity. Thus, rather than HSV-2 being a cause of HHV-8 infection, the association may be a reflection of the sexual activity of the individual.

Furthermore, the multiplicative effect of the model suggests that a man who is both HSV-2 and HIV seropositive is estimated to have $2.206 \times 5.133 = 11.3$ times the odds of HHV-8 infection compared to a man who is seronegative for both, after adjusting for the other infections.

In addition, there is a tendency for a history of syphilis to be associated with HHV-8 serostatus. Although this is marginally non-significant ($P = 0.09$), we should note that the confidence interval does include values for the odds ratio as high as 13.28. In contrast, there is no indication of an independent relationship between a history of gonorrhoea and HHV-8 seropositivity, suggesting that this variable appeared, by the Chi-squared test (Chapter 24), to be associated with HHV-8 serostatus because of the fact that many men who had a history of one of the other sexually transmitted diseases in the past also had a history of gonorrhoea. There is no significant relationship between HHV-8 seropositivity and age; the odds ratio indicates that the estimated odds of HHV-8 seropositivity increases by 0.6% for each additional year of age.

The probability that a 51 year-old man has HHV-8 infection if he has gonorrhoea and is HSV-2 positive (but does not have syphilis and is not HIV positive) is estimated as 0.35, i.e. it is $\exp\{-0.6077\}/\{1 + \exp(-0.6077)\}$ where $-0.6077 = 0.2242 + 0.5093 + 0.7910 + (0.0062 \times 51)$.

31 Rates and Poisson regression

Rates

In any longitudinal study (Chapter 12) investigating the occurrence of an event (such as death), we should take into account the fact that individuals are usually followed for different lengths of time. This may be because some individuals drop out of the study or because individuals are entered into the study at different times, and therefore follow-up times from different people may vary at the close of the study. As those with a longer follow-up time are more likely to experience the event than those with shorter follow-up, we consider the **rate** at which the event occurs *per person per period of time*. Often the unit which represents a convenient period of time is a year (but it could be a minute, day, week, etc.). Then the event rate per person per year (i.e. **per person-year of follow-up**) is estimated by:

$$\textbf{Rate} = \frac{\text{Number of events occurring}}{\text{Total number of years of follow-up for all individuals}} = \frac{\text{Number of events occurring}}{\text{Person-years of follow-up}}$$

Each individual's length of follow-up is usually defined as the time from when s/he enters the study until the time when the event occurs or the study draws to a close if the event does not occur. The total follow-up time is the sum of all the individuals' follow-up times.

The rate is called an **incidence rate** when the event is a new case (e.g. of disease) or the **mortality rate** when the event is death. When the rate is very small, it is often multiplied by a *convenience factor* such as 1,000 and re-expressed as the rate per 1,000 person-years of follow-up.

Features of the rate

- When calculating the rate, we do not distinguish between person-years of follow-up that occur in the same individual and those that occur in different individuals. For example, the person-years of follow-up contributed by 10 individuals, each of whom is followed for 1 year, will be the same as that contributed by 1 person followed for 10 years.
- Whether we also include multiple events from each individual (i.e. when the event occurs on more than one occasion) depends on the hypothesis of interest. If we are only interested in first events, then follow-up must cease at the point at which an individual experiences his/her first event as the individual is no longer at risk of a first event after this time. Where multiple events from the same individual are included in the calculation of the rate, we have a special form of **clustered data** (Chapter 41), and appropriate statistical methods must be used (Chapters 41 and 42).
- A rate cannot be calculated in a cross-sectional study (Chapter 12) since this type of study does not involve time.

Comparing the rate to the risk

The **risk** of an event (Chapter 15) is simply the total number of events divided by the number of individuals included in the study at the start of the investigation, with no allowance for the length of follow-up. As a result, the risk of the event will be greater when individuals are followed for longer, since they will have more opportunity to experience the event. In contrast, the **rate** of the event should remain relatively stable in these circumstances, as the rate takes account of the duration of follow-up.

Relative rates

We may be interested in comparing the rate of disease in a group of individuals exposed to some factor of interest ($\text{Rate}_{\text{exposed}}$) with that in a group of individuals not exposed to the factor ($\text{Rate}_{\text{unexposed}}$).

$$\text{Relative rate} = \frac{\text{Rate}_{\text{exposed}}}{\text{Rate}_{\text{unexposed}}}$$

The relative rate (or **rate ratio**, sometimes referred to as the **incidence rate ratio**) is interpreted in a similar way to the **relative risk** (Chapter 15) and to the **odds ratio** (Chapters 16 and 30); a relative rate of 1 (unity) indicates that the rate of disease is the same in the two groups, a relative rate greater than 1 indicates that the rate is higher in those exposed to the factor than in those who are unexposed, and a relative rate less than one indicates that the rate is lower in the group exposed to the factor.

Although the relative rate is often taken as an estimate of the relative risk, the relative rate and the relative risk will only be similar if the event (e.g. disease) is rare. When the event is not rare and individuals are followed for varying lengths of time, the rate, and therefore the relative rate, will not be affected by the different follow-up times. This is not the case for the relative risk as the risk, and thus the relative risk, will change as individuals are followed for longer periods. Hence, the relative rate is always preferred when follow-up times vary between individuals in the study.

Poisson regression

What is it?

The Poisson distribution (named after a French mathematician) is a probability distribution (Chapter 8) of the count of the number of rare events that occur randomly over an interval of time (or space) at a constant average rate. This forms the basis of Poisson regression which is used to analyse the rate of some event (e.g. disease) when individuals have different follow-up times. This contrasts with logistic regression (Chapter 30) which is concerned only with whether or not the event occurs and is used to estimate odds ratios. In Poisson regression, we assume that the rate of the event among individuals with the same explanatory variables (e.g. age and sex) is constant over the whole study period. We generally want to know which explanatory variables influence the rate at which the event occurs, and may wish to compare this rate in different exposure groups and/or predict the rate for groups of individuals with particular characteristics.

The equation and its interpretation

The Poisson regression model takes a very similar form to the logistic regression model (Chapter 30), each having a (usually) linear combination of explanatory variables on the right hand side of the equation. Poisson regression analysis also mirrors logistic regression analysis in that we transform the outcome variable in order to overcome mathematical difficulties. We use the natural log transformation (ln) of the rate and an iterative process (maximum likeli-

hood, Chapter 32) to produce an estimated regression equation from the sample data of the form:

$$\ln(r) = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

- x_i is the ith explanatory variable ($i = 1, 2, 3, \ldots, k$);
- r is the estimated value of the mean or expected rate for an individual with a particular set of values for $x_1, \ldots, x_k$;
- a is the estimated constant term providing an estimate of the log rate when all x_i's in the equation take the value zero (the log of the baseline rate);
- $b_1, b_2, \ldots, b_k$ are the estimated **Poisson regression coefficients**.

The exponential of a particular coefficient, for example, e^{b_1}, is the estimated **relative rate** associated with the relevant variable. For a particular value of x_1, it is the estimated rate of disease for $(x_1 + 1)$ relative to the estimated rate of disease for x_1, while adjusting for all other x_i's in the equation. If the relative rate is equal to one (unity), then the event rates are the same when x_1 increases by one unit. A value of the relative rate above one indicates an increased event rate, and a value below one indicates a decreased event rate, as x_1 increases by one unit.

As with logistic regression, Poisson regression models are fitted on the log scale. Thus, the effects of the x_i's are *multiplicative* on the rate of disease.

We can manipulate the Poisson regression equation to estimate the event rate for an individual with a particular combination of values of $x_1, \ldots, x_k$. For each set of covariate values for $x_1, \ldots, x_k$, we calculate

$$z = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

Then, the event rate for that individual is estimated as e^z.

Use of an offset

Although we model the rate at which the event occurs (i.e. the number of events divided by the person-years of follow-up), most statistical packages require you to specify the number of events occurring rather than the rate itself as the dependent variable. The log of each individual's person-years of follow-up is then included as an **offset** in the model. Assuming that we are only interested in including a single event per person, the number of events occurring in each individual will either take the value 0 (if the event did not occur) or 1 (if the event did occur). This provides a slightly different formulation of the model which allows the estimates to be generated in a less computationally intensive way. The results from the model, however, are exactly the same as they would be if the rate were modelled.

Entering data for groups

Note that when all of the explanatory variables are categorical, we can make use of the fact that the calculation of the rate does not distinguish between person-years of follow-up that occur in the same individual and those that occur in different individuals to simplify the data entry process. For example, we may be interested in the effect of only two explanatory variables, sex (male or female) and age (<16, 16–20 and 21–25 years), on the rate of some event. Between them, these two variables define six groups (i.e. males aged < 16 years, females aged < 16 years, . . . , females aged 21–25 years). We can simplify the entry of these data by determining the total number of events for all individuals within the same group and the total person-years of follow-up for these individuals. The estimated rate in each group is then calculated as the total number of events divided by the person-years of follow-up in that group. Using this approach, rather than entering data for the n individuals one by one, we enter the data for each of the 6 resulting groups, including in the model the binary and dummy variables (Chapter 29) separately for sex and age. Note that when entering data in this way, it is not possible to accommodate numerical covariates to define the groups or include an additional covariate in the model that takes different values for the individuals in a group.

Incorporating variables that change over time

By splitting the follow-up period into shorter intervals, it is possible to incorporate variables that **change over time** into the model. For example, we may be interested in relating the smoking history of middle-aged men to the rate at which they experience lung cancer. Over a long follow-up period, many of these men may give up smoking and their rates of lung cancer may be lowered as a result. Thus, categorizing men according to their smoking status at the start of the study may give a poor representation of the impact of smoking status on lung cancer. Instead, we split each man's follow-up into short time intervals in such a way that his smoking status remains constant in each interval. We then perform a Poisson regression analysis, treating the relevant information in each short time interval for each man (i.e. the occurrence/non-occurrence of the event, his follow-up time and smoking status) as if it came from a different man.

Computer output

Comprehensive computer output for a Poisson regression analysis includes, for each explanatory variable, the estimated Poisson regression coefficient with standard error, the estimated relative rate (i.e. the exponential of the coefficient) with a confidence interval for its true value, and a Wald test statistic (testing the null hypothesis that the regression coefficient is zero or, equivalently, that the relative rate of 'disease' associated with this variable is unity) and associated P-value. As with the output from logistic regression (Chapter 30), we can assess the adequacy of the model using −2 log likelihood (LRS or deviance) and the model Chi-square or the Chi-square for covariates (see also Chapter 32).

Extra-Poisson variation

One concern when fitting a Poisson regression model is the possibility of *extra-Poisson variation* which usually implies *overdisperson.* This occurs when the residual variance is greater than would be expected from a Poisson model, perhaps because an outlier is present (Chapter 3) or because an important explanatory variable has not been included in the model. Then the standard errors are usually underestimated and, consequently, the confidence intervals for the parameters are too narrow and the P-values too small. A way to investigate the possibility of overdispersion is to divide −2 log likelihood (LRS or deviance) by the degrees of freedom, $n - k$, where n is the number of observations in the data set and k is the number of parameters fitted in the model (including the constant term). This quotient should be approximately equal to 1 if there is no extra-Poisson variation; values substantially above 1 may indicate overdisperson. *Underdispersion*, where the residual variance is less than would be expected from a Poisson model and where the ratio of

–2 log likelihood to $n - k$ is substantially less than 1, may also occur (e.g. if high counts cannot be recorded accurately). Underdispersion and overdispersion may also be a concern when performing logistic regression (Chapter 30) when they are referred to as *extra-Binomial variation*.

Alternative to Poisson analysis

When a group of individuals is followed from a natural 'starting point' (e.g. an operation) until the time that the person develops an endpoint of interest, we may use an alternative approach known as **survival analysis** which, in contrast to Poisson regression, does not assume that the 'hazard' (the rate of the event in a small interval) is constant over time. This approach is described in detail in Chapter 44.

Example

Individuals with HIV infection treated with highly active antiretroviral therapy (HAART) usually experience a decline in HIV viral load to levels below the limit of detection of the assay (an *initial response*). However, some of these individuals may experience virological failure after this stage; this occurs when an individual's viral load becomes detectable again whilst on therapy. Identification of factors that are associated with an increased rate of virological failure may allow steps to be taken to prevent this occurring. There is some concern that the rate of virological failure may increase with longer time on therapy. As patients are followed for different lengths of time, a **Poisson regression analysis** is appropriate.

516 patients who experienced an initial response to therapy were identified and followed until the time of virological failure, or until their last date of follow-up if their viral load remained suppressed at this time. Follow-up started on the first date that their viral load became undetectable. The explanatory variable of primary interest was the duration of time on treatment since an initial response but this was a variable whose values were constantly changing for each patient during the study period. Therefore, to investigate whether the virological failure rate did change over time, the duration of time on treatment since an initial response was split into three time intervals: <1, 1–2 and >2 years (this created 988 sets of observations), with the broad assumption that the virological failure rate was approximately constant within each period. Failure rates in the three time periods were then compared. The data (the length of follow-up in that interval, whether or not virological failure was experienced in that interval, and relevant explanatory variables) were entered on to a spreadsheet for each patient in every interval in which s/he was followed up. The explanatory variables considered included demographics, the stage of disease at the time of starting therapy, the year of starting HAART and whether or not the patient had received treatment in the past.

In order to *limit the number of covariates* in the multivariable Poisson regression model, a separate univariable Poisson regression model for each covariate was used to identify the covariates associated with virological failure (see Chapter 34).

Over a total follow-up of 718 person-years, 61 patients experienced virological failure, an unadjusted event rate of 8.50 per 100 person-years (95% confidence interval: 6.61, 10.92). Unadjusted virological failure rates were 8.13 (6.31, 10.95) in the first year after initial response to therapy, 12.22 (7.33, 17.12) in the second year and 3.99 (1.30, 9.31) in later years. Results from a Poisson regression model that incorporated only two dummy variables (Chapter 29) to reflect the categories of 1–2 and >2 years, each compared to <1 year, since an initial response to therapy suggested that time since initial virological response was significantly associated with virological failure ($P = 0.04$). In addition, the patient's sex ($P = 0.03$), his/her baseline CD8 count ($P = 0.01$) and treatment status at the time of starting the current regimen (previously received treatment, never received treatment, $P = 0.008$) were all significantly associated with virological failure in univariable Poisson models. Thus, a multivariable Poisson regression analysis was performed to assess the relationship between virological failure and duration of time on therapy, after adjusting for these other variables. The results are summarized in Table 31.1; full computer output is shown in Appendix C.

The *results* from this multivariable model suggested that there was a trend towards a higher virological failure rate in the period 1–2 years after initial response compared to that seen in the first year (virological failure rate was increased by 53% in the period 1–2 years), but a *lower* rate after the second year (failure rate was reduced by 44% in this period compared to that seen in the first year after initial response), although neither of these effects was statistically significant. After adjusting for all other variables in the model, patients who were receiving their first treatment had an estimated virological failure rate that was 44% lower than that of patients who had previously received treatment, the estimated virological failure rate in men was 39% less than that seen in women (this was not statistically significant), and the estimated virological failure rate was reduced by 65% if the CD8 count at baseline was 100 cells/mm^3 higher.

See also the Examples in Chapters 32 and 33 for additional analyses relating to this Poisson model, including assessments of overdispersion, goodness of fit and linearity of the covariates.

continued

Table 31.1 Results from multivariable Poisson regression analysis of factors associated with virological failure.

Variable*		Parameter estimate	Standard error	Estimated relative rate	95% Confidence interval for relative rate	Wald *P*-value†
Time since initial response to therapy (years)	<1	reference	—	1	—	—
	1–2	0.4256	0.2702	1.53	0.90, 2.60	0.12
	>2	−0.5835	0.4825	0.56	0.22, 1.44	0.23
Treatment status						
Previously received treatment (0)		reference	—	1		
Never received treatment (1)		−0.5871	0.2587	0.56	0.33, 0.92	0.02
Sex						
Female (0)		reference	—	1	—	
Male (1)		−0.4868	0.2664	0.61	0.36, 1.04	0.07
CD8 count (per 100 cells/mm^3)		−1.0558	0.0267	0.35	0.33, 0.37	0.04

*Codes for binary variables (sex and treatment status) are shown in parentheses. Time since initial response to therapy was included by incorporating dummy variables to reflect the periods 1–2 years and >2 years after initial response.
†An alternative method of assessing the significance of categorical variables with more than two categories is described in Chapters 32 and 33.

Adapted from work carried out by Ms Colette Smith, Department of Primary Care and Population Sciences, Royal Free and University College Medical School, London, UK.

32 Generalized linear models

Statistical modelling includes the use of simple and multiple linear regression (Chapters 27–29), logistic regression (Chapter 30), Poisson regression (Chapter 31) and some methods that deal with survival data (Chapter 44). All these methods rely on generating a **mathematical model** that best describes the relationship between an outcome and one or more explanatory variables. Generation of such a model allows us to determine the extent to which each explanatory variable is related to the outcome after adjusting for all other explanatory variables in the model and, if desired, to predict the value of the outcome from these explanatory variables.

The **generalized linear model** (**GLM**) can be expressed in the form

$$g(Y) = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

where Y is the estimated value of the predicted, mean or expected value of the dependent variable which follows a known probability distribution (e.g. Normal, Binomial, Poisson); $g(Y)$, called the **link function**, is a transformation of Y which produces a linear relationship with $x_1, \ldots, x_k$, the predictor or explanatory variables; $b_1, \ldots, b_k$ are estimated regression coefficients that relate to these explanatory variables; and a is a constant term.

Each of the regression models described in earlier chapters can be expressed as a particular type of GLM (see Table 32.1). The link function is the **logit** of the proportion (i.e. the $\log_e$ of the odds) in logistic regression and the **$\log_e$** of the rate in Poisson regression. No transformation of the dependent variable is required in simple and multiple linear regression; the link function is then referred to as the **identity link**. Once you have specified which type of regression you wish to perform, most statistical packages incorporate the link function into the calculations automatically without any need for further specification.

Which type of model do we choose?

The choice of an appropriate statistical model will depend on the outcome of interest (see Table 32.1). For example, if our dependent variable is a continuous numerical variable, we may use simple or multiple linear regression to identify factors associated with this variable. If we have a binary outcome (eg. patient died or did not die) and all patients are followed for the same amount of time, then logistic regression would be the appropriate choice of model.

Note that we may be able to choose a different type of model by modifying the format of our dependent variable. In particular, if we have a continuous numerical outcome but one or more of the assumptions of linear regression are not met, we may choose to categorize our outcome variable into two groups to generate a new binary outcome variable. For example, if our dependent variable is systolic blood pressure (a continuous numerical variable) after a six-month period of anti-hypertensive therapy, we may choose to dichotomize the systolic blood pressure as high or low using a particular cut-off, and then use logistic regression to identify factors associated with this binary outcome. Whilst dichotomizing the dependent variable in this way may simplify the fitting and interpretation of the statistical model, some information about the dependent variable will usually be discarded. Thus the advantages and disadvantages of this approach should always be considered carefully.

Likelihood and maximum likelihood estimation

When fitting a GLM, we generally use the concept of **likelihood** to estimate the parameters of the model. For any GLM characterized by a known probability distribution, a set of explanatory variables and some potential values for each of their regression coefficients, the likelihood of the model (L) is the *probability* that we would have obtained the *observed results* had the regression coefficients taken those values. We estimate the coefficients of the model by selecting the values for the regression coefficients that maximize L (i.e. they are those values that are *most likely* to have produced our observed results); the process is **maximum likelihood estimation** (**MLE**) and the estimates are **maximum likelihood estimates**. MLE is an iterative process and thus specialized computer software is required. One exception to MLE is in the case of simple and multiple linear regression models (with the identity link function) where we usually estimate the parameters using the **method of least squares** (the estimates are often referred to as **ordinary least squares** (**OLS**) estimates; Chapter 27); the OLS and MLE estimates are identical in this situation.

Assessing adequacy of fit

Although MLE maximizes L for a given set of explanatory variables, we can always improve L further by including additional explanatory variables. At its most extreme, a **saturated** model is one that includes a separate variable for each observation in the data

Table 32.1 Choice of appropriate types of GLM for use with different types of outcome.

Type of outcome	Type of GLM commonly used	See Chapter
Continuous numerical	Simple or multiple linear	28, 29
Binary		
Incidence of disease in longitudinal study (patients followed for equal periods of time)	Logistic	30
Binary outcome in cross-sectional study	Logistic	30
Unmatched case-control study	Logistic	30
Matched case-control study	Conditional logistic	30
Categorical outcome with more than two categories	Multinomial or ordinal logistic regression	30
Event rate or count	Poisson	31
Time to event*	Exponential, Weibull or Gompertz models	44

*Time to event data may also be analysed using a Cox proportional hazards regression model (Chapter 44).

set. Whilst such a model would explain the data perfectly, it is of limited use in practice as the prediction of future observations from this model is likely to be poor. The saturated model does, however, allow us to calculate the value of L that would be obtained if we could model the data perfectly. Comparison of this value of L with the value obtained after fitting our simpler model with fewer variables provides a way of assessing the **adequacy of the fit** of our model. We consider the **likelihood ratio**, the ratio of the value of L obtained from the saturated model to that obtained from the fitted model, in order to compare these two models. More specifically, we calculate the **likelihood ratio statistic** (**LRS**) as:

$$\text{LRS} = -2 \times \frac{\log(L_{\text{saturated}})}{\log(L_{\text{fitted}})}$$
$$= -2 \times \{\log(L_{\text{saturated}}) - \log(L_{\text{fitted}})\}.$$

The LRS, often referred to as **–2log likelihood** (see Chapters 30 and 31) or as the **deviance**, approximately follows a Chi-squared distribution with degrees of freedom equal to the difference in the number of parameters fitted in the two models (i.e. $n - k$ where n is the number of observations in the data set and k is the number of parameters, including the intercept, in the simpler model). The null hypothesis is that the extra parameters in the larger saturated model are all zero; a large value of the LRS will give a significant result indicating that the goodness of fit of the model is poor.

The LRS can also be used in other situations. In particular, the LRS can be used to compare two models, neither of which is saturated, when one model is **nested** within another (i.e. the larger model includes all of the variables that are included in the smaller model, in addition to extra variables). In this situation, the test statistic is the difference between the value of the LRS from the model which includes the extra variables and that from the model which excludes these extra variables. The test statistic follows a Chi-squared distribution with degrees of freedom equal to the number of *additional* parameters included in the larger model, and is used to test the null hypothesis that the extra parameters in the larger model are all zero. The LRS can also be used to test the null hypothesis that all the parameters associated with the covariates of a model are zero by comparing the LRS of the model which includes the covariates to that of the model which excludes them. This is often referred to as the **Model Chi-square** or **the Chi-square for covariates** (see Chapters 30 and 31).

Regression diagnostics

When performing any form of regression analysis, it is important to consider a series of regression diagnostics. These allow us to examine our fitted regression models and look for flaws that may affect our parameter estimates and their standard errors. In particular, we must consider whether the assumptions underlying the model are violated (Chapter 28) and whether our results are heavily affected by influential observations (Chapter 29).

Example

In the example in Chapter 31, we used Wald tests to identify individual factors associated with virological rebound in a group of 516 HIVtve the patients (with 988 sets of observations) who had been treated with highly active antiretroviral therapy (HAART). In particular, we were interested in whether the rate of virological failure increased over time, after controlling for other potentially confounding variables that were related to virological failure. Although the outcome of primary interest was binary (patient experienced virological failure, patient did not experience virological failure), a Poisson regression model rather than a logistic model was chosen as individual patients were followed for different lengths of time. Thus, the outcome variable for the analysis performed was an event rate. In this chapter, P-values for the variables were calculated using **likelihood ratio statistics**. In particular, to calculate the single P-value associated with both dummy variables representing the time since initial response to therapy, two models were fitted. The first included the variables relating to treatment status (previously received treatment, never received treatment), sex and baseline CD8 count (Model 1); the second included these variables as well as the two dummy variables (Model 2). The difference between the values obtained for **–2 log likelihood** (i.e. the LRS or deviance) from each of the models was then considered (Table 32.2). A full computer output is shown in Appendix C.

The inclusion of the two dummy variables was associated with a reduction in the value of –2log likelihood of 5.53 (= 393.12 – 387.59). This test statistic follows the Chi-squared distribution with 2 degrees of freedom (as 2 additional parameters were included in the larger model); the P-value associated with this test statistic was 0.06 indicating that the relationship between virological failure and time since initial response is marginally non-significant. The value of –2log likelihood for Model 2 also allowed us to assess the adequacy of fit of this model by comparing its value of –2log likelihood to a Chi-squared distribution with 982 degrees of freedom. The P–value obtained for this comparison was >0.99, suggesting that the goodness of fit of the model is acceptable. However, it should be noted that after including these five variables in the model, there was some evidence of **underdispersion**, as the ratio of –2log likelihood divided by its degrees of freedom was 0.39 which is substantially less than 1, suggesting that the amount of residual variation was less than would be expected from a Poisson model (see Chapter 31).

Table 32.2 –2Log likelihood values, degrees of freedom and number of parameters fitted in models that exclude and include the time since initial response to therapy.

Model	Variables included	–2log likelihood	Degrees of freedom for the model	Number of parameters fitted in the model, including the intercept
1	Treatment status, sex and baseline CD8 count	393.12	984	4
2	Treatment status, sex, baseline CD8 count and 2 dummy variables for time since initial response to therapy	387.59	982	6

33 Explanatory variables in statistical models

Whatever type of statistical model we choose, we have to make decisions about which explanatory variables to include in the model and the most appropriate way in which they should be incorporated. These decisions will depend on the type of explanatory variable (either nominal categorical, ordinal categorical or numerical) and the relationship between these variables and the dependent variable.

Nominal explanatory variables

It is usually necessary to create **dummy** or **indicator** variables (Chapter 29) to investigate the effect of a nominal categorical explanatory variable in a regression analysis. Note that when assessing the adequacy of fit of a model that includes a nominal variable with more than two categories, or assessing the significance of that variable, it is important to include *all* of the dummy variables in the model *at the same time*; if we do not do this (i.e. if we only include one of the dummy variables for a particular level of the categorical variable) we would only partially assess the impact of that variable on the outcome. For this reason, it is preferable to judge the significance of the variable using the likelihood ratio test statistic (LRS—Chapter 32), rather than by considering individual *P*-values for each of the dummy variables.

Ordinal explanatory variables

In the situation where we have an ordinal variable with more than two categories, we may take one of two approaches.

- Treat the categorical variable as a continuous numerical measurement by allocating a numerical value to each category of the variable. This approach makes full use of the ordering of the categories but it usually assumes a linear relationship (when the numerical values are equally spaced) between the explanatory variable and the dependent variable (or a transformation of it) and this should be validated.
- Treat the categorical variable as a nominal explanatory variable and create a series of dummy or indicator variables for it (Chapter 29). This approach does not take account of the ordering of the categories and is therefore wasteful of information. However, it does not assume a linear relationship with the dependent variable and so may be preferred.

The difference in the values of the LRS from these two models provides a test statistic for a **test of linear trend** (i.e. an assessment of whether the model assuming a linear relationship gives a better fitting model than one for which no linear relationship is assumed). This test statistic follows a Chi-squared distribution with degrees of freedom equal to the difference in the number of parameters in the two models; a significant result suggests non-linearity.

Numerical explanatory variables

When we include a numerical explanatory variable in the model, the estimate of its regression coefficient provides an indication of the impact of a one unit increase in the explanatory variable on the outcome. Thus, for simple and multiple linear regression, the relationship between each explanatory variable and the dependent variable is assumed to be linear. For Poisson and logistic regression, the parameter estimate provides a measure of the impact of a one unit increase in the explanatory variable on the *log* of the dependent variable (i.e. the model assumes a linear relationship between the explanatory variable and the log of the rate or of the odds, but an exponential relationship with the actual rate or odds). It is important to check the appropriateness of the assumption of linearity (see next section) before including numerical explanatory variables in regression models.

Assessing the assumption of linearity

To check the linearity assumption in a simple or multiple linear regression model, we plot the numerical dependent variable, *y*, against the numerical explanatory variable, *x*, or plot the residuals of the model against *x* (Chapter 28). The raw data should approximate a straight line and there should be no discernable pattern in the residuals. We may assess the assumption of linearity in logistic regression (Chapter 30) or Poisson regression (Chapter 31) by categorizing individuals into a small number (5–10) of equally sized subgroups according to their values of *x*. In Poisson regression, we calculate the log of the rate of the outcome in each subgroup and plot this against the midpoint of the range of values for *x* for the corresponding subgroup (see Fig. 33.1). For logistic regression, we calculate the log odds for each subgroup and plot this against the midpoint. In each case, if the assumption of linearity is reasonable, we would expect to see a similarly sized step-wise increase (or decrease) in the log of the rate or odds when moving between adjacent categories of *x*.

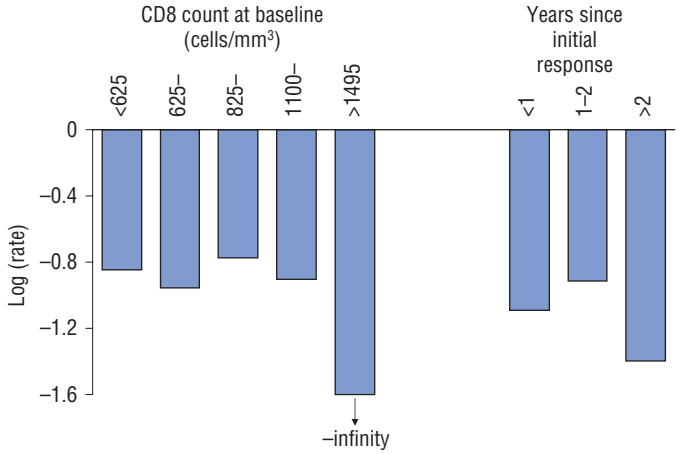


Figure 33.1 Plot of the log(rate) according to the baseline CD8 count and the time since initial response to HAART. Neither variable exhibits linearity.

Dealing with non-linearity

If non-linearity is detected in any of these plots, there are a number of approaches that can be taken.

- Replace x by a set of dummy variables created by categorizing the individuals into three or four subgroups according to the magnitude of x (often defined using the tertiles or quartiles of the distribution). This set of dummy variables can be incorporated into the multivariable regression model as categorical explanatory variables (see Example).
- Transform the x variable in some way (e.g. by taking a log or square root transformation of x; Chapter 9) so that the resulting relationship between the transformed value of x and the dependent variable (or its log for Poisson or its logit for logistic regression) *is* linear.
- Find some algebraic description that approximates the non-linear relationship using higher orders of x (e.g. a quadratic or cubic relationship). This is known as **polynomial regression**. We just introduce terms that represent the relevant higher orders of x into the equation. So, for example, if we have a cubic relationship, our estimated multiple linear regression equation is $Y = a + b_1x + b_2x^2 + b_3x^3$. We fit this model, and proceed with the analysis in exactly the same way as if the quadratic and cubic terms represented different variables (x_2 and x_3, say) in a multiple regression analysis. For example, we may fit a quadratic model that comprises the explanatory 'variables' height and height^2. We can test for linearity by comparing the LRS in the linear and quadratic models (Chapter 32), or by testing the coefficient of the quadratic term.

Selecting explanatory variables

Even if not saturated (Chapter 32), there is always the danger of **over-fitting** models by including a very large number of explanatory variables; this may lead to spurious results that are inconsistent with expectations, especially if the variables are highly correlated. For a multiple linear regression model, a usual rule-of-thumb is to ensure that there are *at least ten times as many individuals as explanatory variables*. For logistic regression, there should be at least 10 times as many *responses* or *events* in each of the two outcome categories as explanatory variables.

Often, we have a large number of explanatory variables that we believe may be related to the dependent variable. For example, many factors may appear to be related to systolic blood pressure, including age, dietary and other lifestyle factors. We should only include explanatory variables in a model if there is reason to suppose, from a biological or clinical standpoint, that they are related to the dependent variable. We can eliminate some variables by performing a *univariable* analysis (perhaps with a less stringent significance level of 0.10 rather than the more conventional 0.05) for each explanatory variable to assess whether it is likely to be related to the dependent variable, e.g. if we have an numerical dependent variable, we may perform a simple regression analysis if the explanatory variable is numerical or an unpaired t-test if it is binary. We then consider only those explanatory variables that were significant at this first stage for our multivariable model (see the Example in Chapter 31).

Automatic selection procedures

When we are particularly interested in using the model for prediction, rather than in gaining insight into whether an explanatory variable influences the outcome or in estimating its effect, computer intensive **automatic selection procedures** provide a means of identifying the optimal model by selecting some of these variables.

- **All subsets**—every combination of explanatory variables is considered; that which provides the best fit, as described by the model R^2 (Chapter 27) or LRS (Chapter 32), is selected.
- **Backwards selection**—all possible variables are included; those that are judged by the model to be least important (where this decision is based on the change in R^2 or the LRS) are progressively removed until none of the remaining variables can be removed without significantly affecting the fit of the model.
- **Forwards selection**—variables that contribute most to the fit of the model (based on the change in R^2 or the LRS) are progressively added until no further variable significantly improves the fit of the model.
- **Stepwise selection**—a combination of forwards and backwards selection that starts by progressing forwards and then, at the end of each 'step', checks backwards to ensure that all of the included variables are still required.

Disadvantages

Although these procedures remove much of the manual aspect of model selection, they have some disadvantages.

- It is possible that two or more models will fit the data equally well, or that changes in the data set will produce different models.
- Because of the multiple testing that occurs when repeatedly comparing one model to another within an automatic selection procedure, the Type I error rate (Chapter 18) is particularly high. Thus, some significant findings may arise by chance. This problem may be alleviated by choosing a more stringent significance level (say 0.01 rather than 0.05).
- If the model is refitted to the data set using the m, say, variables remaining in the final automatic selection model, its estimated parameters may differ from those of the automatic selection model. This is because the automatic selection procedure uses in its analysis only those individuals who have complete information on *all* the explanatory variables, but the sample size may be greater when individuals are included if they have no missing values only for the relevant m variables.
- The resulting models, although mathematically justifiable, may not be sensible. In particular, when including a series of dummy variables to represent a single categorical variable (Chapter 29), automatic models may include only some of the dummy variables, leading to problems in interpretation.

Therefore, a combination of these procedures and common sense should be applied when selecting the best fitting model. Models that are generated using automatic selection procedures should be validated on other external data sets where possible (see 'validating the score', Chapter 34).

Example

In Chapters 31 and 32, we studied the factors associated with virological failure in HIV+ve the patients receiving highly active antiretroviral therapy (HAART). In this multivariable Poisson regression analysis, the individual's CD8 count at baseline was included as a continuous explanatory variable (it was divided by 100 so that each unit increase in the scaled variable reflected a 100 cell/mm^3 increase in the CD8 count); the results indicated that a higher baseline CD8 count was associated with a significantly reduced rate of virological failure. In order to assess the validity of the **linearity assumption** associated with this variable, five groups were defined on the basis of the quintiles of the CD8 distribution, and the failure rate was calculated in each of the five groups. A plot of the log(rate) in each of these groups revealed that the relationship was *not* linear as there was no stepwise progression (Figure 33.1). In particular, whilst the log(rate) was broadly similar in the four lowest groups, no events occurred at all in the highest group (>1495 cells/mm^3), giving a value of minus infinity for the log(rate). For this reason, the two upper groups were combined for the subsequent analysis. Furthermore, it was noted that a substantial number of patients had to be excluded from this analysis as there was no record of their CD8 counts at baseline.

Thus, because of the lack of linearity between the virological failure rate and the actual CD8 count, the continuous explanatory variable representing the CD8 count in the Poisson regression model was replaced by a series of four dummy variables (see Chapter 29). Individuals with baseline CD8 counts in the range 825 < CD8 < 1100 cells/mm^3 were treated as the reference group for these dummies. Each of three dummy variables provided a comparison of one of the remaining CD8 groups with the reference group, and the fourth dummy provided a comparison of those with missing CD8 counts with the reference group. The results are summarized in Table 33.1; a full computer output is shown in Appendix C. A comparison of the value for −2 log likelihood (i.e. the LRS or deviance) from the model that included the four dummy variables for the CD8 count (387.15) to that from the model that included the same variables apart from these dummy variables (392.50) gave a *P*-value of 0.25 (test statistic of 5.35 on 4 degrees of freedom). Thus, after incorporating it in this way, the CD8 count no longer had a statistically significant relationship with virological failure in contrast to the model which, inappropriately, incorporated the CD8 count as a continuous explanatory variable. The relationships between virological failure and treatment status, sex and time since initial response to therapy, however, remained similar.

Table 33.1 Results from multivariable Poisson regression analysis of factors associated with virological failure, after including the CD8 count as a categorical variable in the model.

Variable*		Parameter estimate	Standard error	Estimated relative rate	95% Confidence interval for relative rate	*P*-value†
Time since initial response to therapy (years)	<1	reference	—	1	—	
	1–2	0.4550	0.2715	1.58	0.93, 2.68	
	>2	−0.5386	0.4849	0.58	0.23, 1.51	0.06
Treatment status						
Previously received treatment (0)		reference	—	1	—	
Never received treatment (1)		−0.5580	0.2600	0.57	0.34, 0.95	0.03
Sex						
Female (0)		reference	—	1	—	
Male (1)		−0.4970	0.2675	0.61	0.36, 1.03	0.07
CD8 count (cells/mm^3)						
	<625	−0.2150	0.6221	0.81	0.24, 2.73	
	≥625, <825	−0.3646	0.7648	0.63	0.16, 3.11	
	≥825, <1100	reference	—	1	—	
	≥1100	−0.3270	1.1595	0.78	0.07, 7.00	
	Missing	−0.8264	0.6057	0.44	0.13, 1.43	0.25

* Codes for binary variables (sex and treatment status) are shown in parentheses. Time since initial response to therapy was included by incorporating two dummy variables to reflect the periods 1–2 years and >2 years after initial response. The baseline CD8 count was incorporated as described above.
† *P*-values were obtained using LRS (see Chapter 32); where dummy variables were used to incorporate more than 2 categories of the variable, the *P*-value reflects the combined effect of these dummies.

34 Issues in statistical modelling

Interaction

What is it?

Statistical **interaction** (also known as **effect modification**, Chapter 13) between two explanatory variables in a regression analysis occurs where the relationship between one of the explanatory variables and the dependent variable *is not the same for different levels* of the other explanatory variables, i.e. the two explanatory variables do not act independently on the dependent variable. For example, suppose current smoking status and alcohol status can each be categorized into two levels (smoker/non-smoker and drinker/non-drinker) and that each individual belongs to one category of each variable. If the difference in diastolic blood pressure (the dependent variable) between smokers and non-smokers is greater on average in those who do not consume alcohol than in those who do, then we say that there is an interaction between smoking and alcohol consumption.

Testing for interaction

Testing for statistical interaction in a regression model is usually straightforward, and many statistical packages allow you to request the inclusion of interaction terms. If the package does not provide this facility then an interaction term may be created manually by including the product of the relevant variables as an additional explanatory variable. Thus, to obtain the value of the variable which represents the interaction between two variables (both binary, both numerical or one binary and one numerical), we multiply the individual's values of these two variables. If both variables are numerical, interpretation may be easier if we create an interaction term from the two binary variables obtained by dichotomizing each numerical variable. If one of the two variables is categorical with more than two categories, we create a series of dummy variables from it (Chapter 29) and use each of them, together with the second binary or numerical variable of interest, to generate a series of interaction terms. This procedure can be extended if both variables are categorical and each has more than two categories.

Interaction terms should only be included in the regression model after the **main effects** (the effects of the variables without any interaction) have been included. Note that statistical tests of interaction are usually of **low power** (Chapter 18). This is of particular concern when both explanatory variables are categorical and few events occur in the subgroups formed by combining each level of one variable with every level of the other, or if these subgroups include very few individuals.

Confounding

What is it?

A **confounding variable** or **confounder** is an explanatory variable that is related to both the dependent variable and to one or more of the explanatory variables in the model. For example, we may be interested in studying the effects of smoking status and alcohol consumption on the incidence of coronary heart disease (CHD) in a cohort of middle-aged men. Whilst alcohol consumption and smoking status are both known to be associated with the development of CHD, the two variables are also related to each other (i.e. men who consume alcohol are more likely to smoke than men who do not consume alcohol). Any regression model that considers the effect of one of the explanatory variables on the outcome but does not include the confounder (e.g. a model that relates smoking status to the incidence of CHD *without* adjusting for alcohol consumption) may misrepresent the true role of the explanatory variable. Confounding may either hide a true association or may artificially create a false association between the explanatory variable and the outcome variable. Failure to adjust for confounding factors in regression analyses will lead to *biased* (Chapter 12) estimates of the parameters of the model.

Dealing with confounding

Confounding may be dealt with in one of two ways:

- Create subgroups by stratifying the data set by the levels of the confounding variable (e.g. create two subgroups, drinkers and non-drinkers) and then perform an analysis separately in each subgroup. Whilst this approach is simple and has much to recommend it when there are few confounders, (1) the subgroups may be small, and thus the analyses will have reduced power to detect a significant effect, (2) spurious significant results may arise because of multiple testing (Chapter 18) if a hypothesis test is performed in each subgroup, and (3) it may be difficult to combine the separate estimates of the effect of interest for each subgroup.
- Adjust for the confounding variable in a multivariable regression model. This approach, which is particularly useful when there are many confounders in the study, provides an estimate of the relationship between the explanatory and dependent variables *that cannot be explained* by the relationship between the dependent variable and the confounder.

Confounding in non-randomized studies

We have to be particularly wary of confounding when comparing treatments in a non-randomized clinical cohort study (Chapter 15). In this type of study, the characteristics of the individuals may be unevenly distributed in the different treatment groups. For example, individuals may be selected for a particular treatment on the basis of disease history, demographic or lifestyle factors. Some of these factors may be related to the outcome and will therefore be confounded with the treatment. Whilst multivariable regression models can be used to adjust for any differences in the distribution of the factors in the different treatment groups, this is only possible if the study investigators are aware of the confounding factors and have recorded them in the data set. Randomized controlled trials (Chapter 14) rarely suffer from confounding as patients are randomly allocated to treatment groups and therefore all covariates, both confounders and other explanatory variables, should be evenly distributed in the different treatment groups.

Adjusting for intermediate variables

Where a variable is known to lie on the *causal pathway* between the explanatory variable and the outcome of interest, it is known as an **intermediate variable**. We should be careful when adjusting for intermediate variables in multivariable models. Consider the situation where we are conducting a randomized placebo controlled trial of the effect of a new lipid-lowering drug on the incidence of CHD.

Although we may adjust for any discrepancies between the lipid levels of patients in the two treatment groups at the *start* of the trial (although this should not be necessary if randomization has been successful), we should not adjust for any *changes* in lipids that occur during the trial period. If we adjusted for these changes, we would be controlling for the beneficial effect of the drug, and thus any effect of the drug would probably 'disappear' (although this would provide an indication of how much of the drug's effect *can be explained* by its impact on lipid values which, in itself, may be useful).

Collinearity

When two explanatory variables are highly correlated, it may be difficult to evaluate their individual effects in a multivariable regression model. As a consequence, whilst each variable may be significantly associated with the dependent variable in a univariable model (i.e. when there is a single explanatory variable), neither may be significantly associated with it when both explanatory variables are included in a multivariable model. This **collinearity** (also called **multi-collinearity**) can be detected by examining the correlation coefficients between each pair of explanatory variables (commonly displayed in a correlation matrix) or by visual impression of the standard errors of the regression coefficients in the multivariable model; these will be substantially larger than those in the separate univariable models if collinearity is present. The easiest solution to this problem is to include only one of the variables in the model, although in situations where many of the variables are highly correlated, it may be necessary to seek statistical advice.

Prognostic indices and risk scores for a binary response

Given a large number of demographic or clinical features of an individual, we may want to *predict* whether that individual is likely to develop disease. Models, often fitted using proportional hazards regression (Chapter 44), logistic regression (Chapter 30) or a similar method known as **discriminant analysis**, can be used to identify factors that are significantly associated with outcome. A **prognostic index** or **risk score** can then be generated from the coefficients of this model, and the score calculated for an individual to assess his/her likelihood of disease. However, a model that explains a large proportion of the variability in the data may not necessarily be good at predicting which patients will develop disease. Therefore, once we have derived a predictive score based on a model, we should assess the **validity** of that score.

Validating the score

We can validate our score in a number of ways.

- We produce a prediction table based on our data set, showing the number of individuals in whom we correctly and incorrectly predict the disease status (similar to the table in Chapter 38). Measures, including sensitivity and specificity, can be calculated for this table, *or*
- We categorize individuals according to their score and consider disease rates in the different categories (see Example); we should see a relationship between the categories and disease rates, e.g. with higher scored categories having greater disease rates.

Clearly, any model will always perform well on the data set that was used to generate the model. Therefore, in order to provide a true assessment of the usefulness of the score, it should be validated on other, independent, data sets.

Where this is impractical, we may separate the data into two roughly equally sized subsamples. The first subsample, known as the **training sample**, is used to generate the model. The second subsample, known as the **validation sample**, is used for validating the results from the training sample. As a consequence of the smaller sample size, fewer explanatory variables can be included in the model.

Jackknifing

Jackknifing is a way of estimating parameters and providing confidence intervals in an unbiased manner. Each individual is removed from the sample, one at a time, and the remaining $(n - 1)$ individuals are used to estimate the parameters of the model. This process is repeated for each of the n individuals in the sample, and the resulting estimates of each parameter are averaged over all n samples. Because a score derived in this way is generated from many different data sets, it can be validated on the complete data set without taking subsamples.

Example

Although there are wide differences in prognosis between patients with AIDS, they are often thought of as a single homogeneous group. In order to group patients correctly according to their likely prognosis, a prognostic score was generated on the basis of the clinical experience of 363 AIDS patients at a single centre in London. A total of 159 (43.8%) of these patients died over a follow-up period of 6 years.

The score was the weighted sum of the number of each type (mild, moderate or severe) of AIDS-defining diseases the patient had experienced and his/her minimum CD4 cell count (measured in cells/mm^3) and was equal to:

Score = 300 × number of very severe AIDS events (lymphoma)
+100 × number of severe AIDS events (all events, other than those listed as very severe or mild)
+20 × number of mild AIDS events (oesophageal candida, cutaneous Kaposi's sarcoma, *Pneumocystis carinii* pneumonia, extrapulmonary tuberculosis)
−1 × minimum CD4 cell count measured since AIDS

In order to aid the interpretation of this score, and to validate it, three groups were identified.

AIDS Grade I Score < 0
AIDS Grade II Score 0–99
AIDS Grade III Score ≥ 100

Validation of the score was assessed by considering the death rate (number of deaths divided by the total person-years of follow-up) in each grade.

AIDS grade	Deaths	Follow-up (person-years)	Death rate
I	17	168.0	1.0
II	54	153.9	3.5
III	71	81.2	8.7

Thus there is a clear trend towards increasing death rates as the score increases. The score was also validated on a group of patients from a second London centre.

AIDS grade	Deaths	Follow-up (person-years)	Death rate
I	65	828.5	0.8
II	229	579.6	4.0
III	322	361.3	8.9

Remarkably similar results were seen, thus confirming the value of this scoring system.

Adapted from: Mocroft, A.J., Johnson, M.A., Sabin, C.A., *et al*. (1995) Staging system for clinical AIDS patients. *Lancet* **346**; 12–17 with permission from Elsevier.

35 Checking assumptions

Why bother?

Computer analysis of data offers the opportunity of handling large data sets that might otherwise be beyond our capabilities. However, do not be tempted to 'have a go' at statistical analyses simply because they are available on the computer. The validity of the conclusions drawn rely on the appropriate analysis being conducted in any given circumstance, and a requirement that the underlying assumptions inherent in the proposed statistical analysis are satisfied. We say that an analysis is **robust** to violations of its assumptions if its *P*-value and power (Chapter 18) are not appreciably affected by the violations. Performing a non-robust analysis could lead to misleading conclusions.

Are the data Normally distributed?

Many analyses make assumptions about the underlying distribution of the data. The following procedures verify approximate Normality, the most common of the distributional assumptions.

- We produce a dot plot (for small samples) or a histogram, stem-and-leaf plot (Fig. 4.2) or box plot to show the empirical frequency distribution of the data (Chapter 4). We conclude that the distribution is approximately Normal if it is bell-shaped and symmetrical. The median in a box plot should cut the rectangle defining the first and third quartiles in half, and the two whiskers should be of equal length if the data are Normally distributed.
- Alternatively, we can produce a **Normal plot** (preferably on the computer) which plots the cumulative frequency distribution of the data (on the horizontal axis) against that of the Normal distribution. Lack of Normality is indicated by the resulting plot producing a curve that appears to deviate from a straight line (Fig. 35.1).

Although both approaches are subjective, the Normal plot is more effective for smaller samples. The **Kolmogorov-Smirnov** and **Shapiro-Wilk** tests, both performed on the computer, can be used to assess Normality more objectively.

Are two or more variances equal?

We explained how to use the *t*-test (Chapter 21) to compare two means, and ANOVA (Chapter 22) to compare more than two means. Underlying these analyses is the assumption that the variability of the observations in each group is the same, i.e. we require equal variances, described as **homogeneity of variance** or **homoscedasticity**. We have **heterogeneity of variance** if the variances are unequal.

- We can use **Levene's test**, using a computer program, to test for homogeneity of variance in two or more groups. The null hypothesis is that all the variances are equal. Levene's test has the advantage that it is not strongly dependent on the assumption of Normality. **Bartlett's test** can also be used to compare more than two variances, but it is non-robust to departures from Normality.
- We can use the ***F*-test (variance-ratio test)** described in the box to compare two variances, provided the data in each group are approximately Normally distributed (the test is non-robust to a violation of this assumption). The two estimated variances are s_1^2 and s_2^2, calculated from n_1 and n_2 observations, respectively. By convention, we choose s_1^2 to be the larger of the two variances, if they differ.
- We also assume homogeneity of variance of the *residuals* in simple and multiple regression (Chapters 28, 29) and in random effects models (Chapter 42). We explained how to check this assumption in Chapters 28 and 29.

1 Define the null and alternative hypotheses under study

H_0: the two population variances are equal

H_1: the two population variances are unequal.

2 Collect relevant data from a sample of individuals

3 Calculate the value of the test statistic specific to H_0

$$F = s_1^2 / s_2^2$$

which follows an *F*-distribution with $n_1 - 1$ *df* in the numerator, and $n_2 - 1$ *df* in the denominator. By choosing $s_1^2 \geq s_2^2$, we have ensured that the *F*-ratio will always be ≥1. This allows us to use the tables of the *F*-distribution which are tabulated only for values ≥1.

4 Compare the value of the test statistic to values from a known probability distribution

Refer *F* to Appendix A5. Our two-sided alternative hypothesis leads to a two-tailed test.

5 Interpret the *P*-value and results

Note that we are rarely interested in the variances *per se*, so we do not usually calculate confidence intervals for them.

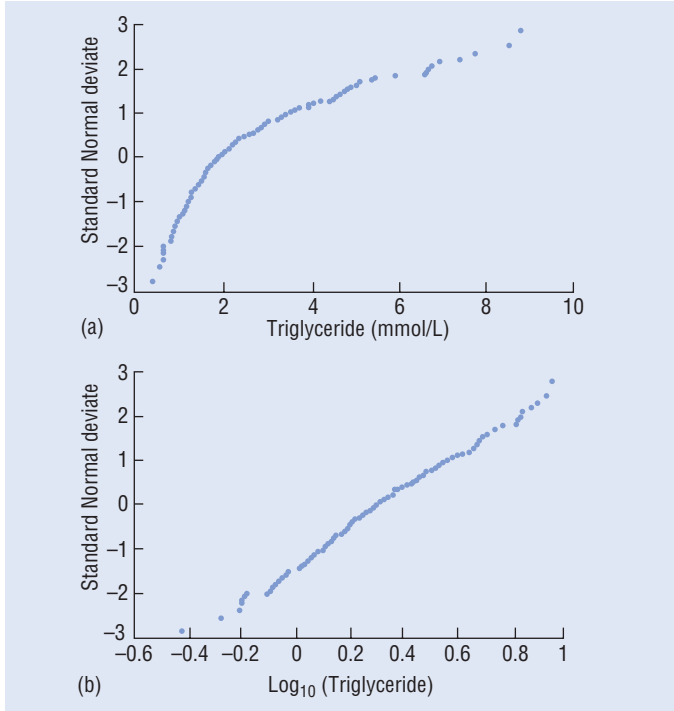


Figure 35.1 (a) Normal plot of untransformed triglyceride levels described in Chapter 19. These are skewed and the resulting Normal plot shows a distinct curve. (b) Normal plot of log (triglyceride levels). The approximately straight line indicates that the log transformation has been successful at removing the skewness in the data.

Are variables linearly related?

Most of the techniques which we discussed in Chapters 26–31 and describe in Chapter 42 assume that there is a **linear** (straight line) relationship between two variables. Any inferences drawn from such analyses rely on the linearity assumption being satisfied. We explained how to check for linearity and how to deal with non-linearity in regression analysis in Chapters 28 and 29 (for simple and multiple regression) and in Chapter 33 (for other generalized linear models, e.g. logistic and Poisson).

What if the assumptions are not satisfied?

We have various options.

- Proceed as planned, recognizing that this may result in a non-robust analysis. Be aware of the implications if you do this. Do not be fooled into an inappropriate analysis just because others, in similar circumstances, have done one in the past!
- Take an appropriate transformation of the raw data so that the transformed data satisfy the assumptions of the proposed analysis (Chapter 9). In regression analysis, this usually means transforming an x variable although other approaches are possible (see Chapter 32);
- If feasible, perform a **non-parametric test** (Chapter 17) that does not make any assumptions about the distribution of the data (e.g. Normality). You may also come across the term *non-parametric regression analysis*[1]; its purpose is to estimate the functional form (rather than the parameters) of the relationship between a response variable and one or more explanatory variables. Using non-parametric regression, we relax the linearity assumption of the model and fit a smooth curve to the data so that we can visualize trends without requiring the specification of a parametric model.

[1] Eubank, R.L. (1999) *Nonparametric Regression and Spline Smoothing*, Marcel Dekker.

Example

Consider the unpaired *t*-test example of Chapter 21. A total of 98 school age children were randomly assigned to receive either inhaled beclomethasone dipropionate or a placebo to determine their effects on wheezing. We used the unpaired *t*-test to compare the mean forced expiratory volume (FEV1) in each group over the 6 months, but need assurance that the underlying assumptions (Normality and constant variance) are satisfied. The stem-and-leaf plots in Fig. 4.2 show that the data in each group are approximately Normally distributed. We performed the ***F*-test** to investigate the assumption of constant variance in the two groups.

1 H_0: the variance of FEV1 measurements in the population of school age children is the same in the two treatment groups

H_1: the variance of FEV1 measurements in the population of school age children is not the same in two treatment groups.

2 Treated group: sample size, $n_1 = 50$, standard deviation, $s_1 = 0.29$ litres

Placebo group: sample size, $n_2 = 48$, standard deviation, $s_2 = 0.25$ litres.

3 The test statistic, $F = \frac{s_1^2}{s_2^2} = \frac{0.29^2}{0.25^2} = \frac{0.0841}{0.0625} = 1.336$ which follows an F-distribution with $50 - 1 = 49$ and $48 - 1 = 47$ *df* in the numerator and denominator, respectively.

4 We refer $F = 1.34$ to Appendix A5 for a two-sided test at the 5% level of significance. Because Appendix A5 is restricted to entries of 25 and infinity *df* in the numerator, and 30 and 50 *df* in the denominator, we have to interpolate (Chapter 21). The required tabulated value at the 5% level of significance lies between 1.57 and 2.12; thus $P > 0.05$ because 1.34 is less than the minimum of these values (computer output gives $P = 0.32$).

5 There is insufficient evidence to reject the null hypothesis that the variances are equal. It is reasonable to use the unpaired *t*-test, which assumes Normality and homogeneity of variance, to compare the mean FEV1 values in the two groups.

36 Sample size calculations

The importance of sample size

If the number of patients in our study is small, we may have inadequate power (Chapter 18) to detect an important existing effect, and we shall have wasted all our resources. On the other hand, if the sample size is unduly large, the study may be unnecessarily time-consuming, expensive and unethical, depriving some of the patients of the superior treatment. We therefore have to choose the optimal sample size which strikes a balance between the implications of making a Type I or Type II error (Chapter 18). Unfortunately, in order to calculate the sample size required, we have to have some idea of the results we expect in the study.

Requirements

We shall explain how to calculate the optimal sample size in simple situations; often more complex designs can be simplified for the purpose of calculating the sample size. If our investigation involves a number of tests, we concentrate on the most important or evaluate the sample size required for each and choose the largest.

Our focus is the calculation of the optimal sample size in relation to a proposed hypothesis test. However, it is possible to base the sample size calculation on other aspects of the study, such as on the precision of an estimate or on the width of a confidence interval (the process usually adopted in equivalence and non-inferiority studies, Chapter 17).

To calculate the optimal sample size for a test, we need to specify the following quantities *at the design stage of the investigation*.

- **Power** (Chapter 18)—the chance of detecting, as statistically significant, a specified effect if it exits. We usually choose a power of at least 80%.
- **Significance level**, α (Chapter 17)—the cut-off level below which we will reject the null hypothesis, i.e. it is the maximum probability of incorrectly concluding that there is an effect. We usually fix this as 0.05 or, occasionally, as 0.01, and reject the null hypothesis if the *P*-value is less than this value.
- **Variability** of the observations e.g. the standard deviation, if we have a numerical variable.
- **Smallest effect of interest**—the magnitude of the effect that is clinically important and which we do not want to overlook. This is often a *difference* (e.g. difference in means or proportions). Sometimes it is expressed as a multiple of the standard deviation of the observations (the **standardized difference**).

It is relatively simple to choose the power and significance level of the test that suit the requirements of our study. The choice is usually governed by the implications of a Type I and a Type II error, but may be specified by the regulatory bodies in some drug licensing studies. Given a particular clinical scenario, it is possible to specify the effect we regard as clinically important. The real difficulty lies in providing an estimate of the variation in a numerical variable before we have collected the data. We may be able to obtain this information from other published studies with similar outcomes or we may need to carry out a **pilot study**.

Methodology

We can calculate sample size in a number of ways, each of which requires essentially the same information (described in Requirements) in order to proceed:

- **General formulae**[1]—these can be complex but may be necessary in some situations {e.g. to retain power in a cluster randomized trial (Chapters 14 and 41), we multiply the sample size that would be required if we were carrying out individual randomization by the design effect equal to $[1 + (m - 1)\,\rho]$, where m is the average cluster size and ρ is the intraclass correlation coefficient (Chapter 42)}.
- **Quick formulae**—these exist for particular power values and significance levels for some hypothesis tests (e.g., Lehr's formulae[2], see below).
- **Special tables**[1]—these exist for different situations (e.g., for *t*-tests, Chi-squared tests, test of the correlation coefficient, comparing two survival curves and for equivalence studies).
- **Altman's nomogram**—this is an easy to use diagram which is appropriate for various tests. Details are given in the next section.
- **Computer software**—this has the advantage that results can be presented graphically or in tables to show the effect of changing the factors (eg. power, size of effect) on the required sample size.

Altman's nomogram

Notation

We show in Table 36.1 the notation for using Altman's nomogram (Appendix B) to estimate the sample size of two *equally sized* groups of observations for three frequently used hypothesis tests of means and proportions.

Method

For each test, we calculate the standardized difference and join its value on the left hand axis of the nomogram to the power we have specified on the right hand vertical axis. The required sample size is indicated at the point at which the resulting line and sample size axis meet.

Note that we can also use the nomogram to evaluate the power of a hypothesis test for a given sample size. Occasionally, this is useful if we wish to know, retrospectively, whether we can attribute lack of significance in a hypothesis test to an inadequately sized sample. Remember, also, that a wide confidence interval for the effect of interest indicates low power (Chapter 11).

Quick formulae

For the unpaired *t*-test and Chi-squared test, we can use **Lehr's formula**[2] for calculating the sample size for a power of 80% and a two-sided significance level of 0.05. The required sample size in each group is:

$$\frac{16}{(\textit{Standardized difference})^2}$$

If the standardized difference is small, this formula overestimates the sample size. Note that a numerator of 21 (instead of 16) relates to a power of 90%.

[1] Machin, D., Campbell, M.J., Fayers, P.M. and Pinol, A.P.Y. (1997) *Sample size Tables for Clinical Studies,* 2nd edn, Blackwell, Oxford.
[2] Lehr, R. (1992) Sixteen s squared over d squared: a relation for crude sample size estimates. Statistic in Medicine **11**:1099–1102.

Table 36.1 Information for using Altman's nomogram.

Hypothesis test	Standardized difference	Explanation of N in nomogram	Terminology
Unpaired *t*-test (Chapter 21)	$\frac{\delta}{\sigma}$	*N*/2 observations in each group	δ: the smallest difference in means which is clinically important. σ: the assumed equal standard deviation of the observations in each of the two groups. You can estimate it using results from a similar study conducted previously or from published information. Alternatively, you could perform a pilot study to estimate it. Another approach is to express δ as a multiple of the standard deviation (e.g, the ability to detect a difference of two standard deviations).
Paired *t*-test (Chapter 20)	$\frac{2\delta}{\sigma_d}$	*N* pairs of observations	δ: the smallest mean difference which is clinically important. σ_d: the standard deviation of the *differences* in response, usually estimated from a pilot study.
Chi-squared test (Chapter 24)	$\frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})}}$	*N*/2 observations in each group	$p_1 - p_2$: the smallest difference in the proportions of 'success' in the two groups that is clinically important. One of these proportions is often known, and the relevant difference evaluated by considering what value the other proportion must take in order to constitute a noteworthy change. $\bar{p} = \frac{p_1 + p_2}{2}$

Power statement

It is often essential and always useful to include a power statement in a study protocol or in the methods section of a paper (see CONSORT statement, Chapter 14) to show that careful thought has been given to sample size at the design stage of the investigation. A typical statement might be '84 patients in each group were required for the unpaired *t*-test to have a 90% chance of detecting a difference in means of 2.5 days (SD = 5 days) at the 5% level of significance' (see Example 1).

Adjustments

We may wish to adjust the sample size:

- to allow for **losses to follow-up** by recruiting more patients into the study at the outset. If we believe that the drop-out rate will be *r*%, then the adjusted sample size is obtained by multiplying the unadjusted sample size by $100/(100 - r)$.
- to have independent **groups of different sizes**. This may be desirable when one group is restricted in size, perhaps because the disease is rare in a case–control study (Chapter 16) or because the novel drug treatment is in short supply. Note, however, that the imbalance in numbers usually results in a larger overall sample size when compared to a balanced design if a similar level of power is to be maintained. If the ratio of the sample sizes in the two groups is k (e.g. $k = 3$ if we require one group to be three times the size of the other), the adjusted overall sample size is

$$N' = N(1+k)^2 / (4k)$$

where N is the unadjusted overall sample size calculated for equally sized groups. Then $N'/(1+k)$ of these patients will be in the smaller group and the remaining patients will be in the larger group.

Example 1

Comparing means in independent groups using the unpaired *t*-test

Objective—to examine the effectiveness of aciclovir suspension (15 mg/kg) for treating 1–7-year-old children with herpetic gingivostomatitis lasting less than 72 h.

Design—randomized, double-blind placebo-controlled trial with 'treatment' administered five times a day for 7 days.

Main outcome measure for the determination of sample size—duration of oral lesions.

Sample size question—how many children are required to have a 90% power of detecting a 2.5 day difference in mean duration of oral lesions between the two groups at the 5% level of significance? The authors assume that the standard deviation of duration of oral lesions is approximately 5 days.

continued

Using the nomogram:
$\delta = 2.5$ days and $\sigma = 5$ days. Thus the standardized

difference equals $\dfrac{\delta}{\sigma} = \dfrac{2.5}{5} = 0.50$

The line connecting a standardized difference of 0.50 and a power of 90% cuts the sample size axis at approximately 160. Therefore about 80 children are required in each group [*Note*: (i) if δ were increased to 3 days, the standardized difference equals 0.6 and the required sample size would decrease to approximately 118 in total, i.e. 59 in each group, and (ii) if, using the original specification, the investigators wanted twice as many children on aciclovir treatment as on placebo (i.e., $k = 2$), then the adjusted sample size would be

$$N' = N(1+k)^2 / (4k) = 160(1+2)^2 / (4 \times 2) = 180,$$

with 180/3 = 60 children on placebo and the remaining 120 children on aciclovir]. Figure 18.1 shows power curves for this example.

Quick formula:
If the power is 90%, the required sample size in each group is:

$$\frac{21}{(\textit{standardized difference})^2} = \frac{21}{(0.50)^2} = 84.$$

Amir, J., Harel, L., Smettana, Z. and Varsano, I. (1997) Treatment of herpes simplex gingivostomatitis with aciclovir in children: a randomized double-blind placebo controlled study *British Medical Journal*, **314**, 1800–1803.

Example 2
Comparing two proportions in independent groups using the Chi-squared test

Objective—to compare the effectiveness of corticosteroid injections with physiotherapy for the treatment of painful stiff shoulder.

Design—Randomized controlled trial (RCT) in which patients are randomly allocated to 6 weeks of treatment, these comprising either a maximum of three injections or twelve 30 min sessions of physiotherapy for each patient.

Main outcome measure for determining sample size—treatment is regarded as a success after 7 weeks if the patient rates him/herself as having made a complete recovery or as having much improvement (on a six-point Likert scale).

Sample size question—how many patients are required in order to have an 80% power of detecting a clinically important difference in success rates of 25% between the two groups at the 5% level of significance? The authors assume a success rate of 40% in the group having the least successful treatment.

Using the nomogram:

$p_1 = 0.40$ and $p_2 = 0.65$, so $\bar{p} = \dfrac{0.40 + 0.65}{2} = 0.525$

Therefore, the standardized difference

$$= \frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})}} = \frac{0.25}{\sqrt{0.525 \times 0.475}} = 0.50$$

The line connecting a standardized difference of 0.50 and a power of 80% cuts the sample size axis at 120. Hence approximately 60 patients are required in each group [*Note*: (i) if the power were increased to 85%, the required sample size would increase to approximately 140 in total, i.e. 70 patients would be required in each group, and (ii) if the drop-out rate was expected to be around 20%, the adjusted overall sample size (with a power of 80%) would be $120 \times 100/(100 - 20) = 150$, with 75 patients in each group). Figure 18.2 shows power curves for this example.

Quick formula:
If the power is 80%, the required sample size in each group is:

$$\frac{16}{(\textit{standardized difference})^2} = \frac{16}{(0.50)^2} = 64.$$

van der Windt, D.A.W.M., Koes, B.W., Devillé, W. de Jong, B.A. and Bouter, M. (1998) Effectiveness of corticosteroid injections with physiotherapy for treatment of painful shoulder in primary care: randomised trial. *British Medical Journal,* **317**, 1292–6.

37 Presenting results

Introduction

An essential facet of statistics is the ability to summarize the important features of the analysis. We must know what to include and how to display our results in a manner that enables others to obtain relevant and important information easily and draw correct conclusions. This chapter describes the key features of presentation.

Numerical results

- Give figures only to the degree of accuracy that is appropriate (as a guideline, one significant figure more than the raw data). If analysing the data by hand, only round up or down at the end of the calculations.
- Give the number of items on which any summary measure (e.g. a percentage) is based.
- Describe any outliers and explain how they are handled (Chapter 3).
- Include the units of measurement.
- When interest is focused on a parameter (e.g. the mean, regression coefficient), always indicate the precision of its estimate. We recommend using a confidence interval for this but the standard error is also acceptable. *Avoid* using the ± symbol, as in mean ± SEM (Chapter 10), because by adding and subtracting the SEM, we create a 67% confidence interval that can be misleading for those used to 95% confidence intervals. It is better to show the standard error in brackets after the parameter estimate [e.g. mean = 16.6 g (SEM 0.5 g)].
- When interest is focused on the distribution of observations, always indicate a measure of the 'spread' of the data. The range of values that excludes outliers (typically, the range of values containing the central 95% of the observations—Chapter 6) is a useful descriptor. If the data are Normally distributed, this range is approximated by the sample mean ±1.96 × standard deviation (Chapter 7). You can quote the mean and the standard deviation [e.g. mean = 35.9 mm (SD 2.8 mm)] instead but this leaves the reader to evaluate the range.

Tables

- Do not give too much information in a table.
- Include a concise, informative, and unambiguous title.
- Label each row and column.
- Remember that it is easier to scan information down columns rather than across rows.

Diagrams

- Keep a diagram simple and avoid unnecessary frills (e.g. making a pie chart three-dimensional).
- Include a concise, informative, and unambiguous title.
- Label all axes, segments, and bars, and explain the meaning of symbols.
- Avoid distorting results by exaggerating the scale on an axis.
- Indicate where two or more observations lie in the same position on a scatter diagram, e.g. by using a different symbol.
- Ensure that all the relevant information is contained in the diagram (e.g. link paired observations).

Presenting results in a paper

When presenting results in a paper, we should ensure that the paper contains enough information for the reader to understand what has been done. S/he should be able to reproduce the results, given the appropriate computer package and data. All aspects of the design of the study and the statistical methodology must be fully described (see also CONSORT Statement—Chapter 14).

Results of a hypothesis test

- Include a relevant diagram, if appropriate.
- Indicate the hypotheses of interest.
- Name the test and state whether it is one- or two-tailed.
- Justify the assumptions (if any) underlying the test (e.g. Normality, constant variance; Chapter 35), and describe any transformations (Chapter 9) required to meet these assumptions (e.g. taking logarithms).
- Specify the observed value of the test statistic, its distribution (and degrees of freedom, if relevant), and, if possible, the *exact P*-value (e.g. $P = 0.03$) rather than an interval estimate of it (e.g. $0.01 < P < 0.05$) or a star system (e.g. *, **, *** for increasing levels of significance). Avoid writing 'n.s.' when $P > 0.05$; an exact *P*-value is preferable even when the result is non-significant.
- Include an estimate of the *relevant* effect of interest (e.g. the difference in means for the two-sample *t*-test, or the mean difference for the paired *t*-test) with a confidence interval (preferably) or standard error.
- Draw conclusions from the results (e.g. reject the null hypothesis), interpret any confidence intervals and explain their implications.

Results of a regression analysis

Here we include simple (Chapters 27 and 28) and multiple linear regression (Chapter 29), logistic regression (Chapter 30), Poisson regression (Chapter 31), proportional hazards regression (Chapter 44) and regression methods for clustered data (Chapter 42). Full details of these analyses are explained in the associated chapters.

- Include relevant diagrams (e.g. a scatter plot with the fitted line for simple regression).
- Clearly state which is the dependent variable and which is (are) the explanatory variable(s).
- Justify underlying assumptions and explain the results of regression diagnostics, if appropriate.
- Describe any transformations, and explain their purpose.
- Where appropriate, describe the possible numerical values taken by any categorical variable (e.g. male = 0, female = 1), how dummy variables were created (Chapter 29), and the units of continuous variables.
- Give an indication of the goodness of fit of the model (e.g. quote R^2 (Chapter 29) or LRS (Chapter 32)).
- If appropriate (e.g. in multiple regression), give the results of the overall *F*-test from the ANOVA table.
- Provide estimates of *all* the coefficients in the model (including those which are not significant, if applicable) together with the confidence intervals for the coefficients or standard errors of their estimates. In logistic regression (Chapter 30), Poisson regression (Chapter

31) and proportional hazards regression (Chapter 44), convert the coefficients to estimated odds ratios, relative rates or relative hazards (with confidence intervals). Interpret the relevant coefficients.

• Show the results of the hypothesis tests on the coefficients (i.e. include the test statistics and the *P*-values). Draw appropriate conclusions from these tests.

Complex analyses

There are no simple rules for the presentation of the more complex forms of statistical analysis. Be sure to describe the design of the study fully (e.g. the factors in the analysis of variance and whether there is a hierarchical arrangement), and include a validation of underlying assumptions, relevant descriptive statistics (with confidence intervals), test statistics and *P*-values. A brief description of what the analysis is doing helps the uninitiated; this should be accompanied by a reference for further details. Specify which computer package has been used.

Example

Informative and unambiguous title

Table 34.1 : Information relating to first births in women with bleeding disorders†, stratified by bleeding disorder

Rows and columns fully labelled

Numbers on which percentages are based

		Bleeding disorder			
	Total	Haem A	Haem B	vWD	FXI deficiency
Number of women with live births	48	14	5	19	10
Mother's age at birth of baby (years)					
Median	27.0	24.9	28.5	27.5	27.1
range	(16.7–37.9)	(16.7–33.0)	(25.6–34.9)	(18.8–36.6)	(22.3–37.9)
Gestational age of baby (weeks)					
Median	40	39	40	40	40.5
(range)	(37–42)	(38–42)	(39–41)	(38–42)	(37–42)
Weight of baby (kg)					
Median	3.64	3.62	3.78	3.64	3.62
(range)	(1.96–4.46)	(1.96–4.46)	(3.15–3.94)	(2.01–4.35)	(2.90–3.84)
Sex of baby*					
Boy	20 (41.7%)	8 (57.1%)	0 (–)	8 (42.1%)	4 (40.0%)
Girl	20 (41.7%)	4 (28.6%)	2 (40.0%)	10 (52.6%)	4 (40.0%)
Not stated	8 (16.7%)	2 (14.3%)	3 (60.0%)	1 (5.3%)	2 (20.0%)
Interventions received during labour*					
Inhaled gas	25 (52.1%)	6 (42.9%)	2 (40.0%)	11 (57.9%)	6 (60.0%)
Intramuscular pethidine	22 (45.8%)	9 (64.3%)	1 (20.0%)	4 (21.1%)	8 (80.0%)
Intravenous pethidine	2 (4.2%)	0 (0.0%)	0 (0.0%)	1 (5.3%)	1 (10.0%)
Epidural	10 (20.8%)	3 (21.4%)	2 (40.0%)	4 (21.1%)	1 (10.0%)

Units of measurement

Estimates of location and spread

Numerical results quoted to appropriate degree of accuracy

*Entries are frequencies (%)
†The study is described in Chapter 2

continued

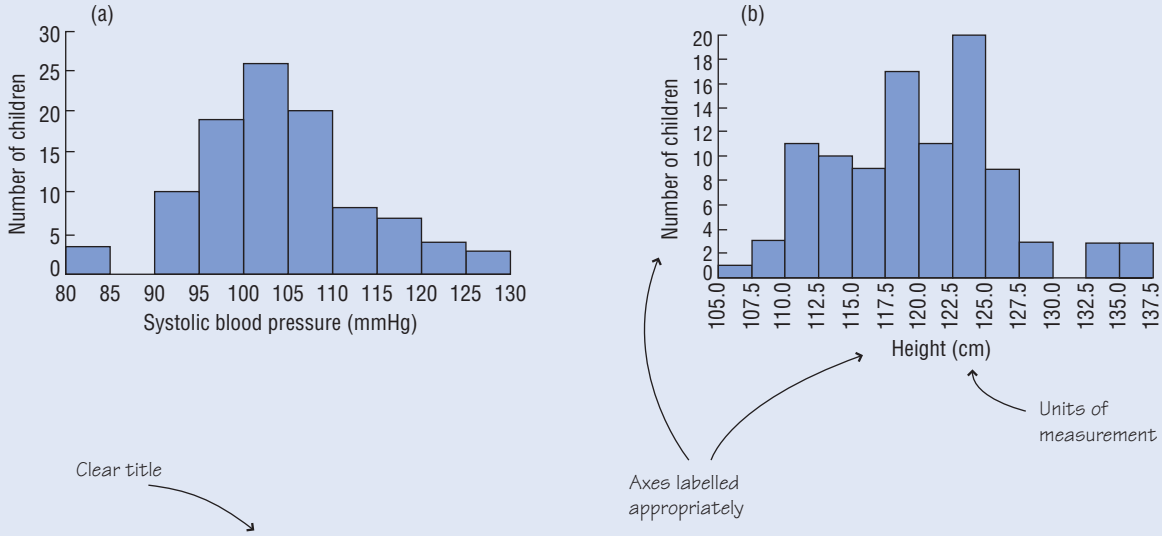


Figure 37.1 Histograms showing the distribution of (a) systolic blood pressure and (b) height in a sample of 100 children (Chapter 26).

38 Diagnostic tools

An individual's state of health is often characterized by a number of numerical or categorical measures. In this context, an appropriate **reference interval** (Chapters 6 and 7) and/or **diagnostic test** may be used:

- by the clinician, together with a clinical examination, to **diagnose** or exclude a particular disorder in his/her patient;
- as a **screening** device to ascertain which individuals in an apparently healthy population are likely to have (or sometimes, not have) the disease of interest. Individuals flagged in this way will then usually be subjected to more rigorous investigations in order to have their diagnoses confirmed. It is only sensible to screen for a disease if adequate facilities for treating the disease at the pre-symptomatic stages exist, this treatment being less costly and/or more effective than when given at a later stage (or, occasionally, if it is believed that individuals who are diagnosed with the disease will modify their behaviour to prevent the disease spreading).

Reference intervals

A **reference interval** (often referred to as a **normal range**) for a single numerical variable, calculated from a very large sample, provides a range of values that are typically seen in healthy individuals. If an individual's value is above the upper limit, or below the lower limit, we consider it to be unusually high (or low) relative to healthy individuals.

Calculating reference intervals

Two approaches can be taken.

- We make the assumption that the data are Normally distributed. Approximately 95% of the data values lie within 1.96 standard deviations of the mean (Chapter 7). We use our data to calculate these two limits (mean $\pm$ 1.96 $\times$ standard deviation).
- An alternative approach, which does not make any assumptions about the distribution of the measurement, is to use a central range which encompasses 95% of the data values (Chapter 6). We put our values in order of magnitude and use the 2.5th and 97.5th percentiles as our limits.

The effect of other factors on reference intervals

Sometimes the values of a numerical variable depend on other factors, such as age or sex. It is important to interpret a particular value only after considering these other factors. For example, we generate reference intervals for systolic blood pressure separately for men and women.

Diagnostic tests

The **gold-standard test** that provides a definitive diagnosis of a particular condition may sometimes be impractical or not routinely available. We would like a simple test, depending on the presence or absence of some marker, which provides a reasonable guide to whether or not the patient has the condition.

We take a group of individuals whose true disease status is known from the gold standard test. We can draw up the 2 $\times$ 2 table of frequencies (Table 38.1):

Table 38.1 Table of frequencies.

	Gold standard test		
Test result	Disease	No disease	Total
Positive	a	b	$a+b$
Negative	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

Of the n individuals studied, $a + c$ individuals have the disease. The **prevalence** (Chapter 12) of the disease in this sample is $= \dfrac{(a+c)}{n}$.

Of the $a + c$ individuals who have the disease, a have positive test results (**true positives**) and c have negative test results (**false negatives**). Of the $b + d$ individuals who do not have the disease, d have negative test results (**true negatives**) and b have positive test results (**false positives**).

Assessing reliability: sensitivity and specificity

Sensitivity = proportion of individuals with the disease who are correctly identified by the test

$$= \frac{a}{(a+c)}$$

Specificity = proportion of individuals without the disease who are correctly identified by the test

$$= \frac{d}{(b+d)}$$

These are usually expressed as percentages. As with all estimates, we should calculate confidence intervals for these measures (Chapter 11).

We would like to have a sensitivity and specificity that are both as close to 1 (or 100%) as possible. However, in practice, we may gain sensitivity at the expense of specificity, and *vice versa*. Whether we aim for a high sensitivity or high specificity depends on the condition we are trying to detect, along with the implications for the patient and/or the population of either a false negative or false positive test result. For conditions that are easily treatable, we prefer a high sensitivity; for those that are serious and untreatable, we prefer a high specificity in order to avoid making a false positive diagnosis. It is important that before screening is undertaken, subjects should understand the implications of a positive diagnosis, as well as having an appreciation of the false positive and false negative rates of the test.

Predictive values

Positive predictive value = proportion of individuals with a positive test result who have the disease

$$= \frac{a}{(a+b)}$$

Negative predictive value = proportion of individuals with a negative test result who do not have the disease

$$= \frac{d}{(c+d)}$$

We calculate confidence intervals for these predictive values, often expressed as percentages, using the methods described in Chapter 11.

These predictive values provide information about how likely it is that the individual has or does not have the disease, given his/her test result. Predictive values are dependent on the prevalence of the disease in the population being studied. In populations where the disease is common, the positive predictive value of a given test will be much higher than in populations where the disease is rare. The converse is true for negative predictive values.

The use of a cut-off value

Sometimes we wish to make a diagnosis on the basis of a continuous measurement. Often there is no threshold above (or below) which disease definitely occurs. In these situations, we need to define a cut-off value ourselves, above (or below) which we believe an individual has a very high chance of having the disease.

A useful approach is to use the upper (or lower) limit of the reference interval. We can evaluate this cut-off value by calculating its associated sensitivity, specificity and predictive values. If we choose a different cut-off, these values may change as we become more or less stringent. We choose the cut-off to optimize these measures as desired.

Receiver operating characteristic curves

These provide a way of assessing whether a particular type of test provides useful information, and can be used to compare two different tests, and to select an optimal cut-off value for a test.

For a given test, we consider all cut-off points that give a unique pair of values for sensitivity and specificity, and plot the *sensitivity* against *one minus the specificity* (thus comparing the probabilities of a positive test result in those with and without disease) and connect these points by lines (Fig. 38.1).

The receiver operating characteristic (ROC) curve for a test that has some use will lie to the left of the diagonal (i.e. the 45° line) of the graph. Depending on the implications of false positive and false negative results, and the prevalence of the condition, we can choose the optimal cut-off for a test from this graph. Two or more tests for the same condition can be compared by considering the area under each curve; this area is given by the C statistic (produced by many statistical packages). The test with the greater area (i.e. the higher C statistic) is better at discriminating between disease outcomes.

Is a test useful?

The **likelihood ratio** (LR) for a positive result is the ratio of the chance of a positive result if the patient has the disease to the chance of a positive result if s/he does not have the disease (see also Chapter 32). Likelihood ratios can also be generated for negative test results. For example, a LR of 2 for a positive result indicates that a positive result is twice as likely to occur in an individual with disease than in one without it. A high likelihood ratio for a positive result suggests that the test provides useful information, as does a likelihood ratio close to zero for a negative result.

It can be shown that:

$$\textbf{LR for a positive result} = \frac{Sensitivity}{(1 - specificity)}$$

We discuss this LR in a Bayesian framework in Chapter 45.

Example

Cytomegalovirus (CMV) is a common viral infection to which approximately 50% of individuals are exposed during childhood. Although infection with the virus does not usually lead to any major problems, individuals who have been infected with CMV in the past may suffer serious disease after certain transplant procedures, such as bone marrow transplantation, if their virus is either reactivated or if they are re-infected by their donors. It is thought that the amount of detectable virus in their blood after transplantation (the viral load) may predict which individuals will get severe disease. In order to study this hypothesis, CMV viral load was measured in a group of 49 bone marrow transplant recipients. Fifteen of the 49 patients developed severe disease during follow-up. Viral load values in all patients ranged from 2.7 $\log_{10}$ genomes/mL to 6.0 $\log_{10}$ genomes/mL. As a starting point, a value in excess of 4.5 $\log_{10}$ genomes/mL was considered an indication of the possible future development of disease. The table of frequencies shows the results obtained; the box contains calculations of estimates of measures of interest.

	Severe disease		
Viral load ($\log_{10}$ genomes/mL)	Yes	No	Total
>4.5	7	6	13
≤4.5	8	28	36
Total	15	34	49

Prevalence = (15/49) × 100% = 31% (95% CI 18% to 45%)

Sensitivity = (7/15) × 100% = 47% (95% CI 22% to 72%)

Specificity = (28/34) × 100% = 82% (95% CI 69% to 95%)

Positive predictive value = (7/13) × 100% = 54% (95% CI 27% to 81%)

Negative predictive value = (28/36) × 100% = 78% (95% CI 65% to 92%)

Likelihood ratio for positive result = 0.47/(1–0.82) = 2.6 (95% CI 1.1% to 6.5%, obtained from computer output)

Therefore, for this cut-off value, we have a relatively high specificity and a moderate sensitivity. The LR of 2.6 indicates that this test is useful, in that a viral load >4.5 $\log_{10}$ genomes/mL is more than twice as likely in an individual with severe disease than in one without severe disease. However, in order to investigate other cut-off values, a ROC curve was plotted (Fig. 38.1). The plotted line falls just to the left of the diagonal of the graph. For this example, the most useful cut-off value (5.0 $\log_{10}$ genomes/mL) is that which gives a sensitivity of 40% and a specificity of 97%; then the LR equals 13.3.

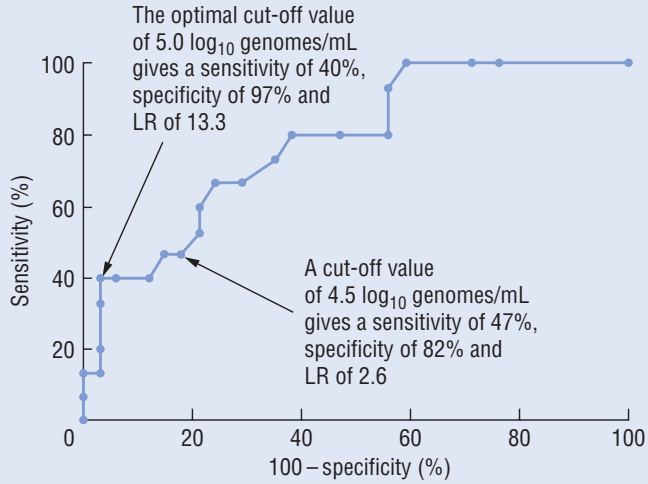


Figure 38.1 Receiver operating characteristic (ROC) curve, highlighting the results from two possible cut-off values, the optimal one and that used in the diagnostic test.

Data kindly provided by Prof. V.C. Emery and Dr D. Gor, Department of Virology, Royal Free and University College Medical School, London, UK.

39 Assessing agreement

Introduction

There are many occasions on which we wish to compare results which should concur. In particular, we may want to assess and, if possible, quantify the following two types of **agreement** or **reliability**:

- **Reproducibility (method/observer agreement)**. Do two techniques used to measure a particular variable, in otherwise identical circumstances, produce the same result? Do two or more observers using the same method of measurement obtain the same results?
- **Repeatability.** Does a single observer obtain the same results when s/he takes repeated measurements in identical circumstances?

Reproducibility and repeatability can be approached in the same way. In each case, the method of analysis depends on whether the variable is **categorical** (e.g. poor/average/good) or **numerical** (e.g. systolic blood pressure). For simplicity, we shall restrict the problem to that of comparing only **paired** results (e.g. two methods/two observers/duplicate measurements).

Categorical variables

Suppose two observers assess the same patients for disease severity using a categorical scale of measurement, and we wish to evaluate the extent to which they agree. We present the results in a two-way contingency table of frequencies with the rows and columns indicating the categories of response for each observer. Table 39.1 is an example showing the results of two observers' assessments of the condition of tooth surfaces. The frequencies with which the observers agree are shown along the **diagonal** of the table. We calculate the corresponding frequencies which would be **expected** if the categorizations were made at random, in the same way as we calculated expected frequencies in the Chi-squared test of association (Chapter 24)—i.e. each expected frequency is the product of the relevant row and column totals divided by the overall total. Then we measure agreement by:

$$\textit{Cohen's kappa, } \kappa = \frac{\left(\frac{O_d}{m} - \frac{E_d}{m}\right)}{\left(1 - \frac{E_d}{m}\right)}$$

which represents the chance corrected proportional agreement, where:

- m = total observed frequency (e.g. total number of patients)
- O_d = sum of observed frequencies *along the diagonal*
- E_d = sum of expected frequencies *along the diagonal*
- 1 in the denominator represents maximum agreement.

$\kappa = 1$ implies perfect agreement and $\kappa = 0$ suggests that the agreement is no better than that which would be obtained by chance. There are no objective criteria for judging intermediate values. However, kappa is often judged as providing agreement[1] which is:

- Poor if $\kappa \leq 0.20$,
- Fair if $0.21 \leq \kappa \leq 0.40$,
- Moderate if $0.41 \leq \kappa \leq 0.60$,
- Substantial if $0.61 \leq \kappa \leq 0.80$,
- Good if $\kappa > 0.80$.

Although it is possible to estimate a standard error for kappa, we do not usually test the hypothesis that kappa is zero since this is not really pertinent or realistic in a reliability study.

Note that kappa is dependent both on the number of categories (i.e., its value is greater if there are less categories) and the prevalence of the condition, so care must be taken when comparing kappas from different studies. For ordinal data, we can also calculate a **weighted kappa**[2] which takes into account the extent to which the observers **disagree** (the non-diagonal frequencies) as well as the frequencies of agreement (along the diagonal). The weighted kappa is very similar to the **intraclass correlation coefficient** (see next section and Chapter 42).

Numerical variables

Suppose an observer takes duplicate measurements of a numerical variable on n individuals (just replace the word 'repeatability' by 'reproducibility' if considering the similar problem of method agreement, but remember to assess the repeatability of each method before carrying out the method agreement study).

- If the average difference between duplicate measurements is zero (assessed by the paired t-test, sign test or signed ranks test—Chapters 19 and 20) then we can infer that there is *no systematic difference* between the pairs of results: if one set of readings represents the true values, as is likely in a method comparison study, this means that there is no **bias.** Then, *on average*, the duplicate readings agree.
- The estimated standard deviation of the differences (s_d) provides a measure of agreement for an *individual*. However, it is more usual to calculate the **British Standards Institution repeatability coefficient** = $2s_d$. This is the maximum difference which is likely to occur between two measurements. Assuming a Normal distribution of differences, we expect approximately 95% of the differences in the population to lie between $\bar{d} \pm 2s_d$ where $\bar{d}$ is the mean of the observed differences. The upper and lower limits of this interval are called the **limits of agreement**; from them, we can decide (subjectively) whether the agreement between pairs of readings in a given situation is acceptable.
- An **index of reliability** commonly used to measure repeatability and reproducibility is the **intraclass correlation coefficient** (ICC, Chapter 42) which takes values from zero (no agreement) to 1 (perfect agreement). When measuring agreement between pairs of observations, the ICC is the proportion of the variability in the observations which is due to the differences between pairs, i.e. it is the between-pair variance expressed as a proportion of the total variance of the observations.

[1] Landis, J.R. and Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**; 159–174.

[2] Cohen, J. (1968). Weighted Kappa: nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bullelin* **70**: 213–220.

When there is no evidence of a systematic difference between the pairs, we may calculate the ICC as the Pearson correlation coefficient (Chapter 26) between the $2n$ pairs of observations obtained by including each pair twice, once when its values are as observed and once when they are interchanged (see Example 2).

If we wish to take the systematic difference between the observations in a pair into account, we estimate the ICC as:

$$\frac{s_a^2 - s_d^2}{s_a^2 + s_d^2 + \frac{2}{n}\left(n\bar{d}^2 - s_d^2\right)}$$

where we determine the difference between and the sum of the observations in each of the n pairs and:

s_a^2 is the estimated variance of the n sums;

s_d^2 is the estimated variance of the n differences;

$\bar{d}$ is the estimated mean of the differences (an estimate of the systematic difference).

We usually carry out a reliability study as part of a larger investigative study. The sample used for the reliability study should be a reflection of that used for the investigative study. We should not compare values of the ICC in different data sets as the ICC is influenced by features of the data, such as its variability (the ICC will be greater if the observations are more variable). Furthermore, the ICC is not related to the actual scale of measurement or to the size of error which is clinically acceptable.

Precautions

• It makes no sense to calculate a *single* measure of repeatability if the extent to which the observations in a pair disagree depends on the magnitude of the measurement. We can check this by determining both the mean of and the difference between each pair of readings, and plotting the n differences against their corresponding means[3] (Fig 39.1). If we observe a random scatter of points (evenly distributed above and below zero if there is no systematic difference between the pairs) then a single measure of repeatability is acceptable. If, however, we observe a funnel effect, with the variation in the differences being greater (say) for larger mean values, then we must reassess the problem. We may be able to find an appropriate transformation of the raw data (Chapter 9), so that when we repeat the process on the transformed observations, the required condition is satisfied. We can also use the plot to detect outliers (Chapter 3).

• Be wary of calculating the correlation coefficient (Chapter 26) between the two sets of readings (e.g. from the first and second occasions or from two methods/observers). We are not really interested in whether the points in the scatter diagram (e.g. of the results from the first occasion plotted against those from the second occasion) lie on a straight line; we want to know whether they conform to the line of equality (i.e., the 45° line when the two scales are the same). This will not be established by a hypothesis test of the null hypothesis that the true correlation coefficient is zero. It would, in any case, be very surprising if the pairs of measurements were not related, given the nature of the investigation. Furthermore, bear in mind the fact that it is possible to increase the magnitude of the correlation coefficient by increasing the range of values of the measurements.

More complex situations

Sometimes you may come across more complex problems when assessing agreement. For example, there may be more than two replicates, or more than two observers, or each of a number of observers may have replicate observations. You can find details of the analysis of such problems in Streiner and Norman[4].

[3] Bland, J.M., Altman, D.G. (1986). Statistical methods for assessing agreement between two pairs of clinical measurement. *Lancet* **i**; 307–310.

[4] Streiner, D.R. and Norman, G.L. (2003). *Health measurement scales: A practical guide to their development and use, 3rd edn*. Oxford University Press, Oxford.

Example 1

Assessing agreement—categorical variable

Two observers, an experienced dentist and a dental student, assessed the condition of 2104 tooth surfaces in school-aged children. Every surface was coded as '0' (sound), '1' (with at least one 'small' cavity), '2' (with at least one 'big' cavity) or '3' (with at least one filling, with or without cavities) by each individual. The observed frequencies are shown in Table 39.1. The bold figures along the diagonal show the observed frequencies of agreement; the corresponding expected frequencies are in brackets. We calculated **Cohen's Kappa** to assess the agreement between the two observers.

We estimate Cohen's kappa as:

$$\kappa = \frac{\left(\frac{1785+154+20+14}{2104}\right) - \left(\frac{1602.1+21.3+0.5+0.2}{2104}\right)}{1 - \left(\frac{1602.1+21.3+0.5+0.2}{2104}\right)}$$

$$= \frac{0.9377 - 0.7719}{1 - 0.7719} = 0.73$$

There appears to be substantial agreement between the student and the experienced dentist in the coding of the children's tooth surfaces.

Table 39.1 Observed (and expected) frequencies of tooth surface assessment.

		Dental student				
	Code	0	1	2	3	Total
Dentist	0	**1785** (1602.1)	46	0	7	1838
	1	46	**154** (21.3)	18	5	223
	2	0	0	**20** (0.5)	0	25
	3	3	1	0	**14** (0.2)	18
	Total	1834	201	43	26	2104

Data kindly provided by Dr R.D.Holt, Eastman Dental Institute, University College London, London, UK.

Example 2

Assessing agreement—numerical variables

The Rosenberg self-esteem index is used to judge a patient's evaluation of his or her own self-esteem. The maximum value of the index (an indication of high self-esteem) for a person is 50, comprising the sum of the individual values from ten questions, each scored from zero to five. Part of a study which examined the effectiveness of a particular type of surgery for facial deformity examined the change in a patient's psychological profile by comparing the values of the Rosenberg index in the patient before and after surgery. The investigators were concerned about the extent to which the Rosenberg score would be reliable for a set of patients, and decided to assess the repeatability of the measure on the first 25 patients requesting treatment for facial deformity. They obtained a value for the Rosenberg index when the patient initially presented at the clinic and then asked the patient for a second assessment 4 weeks later. The results are shown in Table 39.2.

The differences (first value – second value) can be shown to be approximately Normally distributed; they have a mean, $\bar{d} = 0.56$ and standard deviation, $s_d = 1.83$. The test statistic for the paired t-test is equal to 1.53 (degrees of freedom = 24), giving $P = 0.14$. This non-significant result indicates that there is no evidence of any systematic difference between the results on the two occasions.

The British Standards Institution repeatability coefficient is $2s_d = 2 \times 1.83 = 3.7$. Approximately 95% of the differences in the population of such patients would be expected to lie between $\bar{d} \pm 2\,s_d$, i.e. between –3.1 and 4.3. These limits are indicated in Fig. 39.1 which shows that the differences are randomly scattered around a mean of approximately zero.

The index of reliability is estimated as

$$\frac{287.573 - 3.340}{287.573 + 3.340 + \frac{2}{25}(25(0.56^2) - 3.340)} = 0.98$$

Since the systematic difference is negligible, this value for the ICC is the same as the one we get by calculating the Pearson correlation coefficient for the 50 pairs of results obtained by using each pair of results twice, once with the order reversed. As an illustration of the technique, consider the first 5 pairs of pre-treatment values: (30, 27), (39, 41), (50, 49), (45, 42) and (25, 28). If we reverse the order of each pair, we obtain a second set of 5 pairs: (27, 30), (41, 39), (49, 50), (42, 45) and (28, 25). By repeating this process for the remaining 20 pairs, we obtain a total of 50 pairs which we use to calculate the correlation coefficient, an estimate of the ICC.

Since the maximum likely difference between repeated measurements is around 3.7, and since virtually all (i.e., 98%) of the variability in the results can be attributed to differences between patients, the investigators felt that the Rosenberg index was reliable, and used it to evaluate the patients' perceptions of the effectiveness of the facial surgery.

Table 39.2 The pre-treatment values (1st and 2nd) of the Rosenberg index obtained on 25 patients.

1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
30	27	41	39	37	39	43	43	21	20
39	41	41	41	42	42	40	39	41	39
50	49	50	49	46	44	31	30	29	28
45	42	38	40	49	48	45	46	26	27
25	28	41	39	21	23	46	42	32	30

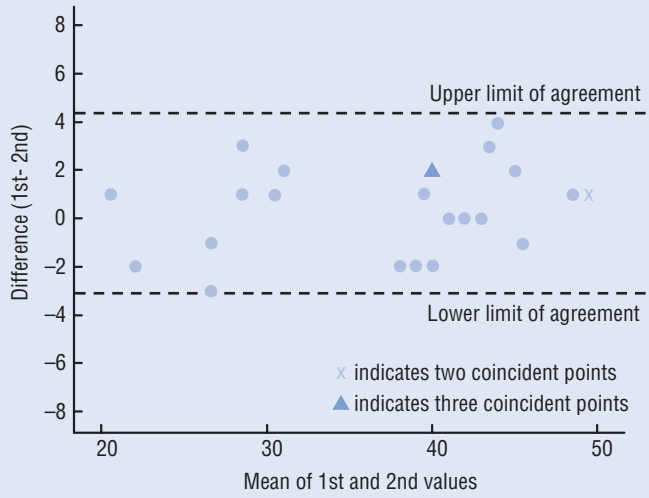


Figure 39.1 Difference between first and second Rosenberg self-esteem values plotted against their mean for 25 patients.

Cunningham, S.J., Hunt, N.P. and Feinnman, C. (1996) Perceptions of outcome following othognathic surgery *British Journal of Oral and Maxillofacial Surgery* **34**, 210–213, 1996

40 Evidence-based medicine

Sackett *et al.*[1] describe **evidence-based medicine (EBM)** as 'the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients'. To practice EBM, you must be able to locate the research relevant to the care of your patients, and judge its quality. Only then can you think about applying the findings in clinical practice.

Sackett *et al*. suggest the following approach to EBM. For convenience, we have phrased the third and fourth points below in terms of clinical trials (Chapter 14) and observational studies (Chapters 15 and 16), but they can be modified to suit other forms of investigations (e.g. diagnostic tests, Chapter 38).

1 Formulate the problem

You must decide what is of interest to you —how you define the patient population, which intervention (e.g. treatment) or comparison is relevant, and what outcome you are looking at (e.g. reduced mortality).

2 Locate the relevant information (e.g. on diagnosis, prognosis or therapy)

Often the relevant information will be found in published papers, but you should also consider other possibilities, such as conference abstracts. You must know what databases (e.g. Medline) and other sources of evidence are available, how they are organized, which search terms to use, and how to operate the searching software.

3 Critically appraise the methods in order to assess the validity (closeness to the truth) of the evidence

The following questions should be asked.

- Have all **important** *outcomes* been considered?
- Was the study conducted using an **appropriate spectrum of patients**?
- Do the results make **biological sense**?
- Was the study designed to eliminate **bias**? For example, in a clinical trial, was the study **controlled**, was **randomization** used in the assignment of patients, was the assessment of response '**blind**', were any patients lost to follow-up, were the groups treated in similar fashion, aside from the fact that they received different treatments, and was an '**intention-to-treat**' analysis performed?
- Are the statistical methods appropriate (e.g. have underlying assumptions been verified; have dependencies in the data (e.g. pairing) been taken into account in the analysis)?

[1] Sackett, D.L., Straus, S., Richardson, S., Rosenberg. W. and Haynes, R.B. (2000) *Evidence-based Medicine: How to Practice and Teach EBM.*2nd Edn Churchill-Livingstone, London.

4 Extract the most useful results and determine whether they are important

Extracting the most useful results

You should ask the following questions:

(a) What is the **main outcome variable** (i.e. that which relates to the major objective)?

(b) How **large** is the **effect of interest**, expressed in terms of the main outcome variable? If this variable is:

- **Binary** (e.g. died/survived)
  - (i) What are the rates/risks/odds of occurrence of this event (e.g. death) in the (two) comparison groups?
  - (ii) The effect of interest may be the difference in rates or risks (the absolute reduction) or a ratio (the relative rate or risk or odds ratio) —what is its magnitude?
- **Numerical** (e.g. systolic blood pressure)
  - (i) What is the mean (or median) value of the variable in each of the comparison groups?
  - (ii) What is the effect of interest, i.e. the difference in means (medians)?

(c) How **precise** is the **effect of interest**? Ideally, the research being scrutinized should include the confidence interval for the true effect (a wide confidence interval is an indication of poor precision). Is this confidence interval quoted? If not, is sufficient information (e.g. the standard error of the effect of interest) provided so that the confidence interval can be determined?

Deciding whether the results are important

- Consider the **confidence interval** for the effect of interest (e.g. the difference in treatment means):
  - (i) Would you regard the observed effect clinically important (irrespective of whether or not the result of the relevant hypothesis test is statistically significant) if the lower limit of the confidence interval represented the true value of the effect?
  - (ii) Would you regard the observed effect clinically important if the upper limit of the confidence interval represented the true value of the effect?
  - (iii) Are your answers to the above two points sufficiently similar to declare the results of the study unambiguous and important?
- To assess therapy in a randomized controlled trial, evaluate the **number of patients you need to treat** (NNT) with the experimental treatment rather than the control treatment in order to prevent one of them developing the 'bad' outcome (such as post-partum haemorrhage, see Example). The NNT can be determined in various ways depending on the information available. It is, for example, the reciprocal of the difference in the proportions of individuals with the bad outcome in the control and experimental groups (see Example).

5 Apply the results in clinical practice

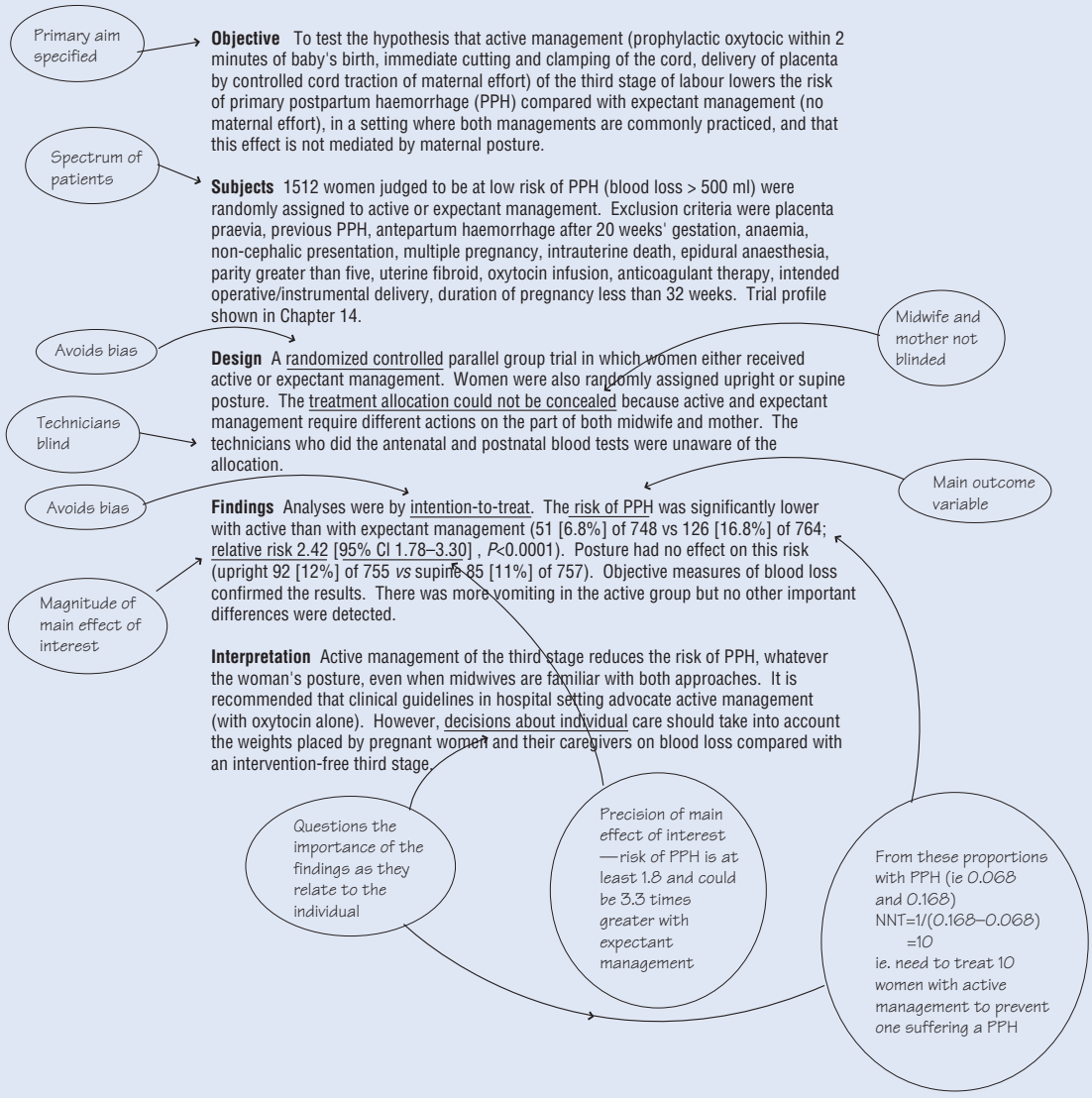
If the results are to help you in caring for your patients, you must ensure that:

- your patient is similar to those on whom the results were obtained;
- the results can be applied to your patient;
- all clinically important outcomes have been considered;
- the likely benefits are worth the potential harms and costs.

6 Evaluate your performance

Self-evaluation involves questioning your abilities to complete tasks 1 to 5 successfully. Are you then able to integrate the critical appraisal into clinical practice, and have you audited your performance? You should also ask yourself whether you have learnt from past experience so that you are now more efficient and are finding the whole process of EBM easier.

Example



Adapted from Rogers, J., Wood, J., McCandish, R., Ayers, S., Truesdale, A., Elbourne, D. (1998) Active versus expectant management of third stage of labour: the Hinchingbrooke randomised controlled trial. *Lancet*, **351**, 693–699, with permission from Elsevier.

41 Methods for clustered data

Clustered data conform to a hierarchical or nested structure in which, in its simplest form (the univariable **two-level structure**), the value of a single response variable is measured on a number of *level 1 units* contained in different groups or clusters (*level 2 units*). For example, the level 1 and level 2 units, respectively, may be teeth in a mouth, knees in a patient, patients in a hospital, clinics in a region, children in a class, successive visit times for a patient (i.e. *longitudinal* data, Fig. 41.1), etc. The statistical analysis of such **repeated measures** data should take into account the fact that the observations in a cluster tend to be correlated, i.e. they are *not independent*. Failure to acknowledge this usually results in underestimation of the standard errors of the estimates of interest and, consequently, increased Type I error rates and confidence intervals which are too narrow.

For the purposes of illustration, we shall assume, in this chapter, that we have longitudinal data and our repeated measures data comprise each patient's values of the variable at different time points, i.e. the patient is the cluster. We summarize the data by describing the patterns in individual patients, and, if relevant, assess whether these patterns differ between two or more groups of patients.

Displaying the data

A plot of the measurement against time for each patient in the study provides a visual impression of the pattern over time. When we are studying only a small group of patients, it may be possible to show all the individual plots in one diagram. However, when we are studying large groups this becomes difficult, and we may illustrate just a selection of 'representative' individual plots (Fig. 41.3), perhaps in a grid for each treatment group. Note that the average pattern generated by plotting the means over all patients at each time point may be very different from the patterns seen in individual patients.

Comparing groups: inappropriate analyses

It is **inappropriate** to use all the values in a group to fit a *single* linear regression line (Chapters 27, 28) or perform a one-way analysis of variance (ANOVA; Chapter 22) to compare groups because these methods do not take account of the repeated measurements on the same patient. Furthermore, it is also **incorrect** to compare the means in the groups *at each time point separately* using unpaired *t*-tests (Chapter 21) or one-way ANOVA for a number of reasons:

- The measurements in a patient from one time point to the next are not independent, so interpretation of the results is difficult. For example, if a comparison is significant at one time point, then it is likely to be significant at other time points, irrespective of any changes in the values in the interim period.
- The large number of tests carried out implies that we are likely to obtain significant results purely by chance (Chapter 18).
- We lose information about within-patient changes.

Comparing groups: appropriate analyses

Using summary measures

We can base our analysis on a **summary measure** that captures the important aspects of the data, and calculate this summary measure *for each patient*. Typical summary measures are:

- change from baseline at a pre-determined time point;
- maximum (peak) or minimum (nadir) value reached;
- time to reach the maximum (or minimum) value;
- time to reach some other pre-specified value;
- average value (e.g. mean);
- area under the curve (AUC, Fig. 41.2);
- slope or intercept of the patient's regression line (describing the relationship between the measurement and time).

If the parameter (e.g. the mean or slope) is estimated more precisely in some patients than others (perhaps because there are more observations for these patients), we should take account of this in the analysis by giving more weight to those measures which are estimated more precisely.

The choice of summary measure depends on the main question of interest and should be made in advance of collecting the data. For example, if we are considering drug concentrations following treatment with two therapies, we may consider time to maximum drug concentration (C_{max}) or AUC. However, if we are interested in

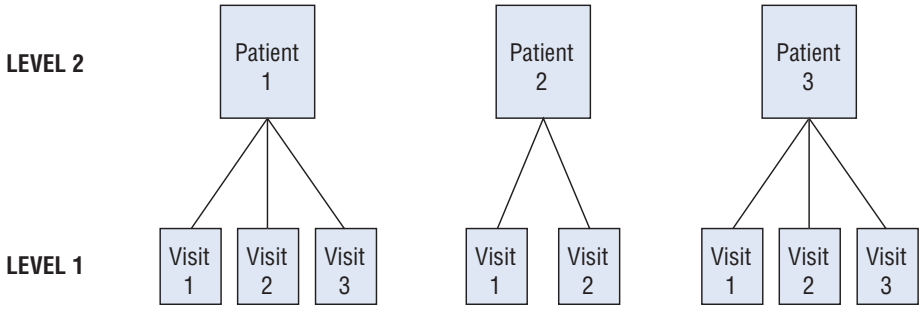


Figure 41.1 Diagrammatic representation of a two-level hierarchical structure for longitudinal data.

antibody titres following vaccination, then we may be interested in the time it takes the antibody titre to drop below a particular protective level.

We compare the values of the summary measure in the different groups using standard hypothesis tests [e.g. Wilcoxon rank sum (Chapter 21) or Kruskal–Wallis (Chapter 22)]. Because we have reduced a number of dependent measurements on each individual to a single quantity, the values included in the analysis are now independent.

Whilst analyses based on summary measures are simple to perform, it may be difficult to find a suitable measure that adequately describes the data, and we may need to use two or more summary measures. In addition, these approaches do not use all data values fully.

Repeated measures ANOVA

We can perform a particular type of ANOVA (Chapter 22), called **repeated measures ANOVA**, in which the different time points are considered as the levels of one factor in the analysis and the grouping variable is a second factor in the analysis. We can regard the repeated measures ANOVA as an extension of the paired *t*-test when we have more than two related observations. If the repeated measures ANOVA produces significant differences between the groups, then paired *t*-tests, which take account of the dependence in the data and have *P*-values adjusted for multiple testing (Chapter 18), can be performed to identify at what time points these differences become apparent[1].

However, repeated measures ANOVA has several disadvantages:

- It is often difficult to perform.
- The results may be difficult to interpret.
- It generally assumes that values are measured at regular time intervals and that there are no missing data, i.e. the design of the study is assumed to be **balanced**. In reality, values are rarely measured at all time points because patients often miss appointments or come at different times to those planned.

Regression methods

Various regression methods, such as those which provide parameter estimates with robust standard errors or use generalized estimating equations (GEE) or random effects models, may be used to analyse clustered data (see Chapter 42).

[1] Mickey, R.M., Dunn, O.J., Clark, V.A. (2004) *Applied Statistics: Analysis of Variance and Regression* 3rd Edn. Wiley.

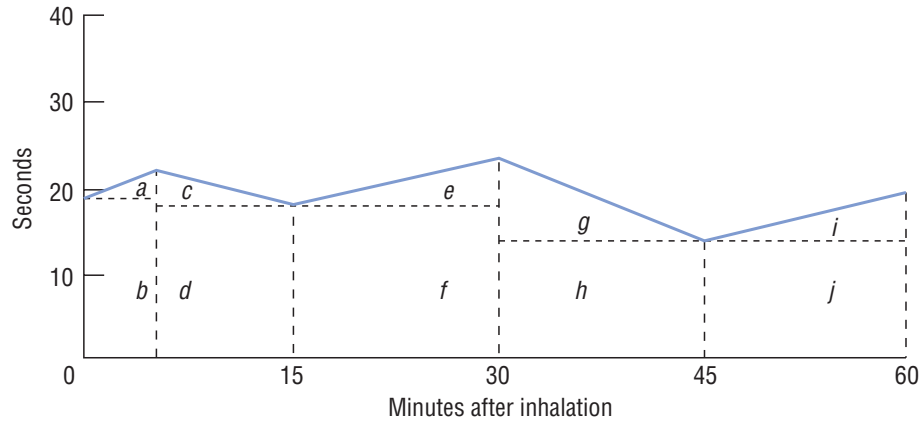


Figure 41.2 Calculation of the AUC for a single student. The total area under the line can be divided into a number of rectangles and triangles (marked *a* to *j*). The area of each can easily be calculated. Total AUC = Area (*a*) + Area (*b*) + . . . + Area (*j*).

Example

As part of a practical class designed to assess the effects of two inhaled bronchodilator drugs, fenoterol hydrobromide and ipratropium bromide, 99 medical students were randomized to receive one of these drugs ($n = 33$ for each drug) or placebo ($n = 33$). Each student inhaled four times in quick succession. Tremor was assessed by measuring the total time (in seconds) taken to thread five sewing needles mounted on a cork; measurements were made at baseline before inhalation and at 5, 15, 30, 45 and 60 mins afterwards. The measurements of a representative sample of the students in each treatment group are shown in Fig. 41.2.

It was decided to compare the values in the three groups using the 'area under the curve' (AUC) as a summary measure. The calculation of AUC for one student is illustrated in Fig. 41.3.

The median (range) AUC was 1552.5 (417.5–3875), 1215 (457.5–2500) and 1130 (547.5–2625) seconds[2] in those receiving fenoterol hydrobromide, ipratropium bromide and placebo, respectively. The values in the three groups were compared using the Kruskal–Wallis test which gave $P = 0.008$. There was thus strong evidence that the AUC measures were different in the three groups. Non-parametric *post-hoc* comparisons, adjusted for multiple testing, indicated that values were significantly greater in the group receiving fenoterol hydrobromide, confirming pharmacological knowledge that this drug, as a β_2-adrenoceptor agonist, induces tremor by the stimulation of β_2-adrenoceptors in skeletal muscle.

Data were kindly provided by Dr R. Morris, Department of Primary Care and Population Sciences, and were collected as part of a student practical class organized by Dr T.J. Allen, Department of Pharmacology, Royal Free and University College Medical School, London, UK.

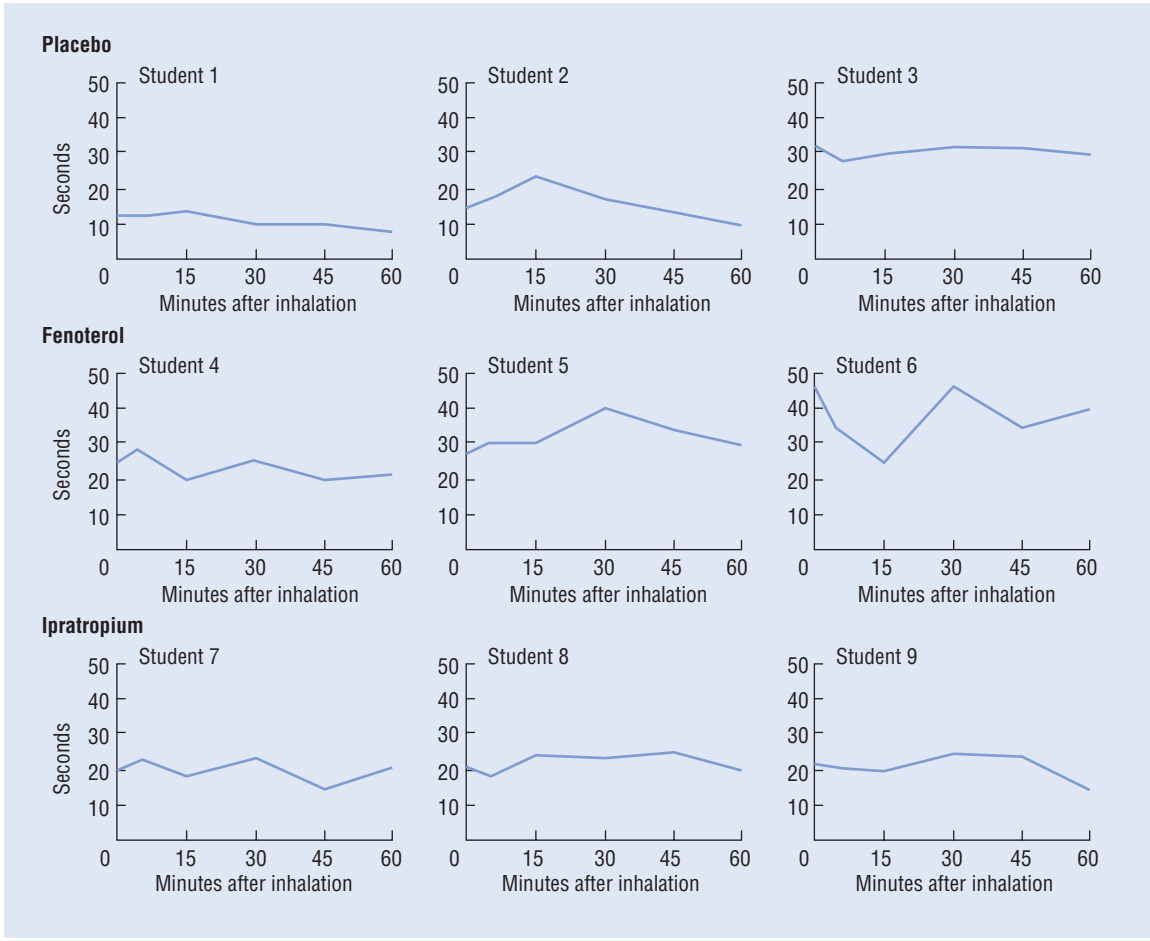


Figure 41.3 Time taken to thread five sewing needles for three representative students in each treatment group.

42 Regression methods for clustered data

Various **regression methods** can be used for the analysis of the two-level hierarchical structure described in Chapter 41, in which each cluster (level 2 unit) contains a number of individual level 1 units. For example, in a study of rheumatoid arthritis, we may measure the flexion angle on both the left and right knees (level 1) of every patient (level 2). Alternatively, we may have a longitudinal data set with a measurement (e.g. total cholesterol) observed at successive times (level 1) on each patient (level 2). The main advantages and disadvantages of each method are summarized in Table 42.1. Most of these methods are unreliable unless there are sufficient clusters, and they can be complicated to perform and interpret correctly; we therefore suggest you consult a specialist statistician for advice.

Aggregate level analysis

A very simple approach is to *aggregate* the data and perform an analysis using an appropriate numerical **summary measure** (e.g. the mean) for each cluster (e.g. the patient) (Chapter 41). The choice of this summary measure will depend on features of the data and on the hypotheses being studied. We perform an ordinary least squares (OLS) multiple regression analysis using the *cluster as the unit of investigation* and the *summary measure as the outcome variable*. If each cluster has been allocated a particular treatment (in the knee example, the patient may be randomly allocated one of two treatments—an exercise regimen or no exercise) then, together with other cluster level covariates (e.g. gender, age), we can incorporate 'treatment' in the regression model as a dummy variable using codes such as 0 and 1 (or as a series of dummy variables if we have more than 2 treatments (Chapter 29)).

Robust standard errors

If the clustering is ignored in the regression analysis of a two-level structure, an important assumption underlying the linear regression model—that of independence between the observations (see Chapters 27 and 28)—is violated. As a consequence, the standard errors of the parameter estimates are likely to be too small and, hence, results may be spuriously significant.

To overcome this problem, we may determine **robust standard errors** of the parameter estimates, basing our calculation of them on the variability in the data (evaluated by appropriate residuals) rather than on that assumed by the regression model. In a multiple regression analysis with robust standard errors, the estimates of the regression coefficients are the same as in OLS linear regression but the standard errors are more robust to violations of the underlying assumptions, our particular concern being lack of independence when we have clustered data.

Random effects models

Random effects models[1] are also known as **hierarchical**, **multilevel**, **mixed**, **cluster-specific** or **cross-sectional** time series models. They can be fitted using various comprehensive statistical computer packages, such as SAS and Stata, or specialist software such as MLwiN (http://multilevel.ioe.ac.uk), all of which use a version of maximum likelihood estimation. The estimate of the effect for each cluster is derived using both the individual cluster information as well as that of the other clusters so that it benefits from the 'shared' information. In particular, *shrinkage* estimates are commonly determined whereby, using an appropriate shrinkage factor, each cluster's estimate of the effect of interest is 'shrunk' towards the estimated overall mean. The amount of shrinkage depends on the cluster size (smaller clusters have greater shrinkage) and on the variation in the data (shrinkage is greater for the estimates when the variation within clusters is large when compared to that between clusters).

A random effects model regards the clusters as a sample from a real or hypothetical population of clusters. The individual clusters are not of primary interest; they are assumed to be broadly similar with differences between them attributed to random variation or to other 'fixed' factors such as gender, age, etc. The two-level random effects model differs from the model which takes no account of clustering in that, although both incorporate random or unexplained error due the variation between level 1 units (the *within* cluster variance, σ^2), the random effects model also includes random error which is due to the variation *between* clusters (σ_c^2). The variance of an individual observation in this random effects model is therefore the sum of the two components of variance, i.e. it is $\sigma^2 + \sigma_c^2$.

Particular models

When the outcome variable, y, is numerical and there is a single explanatory variable, x, of interest, the simple **random intercepts** linear two-level model assumes that there is a linear relationship between y and x in each cluster, with all the cluster regression lines having a common slope, β, but different intercepts (Fig. 42.1a). The mean regression line has a slope equal to β and an intercept equal to α, which is the mean intercept averaged over all the clusters. The random error (residual) for each cluster is the amount by which the intercept for that cluster regression line differs, in the vertical direction, from the overall mean intercept, α (Fig. 42.1a). The cluster residuals are assumed to follow a Normal distribution with zero mean and variance = σ_c^2. Within each cluster, the residuals for the level 1 units are assumed to follow a Normal distribution with zero mean and the same variance, σ^2. If the cluster sizes are similar, a simple approach to checking for Normality and constant variance of the residuals for both the level 1 units and clusters is to look for Normality in a histogram of the residuals, and to plot the residuals against the predicted values (see Chapter 28).

This model can be *modified* in a number of ways (see also Table 42.1), for example, by allowing the slope, β, to vary randomly between clusters. The model is then called a **random slopes model** in which case the cluster specific regression lines are not parallel to the mean regression line (Fig 42.1b).

[1] Goldstein, H. (2003) *Multilevel Statistical Models* 3rd edn, Kendall Library of Statistics 3, Arnold.

Assessing the clustering effect

The effect of clustering can be assessed by:

• Calculating the **intraclass correlation coefficient** (ICC**,** sometimes denoted by ρ—see also Chapter 39) which, in the two-level structure, represents the correlation between two randomly chosen level 1 units in one randomly chosen cluster.

$$\text{ICC} = \frac{\sigma_c^2}{\sigma^2 + \sigma_c^2}$$

The ICC expresses the variation between the clusters as a proportion of the total variation; it is often presented as a percentage. The ICC = 1 when there is no variation *within* the clusters and all the variation is attributed to differences *between* clusters; the ICC = 0 when there is no variation between the clusters. We can use the ICC to make a subjective decision about the importance of clustering.

• Comparing two models where one model is the full random effects model and the other is a regression model with the same explanatory variable(s) but which does not take clustering into account. The relevant *likelihood ratio test* has a test statistic equal to the *difference* in the likelihood ratio statistics of the two models (see Chapter 32) and it follows the Chi-squared distribution with one degree of freedom.

[2] Liang, K.-Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models *Biometrika* **73**; 13–22.

Generalized estimating equations (GEE)

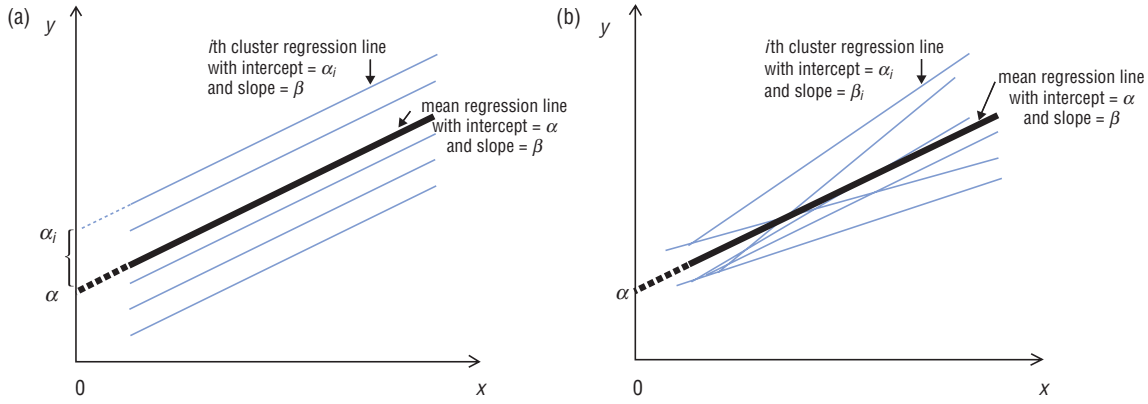
In the GEE approach[2] to estimation, we adjust both the parameter estimates of a GLM and their standard errors to take into account the clustering of the data in a two-level structure. We make distributional assumptions about the dependent variable but, in contrast to the random effects model, do not assume that the between-cluster residuals are Normally distributed. We regard the clustering as a nuisance rather than of intrinsic interest, and proceed by postulating a 'working' structure for the correlation between the observations within each cluster. This does not have to be correct since, provided there are enough clusters, the robust standard errors and parameter estimates will be acceptable. However, we will obtain better parameter estimates if the structure is plausible. We commonly adopt an *exchangeable* correlation structure which assumes that exchanging two level 1 units within a cluster will not affect the estimation.

The GEE approach is sometimes called **population-averaged** (referring to the population of clusters) or **marginal** because the parameter estimates represent the effects averaged across the clusters (even though all level 1 unit information is included in the analysis). The GEE approach is often preferred to the more complex random effects model analysis for logistic (Chapter 30) and, sometimes, Poisson (Chapter 31) regression, even though the exchangeable correlation structure is known to be incorrect in these situations.

Table 42.1 Main advantages and disadvantages of regression methods for analysing clustered data.

Method	Advantages	Disadvantages
Aggregate level analysis	• Simple • Easy to perform with basic software	• Does not allow for effects of covariates for level 1 units • Ignores differences in cluster sizes and in precision of the estimate of each cluster summary measure • May not be able to find an appropriate summary measure
Robust standard errors that allow for clustering	• Relatively simple • Can include covariates which vary for level 1 units • Adjusts standard errors, confidence intervals and *P*-values to take account of clustering • Allows for different numbers of level 1 units per cluster	• Unreliable unless number of clusters large, say >30 • Does not adjust parameter estimates for clustering
Random effects model	• Explicitly allows for clustering by including both inter- and intra-cluster variation in model • Cluster estimates benefit from shared information from all clusters • Adjusts parameter estimates, standard errors, confidence intervals and *P*-values to take account of clustering • Can include covariates which vary for level 1 units • Allows for different numbers of level 1 units per cluster • Can extend hierarchy from 2 levels to multi-levels • Can accommodate various forms of GLM (e.g. Poisson)	• Unreliable unless there are sufficient clusters • Parameter estimates often biased • Complex modelling skills required for extended models • Estimation of random effects logistic model problematic
GEE	• Relatively simple • No distributional assumptions of random effects (due to clusters) required • Can include covariates which vary for level 1 units • Allows for different numbers of level 1 units per cluster • Adjusts parameter estimates, standard errors, confidence intervals and *P*-values to take account of clustering	• Unreliable unless number of clusters large, say >30 • Treats clustering as a nuisance of no intrinsic interest* • Requires specification of working correlation structure* • Parameter estimates are cluster averages and do not relate to individuals in population*

*These points may sometimes be regarded as advantages, depending on the question of interest.



(a) **Random intercepts model** The bold line represents the mean regression line for all the clusters and each of the lighter lines represents the regression line for a different cluster. The intercept of the ith cluster specific regression line differs from that of the mean line by a residual = $\alpha_i - \alpha$, where these residuals are Normally distributed with zero mean and variance, σ_c^2. Every line has a slope = β.

(b) **Random slopes model** The bold line represents the mean regression line for all the clusters and each of the lighter lines represents the regression line for a different cluster. The intercept of the ith cluster specific regression line differs from that of the mean line by a residual = $\alpha_i - \alpha$, and the slope of the ith specific regression line differs from that of the mean line by a residual = $\beta_i - \beta$, where the residuals are Normally distributed with zero mean and variances σ_c^2 and σ_d^2, respectively.

Figure 42.1 Two-level random effects linear regression models with a single covariate, x.

Example

Data relating to periodontal disease were obtained on 96 white male trainee engineers aged between 16 and 20 entering the apprentice training school at Royal Air Force Halton, England (see also Chapter 20). Each of the possible 28 teeth (excluding wisdom teeth) in every trainee's mouth was examined at four sites (the mesiobuccal, mesiolingual, distobuccal and mesiobuccal). To simplify the analysis, we have considered a subset of the data, namely, only (1) the mesiobuccal site in each tooth; this leads to a two-level structure of teeth within subjects (each subject corresponds to a cluster), and (2) two variables of interest; loss of attachment (loa, measured in mm) between the tooth and the jawbone evaluated at the mesiobuccal site, and the current cigarette smoking status of the trainee (yes = 1, no = 0). We wish to assess whether smoking is a risk factor for gum disease (where greater loss of attachment indicates worse disease).

Table 42.2 shows extracts of the results from different regression analyses in which the outcome variable is loss of attachment (mm) and the covariate is smoking. Full computer output is given in Appendix C. The estimates of the regression coefficients for smoking and/or their standard errors vary according to the type of analysis performed. The two OLS analyses have identical estimated regression coefficients (which are larger than those of the other three analyses) but their standard errors are different. The standard error of the estimated regression coefficient in the OLS analysis which ignores clustering is substantially smaller that the standard errors in the other four analyses, i.e. *ignoring clustering results in an underestimation of the standard error of the regression coefficient and, consequently, a confidence interval that is too narrow and a P-value that is too small.* The intracluster correlation coefficient from the random effects model is estimated as 0.224. Thus approximately 22% of the variation in loss of attachment, after taking account of smoking, was between trainees rather than within trainees.

In this particular example, we conclude from all five analyses that smoking is not significantly associated with loss of attachment. This lack of significance for smoking is an unexpected finding and may be explained by the fact that these trainees were very young and so the smokers amongst them would not have smoked for a long period.

Table 42.2 Summary of results of regression analyses in which loa (mm) is the outcome variable.

Analysis	Estimated coefficient (smoking)	Standard Error (SE)	95% CI for coefficient	Test statistic*	*P*-value
OLS+ regression ignoring clustering	–0.0105	0.0235	–0.057 to 0.036	$t = -0.45$	0.655
OLS regression with robust SEs	–0.0105	0.0526	–0.115 to 0.094	$t = -0.20$	0.842
Aggregate analysis (OLS regression on group means)	–0.0046	0.0612	–0.126 to 0.117	$t = -0.07$	0.941
Random effects model	–0.0053	0.0607	–0.124 to 0.114	$z = -0.09$	0.930
GEE with robust SEs & exchangeable correlation structure	–0.0053	0.0527	–0.108 to 0.098	$z = -0.10$	0.920

* t = test statistic following t-distribution; z = Wald test statistic following standard Normal distribution.
+ OLS = ordinary least squares.

Data kindly provided by Dr Gareth Griffiths, Dept of Periodontology, Eastman Dental Institute, University College London, UK.

43 Systematic reviews and meta-analysis

The systematic review

What is it?

A **systematic review**[1] is a formalized and stringent process of combining the information from *all* relevant studies (both published and unpublished) of the same health condition; these studies are usually clinical trials (Chapter 14) of the same or similar treatments but may be observational studies (Chapters 15 and 16). A systematic review is an integral part of **evidence-based medicine** (EBM; Chapter 40) which applies the results of the best available evidence, together with clinical expertise, to the care of patients. So important is its role in EBM that it has become the focus of an international network of clinicians, methodologists and consumers who have formed the **Cochrane Collaboration**. This has produced the Cochrane Library containing regularly updated evidence-based healthcare databases including the Cochrane Database of Systematic Reviews; full access to these reviews requires subscription but the abstracts are freely available on the internet (www.cochrane.org/reviews).

What does it achieve?

- **Refinement and reduction**—large quantities of information are refined and reduced to a manageable size.
- **Efficiency**—the systematic review is usually quicker and less costly to perform than a new study. It may prevent others embarking on unnecessary studies, and can shorten the time lag between medical developments and their implementation.
- **Generalizability and consistency**—results can often be generalized to a wider patient population in a broader setting than would be possible from a single study. Consistencies in the results from different studies can be assessed, and any inconsistencies determined.
- **Reliability**—the systematic review aims to reduce errors, and so tends to improve the reliability and accuracy of recommendations when compared with haphazard reviews or single studies.
- **Power and precision**—the quantitative systematic review (see meta-analysis) has greater power (Chapter 18) to detect effects of interest and provides more precise estimates of them than a single study.

Meta-analysis

What is it?

A **meta-analysis** or **overview** is a particular type of systematic review that focuses on the numerical results. The main aim of a meta-analysis is to combine the results from individual studies to produce, if appropriate, an estimate of the overall or average effect of interest (e.g. the relative risk, RR; Chapter 15). The direction and magnitude of this average effect, together with a consideration of the associated confidence interval and hypothesis test result, can be used to make decisions about the therapy under investigation and the management of patients.

Statistical approach

1. We decide on the **effect of interest** and, if the raw data are available, evaluate it for each study. However, in practice, we may have to extract these effects from published results. If the outcome in a clinical trial comparing two treatments is:

- **numerical**—the effect may be the difference in treatment means. A zero difference implies no treatment effect;
- **binary** (e.g. died/survived)—we consider the risks, say, of the outcome (e.g. death) in the treatment groups. The effect may be the difference in risks or their ratio, the RR. If the difference in risks equals zero or RR = 1, then there is no treatment effect.

2. Check for statistical homogeneity and obtain an estimate of statistical heterogeneity—we have **statistical heterogeneity** when there is genuine variation between the effects of interest from the different studies. We can perform a hypothesis **test of homogeneity** to investigate whether the variation in the individual effects is compatible with chance alone. However, this test has low power (Chapter 18) to detect heterogeneity if there are few studies in the meta-analysis and may, conversely, give a highly significant result if it comprises many large studies, even when the heterogeneity is unlikely to affect the conclusions. An **index**, I^2, which does not depend on the number of studies, the type of outcome data or the choice of treatment effect (e.g. relative risk), can be used to quantify the impact of heterogeneity and assess inconsistency[2] (see Example). I^2 represents the percentage of the total variation across studies due to heterogeneity; it takes values from 0% to 100%, with a value of 0% indicating no observed heterogeneity. If there is evidence of statistical heterogeneity, we should proceed cautiously, investigate the reasons for its presence and modify our approach accordingly, perhaps by dividing the studies into subgroups of those with similar characteristics.

3. Estimate the average effect of interest (with a confidence interval), and perform the appropriate hypothesis test on the effect (e.g. that the true RR = 1)—you may come across the terms 'fixed effects' and 'random effects' models in this context (see also Chapter 42). If there is no evidence of statistical heterogeneity, we generally use a fixed effects model (which assumes the treatment effect is the same in every study and any observed variation in it is due to sampling error); otherwise we use a random effects model (which assumes that the separate studies represent a random sample from a population of studies which has a mean treatment effect about which the individual study effects vary).

4. Interpret the results and present the findings—it is helpful to summarize the results from each trial (e.g. the sample size, baseline characteristics, effect of interest such as the RR, and related confidence interval, CI) in a table (see Example). The most common graphical display is a **forest plot** (Fig. 43.1) in which the estimated effect (with CI) for each trial and their average are marked along the length of a vertical line which represents 'no treatment effect' (e.g. this line corresponds to the value 'one' if the effect is a RR). The plotting symbol for the estimated effect for each study is often a box which has an area proportional to the size of that study. Initially, we examine whether the estimated effects from the different studies are on the same side of the line. Then we can use the CIs to judge whether the results are compatible (if the CIs overlap), to

[1] Chalmers, I. and Altman, D.G. (eds) (1995) *Systematic Reviews*. British Medical Journal Publishing Group, London.

[2] Higgins, P.T., Thompson, S.G., Deeks, J.J. and Altman, D.G. (2003) Measuring inconsistency in meta-analysis *British Medical Journal* **237**, 557–560.

determine whether incompatible results can be explained by small sample sizes (if CIs are wide) and to assess the significance of the individual and overall effects (by observing whether the vertical line crosses some or all of the CIs).

Advantages and disadvantages

As a meta-analysis is a particular form of systematic review, it offers all the **advantages** of the latter (see 'what does it achieve?'). In particular, a meta-analysis, because of its inflated sample size, is able to detect treatment effects with **greater power** and estimate these effects with **greater precision** than any single study. Its advantages, together with the introduction of meta-analysis software, have led meta-analyses to proliferate. However, improper use can lead to erroneous conclusions regarding treatment efficacy. The following principal **problems** should be thoroughly investigated and resolved before a meta-analysis is performed.

- **Publication bias**—the tendency to include in the analysis only the results from published papers; these favour statistically significant findings. We can decide if publication bias is an issue by drawing a **funnel plot**, a scatter diagram which usually has the study sample size on the horizontal axis and the treatment effect (e.g. odds ratio) on the vertical axis, In the absence of publication bias, the scatter of points (each point representing one study) in the funnel plot will be substantial at the bottom where the study size is small, and will narrow (in the shape of a funnel) towards the top where the study size is large. If publication bias is present, the funnel plot will probably be skewed and asymmetrical, with a gap towards the bottom left hand corner where both the treatment effect and study size are small (i.e. when the study has low power to detect a small effect).
- **Clinical heterogeneity**—in which differences in the patient population, outcome measures, definition of variables, and/or duration of follow-up of the studies included in the analysis create problems of non-compatibility.
- **Quality differences**—the design and conduct of the studies may vary in their quality. Although giving more weight to the better studies is one solution to this dilemma, any weighting system can be criticized on the grounds that it is arbitrary.
- **Dependence**—the results from studies included in the analysis may not be independent, e.g. when results from a study are published on more than one occasion.

Example

A patient with severe angina will often be eligible for either percutaneous transluminal coronary angioplasty (PTCA) or coronary artery bypass graft (CABG) surgery. Results from eight published randomized trials were combined in a collaborative **meta-analysis** of 3371 patients (1661 CABG, 1710 PTCA) with a mean follow-up of 2.7 years. The main features of the trials are shown in Table 43.1. Results for the composite endpoint of cardiac death plus non-fatal myocardial infarction (MI) in the first year of follow-up are shown in Fig. 43.1. The estimated relative risks (RR) are for the PTCA group compared with the CABG group. The figure uses a logarithmic scale for the RR to achieve symmetrical confidence intervals (CI). Although the individual estimates of relative risk vary quite considerably, from reductions in risk to quite large increases in risk, all the confidence intervals overlap to some extent. A more formal assessment of heterogeneity is provided by **Cochran's Chi-squared test for homogeneity** which gives a non-significant result (test statistic $Q = 13.2$, degrees of freedom $df = 8 - 1 = 7$, $P = 0.07$). However, $I^2 = 100\times(Q - df)/Q = 100\times(13.2 - 7)/13.2 = 47\%$ (95% CI 0% to 76%) which suggests moderate inconsistency across the studies and advocates a cautious approach to interpreting the combined estimate of relative risk for all trials. This relative risk was estimated as 1.03 (95% CI 0.79 to 1.50), indicating that there was no evidence of a real overall difference between the two revascularization strategies. It may be of interest to note that during early follow-up, the prevalence of angina was higher in PTCA patients than in CABG patients.

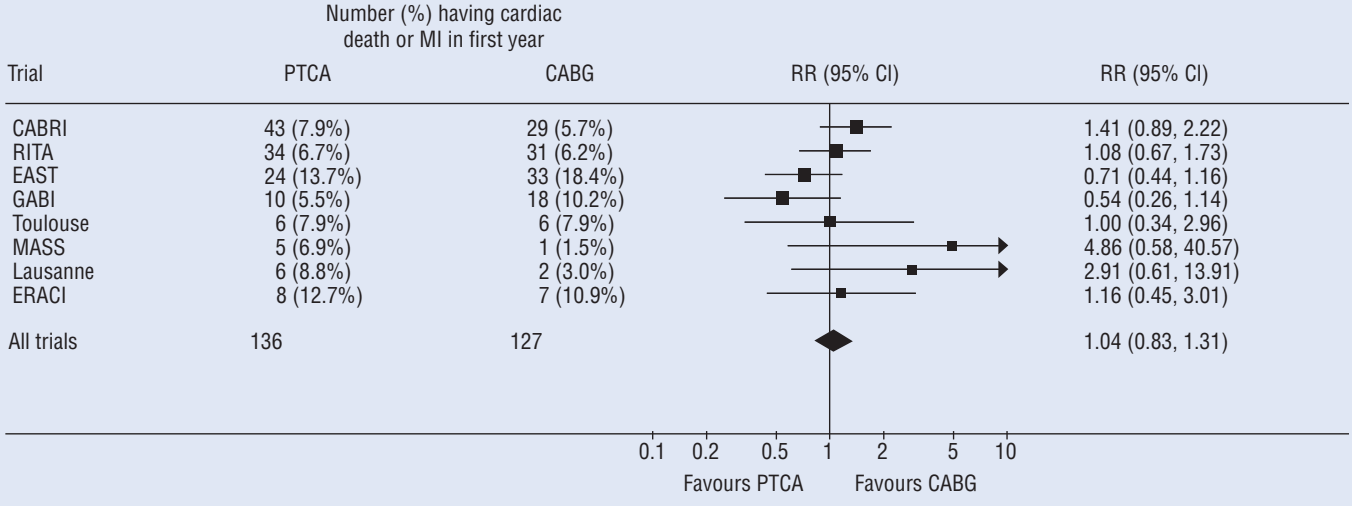


Figure 43.1 Forest plot of relative risk (RR) with 95% CI of cardiac death or myocardial infarction for PTCA group compared with CABG group in first year since randomization.

continued

Table 43.1 Characteristics of eight randomized trials comparing percutaneous transluminal coronary angioplasty and coronary artery bypass graft.

	Country	Principal investigator	Single- or multi-vessel	Number of patients		Follow-up (years)
				CABG	PTCA	
Coronary Angioplasty Bypass Revascularisation Investigation (CABRI)	Europe	A.F. Rickards	Multi	513	541	1
Randomised Intervention on Treatment of Angina Trial (RITA)	UK	J.R. Hampton	Single ($n = 456$) Multi ($n = 555$)	501	510	4.7
Emory Angioplasty versus Surgery Trial (EAST)	USA	S.B. King	Multi	194	198	3+
German Angioplasty Bypass Surgery Investigation (GABI)	Germany	C.W. Hamm	Multi	177	182	1
The Toulouse Trial (Toulouse)	France	J. Puel	Multi	76	76	2.8
Medicine Angioplasty or Surgery study (MASS)	Brazil	W. Hueb	Single	70	72	3.2
The Lausanne trial (Lausanne)	Switzerland	J.-J. Goy	Single	66	68	3.2
Argentine Trial of PTCA versus CABG (ERACI)	Argentina	A. Rodriguez	Multi	64	63	3.8

Adapted from Pocock, S.J., Henderson, R.A., Rickards, A.F., *et al.* (1995) A meta-analysis of randomised trials comparing coronary angioplasty with bypass surgery. *Lancet*, **346**, 1184–1189, with permission from Elsevier.

44 Survival analysis

Survival data are concerned with the time it takes an individual to reach an endpoint of interest (often, but not always, death) and are characterized by the following two features.

- It is the **length of time** for the patient to reach the endpoint, rather than whether or not s/he reaches the endpoint, that is of primary importance. For example, we may be interested in length of survival in patients admitted with cirrhosis.
- Data may often be **censored** (see below).

Standard methods of analysis, such as logistic regression or a comparison of the mean time to reach the endpoint in patients with and without a new treatment, can give misleading results because of the censored data. Therefore, a number of statistical techniques, known as **survival methods**[1], have been developed to deal with these situations.

Censored data

Survival times are calculated from some baseline date that reflects a natural 'starting point' for the study (e.g. time of surgery or diagnosis of a condition) until the time that a patient reaches the endpoint of interest. Often, however, we may not know when the patient reached the endpoint, only that s/he remained free of the endpoint while in the study. For example, patients in a trial of a new drug for HIV infection may remain AIDS-free when they leave the study. This may either be because the trial ended while they were still AIDS-free, or because these individuals withdrew from the trial early before developing AIDS, or because they died of non-AIDS causes before the end of follow-up. Such data are described as **right-censored**. These patients were known *not* to have reached the endpoint when they were last under follow-up, and this information should be incorporated into the analysis.

Where follow-up does not begin until after the baseline date, survival times can also be **left-censored**.

Displaying survival data

- A separate horizontal line can be drawn for each patient, its length indicating the survival time. Lines are drawn from left to right, and patients who reach the endpoint and those who are censored can be distinguished by the use of different symbols at the end of the line (Fig. 44.1). However, these plots do not summarize the data and it is difficult to get a feel for the survival experience overall.
- **Survival curves**, usually calculated by the **Kaplan–Meier** method, display the cumulative probability (the **survival probability**) of an individual remaining free of the endpoint at any time after baseline (Fig. 44.2). The survival probability will only change when an endpoint occurs, and thus the resulting 'curve' is drawn as a series of steps. An alternative method of calculating survival probabilities, using a **lifetable** approach, can be used when the time to reach the endpoint is only known to within a particular time interval (e.g. within a year). The survival probabilities using either method are simple but time-consuming to calculate, and can be easily obtained from most statistical packages.

Summarizing survival

We often summarize survival by quoting survival probabilities (with confidence intervals) at certain time points on the curve, for example, the 5 year survival rates in patients after treatment for breast cancer. Alternatively, the median time to reach the endpoint (the time at which 50% of the individuals have *progressed*) can be quoted.

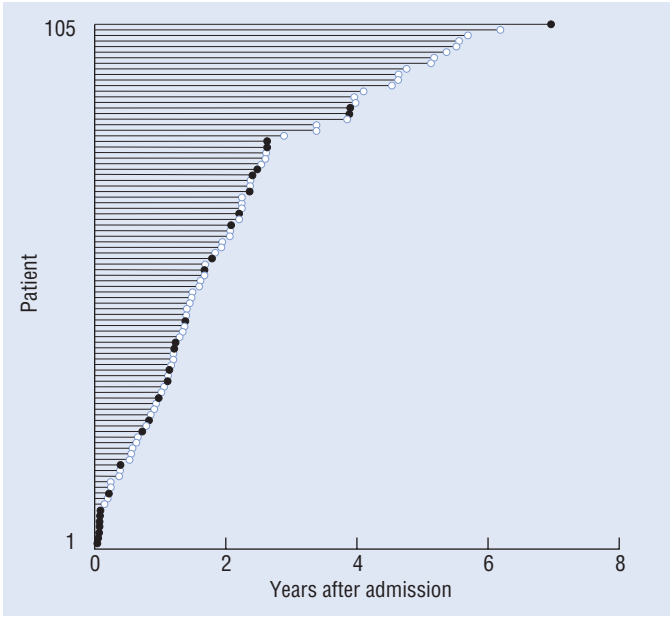


Figure 44.1 Survival experience of 105 patients following admission with cirrhosis. Filled blank circles indicate patients who died, open circles indicate those who remained alive at the end of follow-up.

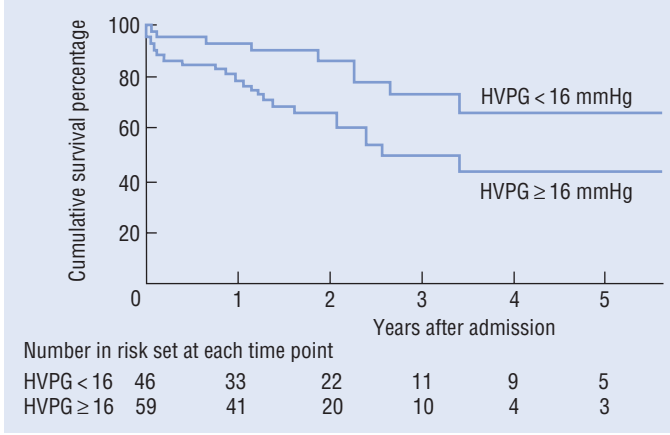


Figure 44.2 Kaplan–Meier curves showing the survival probability, expressed as a percentage, following admission for cirrhosis, stratified by baseline HVPG measurement.

[1] Collett, D. (2003). *Modelling Survival Data in Medical Research*. 2nd Edn. Chapman and Hall/CRC, London.

Comparing survival

We may wish to assess the impact of a number of factors of interest on survival, e.g. treatment, disease severity. Survival curves can be plotted separately for subgroups of patients; they provide a means of assessing visually whether different groups of patients reach the endpoint at different rates (Fig. 44.2). We can test formally whether there are any significant differences in progression rates between the different groups by, for example, using the log-rank test or regression models.

The log-rank test

This non-parametric test addresses the null hypothesis that there are no differences in survival times in the groups being studied, and compares events occurring at all time points on the survival curve. We cannot assess the independent roles of more than one factor on the time to the endpoint using the log-rank test.

Regression models

We can generate a regression model to quantify the relationships between one or more factors of interest and survival. At any point in time, t, an individual, i, has an instantaneous risk of reaching the endpoint (often known as the **hazard**, or $\lambda_i(t)$), given that s/he has not reached it up to that point in time. For example, if death is the endpoint, the hazard is the risk of dying at time t. This instantaneous hazard is usually very small and is of limited interest. However, we may want to know whether there are any systematic differences between the hazards, over all time points, of individuals with different characteristics. For example, is the hazard generally reduced in individuals treated with a new therapy compared with those treated with a placebo, when we take into account other factors, such as age or disease severity?

We can use the **Cox proportional hazards model** to test the independent effects of a number of explanatory variables (factors) on the hazard. It is of the form:

$$\lambda_i(t) = \lambda_0(t)\exp\{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k\}$$

where $\lambda_i(t)$ is the hazard for individual i at time t, $\lambda_0(t)$ is an arbitrary baseline hazard (in which we are not interested), $x_1, \ldots, x_k$ are explanatory variables in the model and $\beta_1, \ldots, \beta_k$ are the corresponding coefficients. We obtain estimates, $b_1, \ldots, b_k$, of these parameters using a form of maximum likelihood known as **partial likelihood**. The exponential of these values (e.g. $\exp\{b_1\} = e^{b_1}$) are the estimated **relative hazards** or **hazard ratios**. For a particular value of x_1, the hazard ratio is the estimated hazard of disease for $(x_1 + 1)$ relative to the estimated hazard of disease for x_1, while adjusting for all other x's in the equation. The relative hazard is interpreted in a similar manner to the odds ratio in logistic regression (Chapter 30) or the relative rate in Poisson regression (Chapter 31); therefore values above one indicate a raised hazard, values below one indicate a decreased hazard and values equal to one indicate that there is no increased or decreased hazard of the endpoint. A confidence interval can be calculated for the relative hazard and a significance test performed to assess its departure from 1.

The relative hazard is assumed to be constant over time in this model (i.e. the hazards for the groups to be compared are assumed to be *proportional*). It is important to check this assumption either by using graphical methods or by incorporating an interaction between the covariate and log(time) in the model and ensuring that it is non-significant[1].

Other models can be used to describe survival data, e.g. the **Exponential, Weibull** or **Gompertz** models, each of which assumes a specific probability distribution for the hazard function. However, these are beyond the scope of this book[1].

Example

Height of portal pressure (HVPG) is known to be associated with the severity of alcoholic cirrhosis but is rarely used as a predictor of survival in patients with cirrhosis. In order to assess the clinical value of this measurement, 105 patients admitted to hospital with cirrhosis, undergoing hepatic venography, were followed for a median of 566 days. The experience of these patients is illustrated in Fig. 44.1. Over the follow-up period, 33 patients died. **Kaplan–Meier curves** showing the cumulative survival percentage at any time point after baseline are displayed separately for individuals in whom HVPG was less than 16 mmHg (a value previously suggested to provide prognostic significance) and for those in whom HVPG was 16 mmHg or greater (Fig. 44.2).

The computer output for the **log-rank test** contained the following information:

Test	Chi-square	*df*	*P*-value
Log-rank	5.2995	1	0.0213

Thus there is a significant difference ($P = 0.02$) between survival times in the two groups. By 3 years after admission, 73.1% of those with a low HVPG measurement remained alive, compared with 49.6% of those with a higher measurement (Fig. 44.2).

A **Cox proportional hazards regression model** was used to investigate whether this relationship could be explained by differences in any known prognostic or demographic factors at baseline. Twenty variables were considered for inclusion in the model, including demographic, clinical and laboratory markers. Graphical methods suggested that the proportional hazards assumption was reasonable for these variables. A stepwise selection procedure (Chapter 33) was used to select the final optimal model, and the results are shown in Table 44.1.

The results in Table 44.1 indicate that raised HVPG remains independently associated with shorter survival after adjusting for other factors known to be associated with a poorer outcome. In particular, individuals with HVPG of 16 mmHg or higher had 2.46 (= exp{0.90}) times the hazard of death compared with those with lower levels ($P = 0.04$) after adjusting for other factors. In other words, the hazard of death is increased by 146% in these individuals. In addition, increased prothrombin time (hazard increases by 5% per additional second), increased bilirubin level (hazard increases by 5% per 10 additional mmol/L), the presence of ascites (hazard increases by 126% for a one level increase) and previous long-term endoscopic treatment (hazard increases by 246%) were all independently and significantly associated with outcome.

Table 44.1 Results of Cox proportional hazards regression analysis.

Variable (and coding)	*df*	Parameter estimate	Standard error	*P*-value	Estimated relative hazard	95% CI for relative hazard
HVPG* (0 = <16, 1 = ≥16 mmHg)	1	0.90	0.44	0.04	2.46	(1.03–5.85)
Prothrombin time (secs)	1	0.05	0.01	0.0002	1.05	(1.02–1.07)
Bilirubin (10 mmol/L)	1	0.05	0.02	0.04	1.05	(1.00–1.10)
Ascites (0 = none, 1 = mild, 2 = moderate/severe)	1	0.82	0.18	0.0001	2.26	(1.58–3.24)
Previous long-term endoscopic treatment (0 = no, 1 = yes)	1	1.24	0.41	0.003	3.46	(1.54–7.76)

HVPG*, Height of portal pressure.

Data kindly provided by Dr D. Patch and Prof. A.K. Burroughs, Liver Unit, Royal Free Hospital, London, UK.

45 Bayesian methods

The frequentist approach

The hypothesis tests described in this book are based on the **frequentist** approach to probability (Chapter 7) and inference that considers the number of times an event would occur if we were to repeat the experiment a large number of times. This approach is sometimes criticized for the following reasons.

- It uses only information obtained from the current study, and does not incorporate into the inferential process any other information we might have about the effect of interest, e.g. a clinician's views about the relative effectiveness of two therapies before a clinical trial is undertaken.
- It does not directly address the issues of greatest interest. In a drug comparison, we are usually really interested in knowing whether one drug is *more effective* than the other. However, the frequentist approach tests the hypothesis that the two drugs are *equally effective*. Although we conclude that one drug is superior to the other if the *P*-value is small, this probability (i.e. the *P*-value) describes the chance of getting the observed results if the drugs are equally effective, rather than the chance that one drug is more effective than the other (our real interest).
- It tends to over-emphasize the role of hypothesis testing and whether or not a result is significant, rather than the implications of the results.

The Bayesian approach

An alternative, **Bayesian**[1], approach to inference reflects an individual's personal degree of belief in a hypothesis, possibly based on information already available. Individuals usually differ in their degrees of belief in a hypothesis; in addition, these beliefs may change as new information becomes available. The Bayesian approach calculates the probability that a hypothesis is *true* (our focus of interest) by updating *prior* opinions about the hypothesis as new data become available.

Conditional probability

A particular type of probability, known as **conditional probability**, is fundamental to Bayesian analyses. This is the probability of an event, given that another event has already occurred. As an illustration, consider an example. The incidence of haemophilia A in the general population is approximately 1 in 10000 male births. However, if we know that a woman is a carrier for haemophilia, this incidence increases to around 1 in 2 male births. Therefore, the probability that a male child has haemophilia, given that his mother is a carrier, is very different to the unconditional probability that he has haemophilia if his mother's carrier status is unknown.

Bayes theorem

Suppose we are investigating a hypothesis (e.g. that a treatment effect equals some value). Bayes theorem converts a **prior probability**, describing an individual's belief in the hypothesis *before* the study is carried out, into a **posterior probability**, describing his/her belief *afterwards*. The posterior probability is, in fact, the conditional probability of the hypothesis, given the results from the study. **Bayes theorem** states that the **posterior probability** is proportional to the **prior probability** multiplied by a value, the **likelihood** of the observed results which describes the plausibility of the observed results if the hypothesis is true (Chapter 32).

Diagnostic tests in a Bayesian framework

Almost all clinicians intuitively use a Bayesian approach in their reasoning when making a diagnosis. They build a picture of the patient based on clinical history and/or the presence of symptoms and signs. From this, they decide on the *most likely* diagnosis, having eliminated other diagnoses on the presumption that they are unlikely to be true, given what they know about the patient. They may subsequently confirm or amend this diagnosis in the light of new evidence, e.g. if the patient responds to treatment or a new symptom develops.

When an individual attends a clinic, the clinician usually has some idea of the probability that the individual has the disease —the

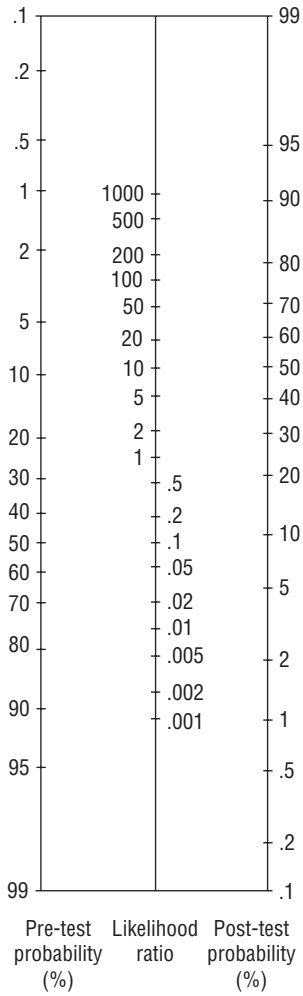


Figure 45.1 Fagan's nomogram for interpreting a diagnostic test result. Adapted from Sackett, D.L., Richardson, W.S., Rosenberg, W., Haynes, R.B. (1997) *Evidence-based Medicine: How to Practice and Teach EBM*. Churchill-Livingstone, London, with permission.

[1] Freedman, L. (1996) Bayesian statistical methods. A natural way to assess clinical evidence. *British Medical Journal*, **313**, 569–570.

prior or **pre-test probability**. If nothing else is known about the patient, this is simply the **prevalence** (Chapters 12 and 38) of the disease in the population. We can use Bayes theorem to change the prior probability into a posterior probability. This is most easily achieved if we incorporate the **likelihood ratio** (Chapter 32), based on information obtained from the most recent investigation (e.g. a diagnostic test result), into Bayes theorem. The likelihood ratio of a positive test result is the chance of a positive test result if the patient has disease, divided by that if s/he is disease-free. We discussed the likelihood ratio in this context in Chapter 38, and showed that it could be used to indicate the usefulness of a diagnostic test. We now use it to express Bayes theorem in terms of odds (Chapter 16):

Posterior odds of disease = prior odds × likelihood ratio of a positive test result

where

$$\text{Prior odds} = \frac{\text{prior probability}}{(1 - \text{prior probability})}$$

The posterior odds is simple to calculate, but for easier interpretation, we convert the odds back into a probability using the relationship:

$$\text{Posterior probability} = \frac{\text{posterior odds}}{(1 + \text{posterior odds})}$$

This **posterior** or **post-test probability** is the probability that the patient has the disease, given a positive test result. It is similar to the positive predictive value (Chapter 38) but takes account of the prior probability that the individual has the disease.

A simpler way to perform these calculations is to use **Fagan's nomogram** (see Fig. 45.1); by connecting the pre-test probability (expressed as a percentage) to the likelihood ratio and extending the line, we can evaluate the post-test probability.

Disadvantages of Bayesian methods

As part of any Bayesian analysis, it is necessary to specify the prior probability of the hypothesis (e.g. the pre-test probability that a patient has disease). Because of the subjective nature of these priors, individual researchers and clinicians may choose different values for them. For this reason, Bayesian methods are often criticized as being arbitrary. Where the most recent evidence from the study (i.e. the likelihood) is very strong, however, the influence of the prior information is minimized (at its extreme, the results will be completely uninfluenced by the prior information).

The calculations involved in many Bayesian analyses are complex, usually requiring sophisticated statistical packages that are highly computer intensive. Therefore, despite being intuitively appealing, Bayesian methods have not been used widely. However, the availability of powerful personal computers means that their use is becoming more common.

Example

In the example in Chapter 38 we showed that in bone marrow transplant recipients, a viral load above 5 $\log_{10}$ genomes/mL gave the optimal sensitivity and specificity of a test to predict the development of severe clinical disease. The likelihood ratio for a positive test for this cut-off value was 13.3.

If we believe that the prevalence of severe disease as a result of cytomegalovirus (CMV) infection after bone marrow transplantation is approximately 33%, the prior probability of severe disease in these patients equals 0.33.

$$\text{Prior odds} = \frac{0.33}{0.67} = 0.493$$

$$\begin{aligned}\text{Posterior odds} &= 0.493 \times \text{likelihood ratio} \\ &= 0.493 \times 13.3 \\ &= 6.557\end{aligned}$$

$$\text{Posterior probability} = \frac{6.557}{(1 + 6.557)} = \frac{6.557}{7.557} = 0.868$$

Therefore, if the individual has a CMV viral load above 5 $\log_{10}$ genomes/mL, and we assume that the pre-test probability of severe disease is 0.33 (i.e. 33%), then we believe that the individual has an 87% chance of developing severe disease. This can also be estimated directly from Fagan's nomogram (Fig. 45.1) by connecting the pre-test probability of 33% to a likelihood ratio of 13.3 and extending the line to cut the post-test probability axis. In contrast, if we believe that the probability that an individual will get severe disease is only 0.2 (i.e. pre-test probability equals 20%), then the post-test probability will equal 77%.

In both cases, the post-test probability is much higher than the pre-test probability, indicating the usefulness of a positive test result. Furthermore, both results indicate that the patient is at high risk of developing severe disease after transplantation and that it may be sensible to start anti-CMV therapy. Therefore, despite having very different prior probabilities, the general conclusion remains the same in each case.

Appendix A: Statistical tables

This appendix contains statistical tables discussed in the text. We have provided only limited *P*-values because data are usually analysed using a computer, and *P*-values are included in its output. Other texts, such as that by Fisher and Yates[1], contain more comprehensive tables. You can also obtain the *P*-value directly from some computer packages, given a value of the test statistic. Empty cells in a table are an indication that values do not exist.

Table A1 contains the probability in the two tails of the distribution of a variable, z, which follows the Standard Normal distribution. The *P*-values in Table A1 relate to the absolute values of z, so if z is negative, we ignore its sign. For example, if a test statistic that follows the Standard Normal distribution has the value 1.1, $P = 0.271$.

Table A2 and **Table A3** contain the probability in the two tails of a distribution of a variable that follows the *t*-distribution (Table A2) or the Chi-squared distribution (Table A3) with given degrees of freedom (*df*). To use Table A2 or Table A3, if the absolute value of the test statistic (with given *df*) lies between the tabulated values in two columns, then the two-tailed *P*-value lies between the *P*-values specified at the top of these columns. If the test statistic is to the right of the final column, $P < 0.001$; if it is to the left of the second column, $P > 0.10$. For example, (i) Table A2: if the test statistic is 2.62 with $df = 17$, then $0.01 < P < 0.05$; (ii) Table A3: if the test statistic is 2.62 with $df = 17$, then $P < 0.001$.

Table A4 contains often used *P*-values and their corresponding values for z, a variable with a Standard Normal distribution. This table may be used to obtain multipliers for the calculation of confidence intervals (CI) for Normally distributed variables. For example, for a 95% confidence interval, the multiplier is 1.96.

Table A5 contains *P*-values for a variable that follows the *F*-distribution with specified degrees of freedom in the numerator and denominator. When comparing variances (Chapter 35), we usually use a two-tailed *P*-value. For the analysis of variance (Chapter 22), we use a one-tailed *P*-value. For given degrees of freedom in the numerator and denominator, the test is significant at the level of *P* quoted in the table if the test statistic is greater than the tabulated value. For example, if the test statistic is 2.99 with $df = 5$ in the numerator and $df = 15$ in the denominator, then $P < 0.05$ for a one-tailed test.

Table A6 contains two-tailed *P*-values of the sign test of r responses of a particular type out of a total of n' responses. For a one-sample test, r equals the number of values above (or below) the median (Chapter 19). For a paired test, r equals the number of positive (or negative) differences (Chapter 20) or the number of preferences for a particular treatment (Chapter 23). n' equals the number of values not equal to the median, non-zero differences or actual preferences, as relevant. For example, if we observed three positive differences out of eight non-zero differences, then $P = 0.726$.

Table A7 contains the ranks of the values which determine the upper and lower limits of the approximate 90%, 95% and 99% confidence intervals (CI) for the median. For example, if the sample size is 23, then the limits of the 95% confidence interval are defined by the 7th and 17th ordered values.

For sample sizes greater than 50, find the observations that correspond to the ranks (to the nearest integer) equal to: (i) $n/2 - z\sqrt{n}/2$; and (ii) $1 + n/2 + z\sqrt{n}/2$; where n is the sample size and $z = 1.64$ for a 90% CI, $z = 1.96$ for a 95% CI, and $z = 2.58$ for a 99% CI (the values of z being obtained from the Standard Normal distribution, Table A4). These observations define (i) the lower, and (ii) the upper confidence limits for the median.

Table A8 contains the range of values for the sum of the ranks (T_+ or T_-), which determines significance in the Wilcoxon signed ranks test (Chapter 20). If the sum of the ranks of the positive (T_+) or negative (T_-) differences, out of n' non-zero differences, is equal to or outside the tabulated limits, the test is significant at the *P*-value quoted. For example, if there are 16 non-zero differences and $T_+ = 21$, then $0.01 < P < 0.05$.

Table A9 contains the range of values for the sum of the ranks (T) which determines significance for the Wilcoxon rank sum test (Chapter 21) at (a) the 5% level and (b) the 1% level. Suppose we have two samples of sizes n_S and n_L, where $n_S \leq n_L$. If the sum of the ranks of the group with the smaller sample size, n_S, is equal to or outside the tabulated limits, the test is significant at (a) the 5% level or (b) the 1% level. For example, if $n_S = 6$ and $n_L = 8$, and the sum of the ranks in the group of six observations equals 39, then $P > 0.05$.

Tables A10 and **Table A11** contain two-tailed *P*-values for Pearson's (Table A10) and Spearman's (Table A11) correlation coefficients when testing the null hypothesis that the relevant correlation coefficient is zero (Chapter 26). Significance is achieved, for a given sample size, at the stated *P*-value if the absolute value (i.e. ignoring its sign) of the sample value of the correlation coefficient exceeds the tabulated value. For example, if the sample size equals 24 and Pearson's $r = 0.58$, then $0.001 < P < 0.01$. If the sample size equals 7 and Spearman's $r_s = -0.63$, then $P > 0.05$.

Table A12 contains the digits 0–9 arranged in random order.

[1] Fisher, R.A. and Yates, F. (1963) *Statistical Tables for Biological, Agricultural and Medical Research*, 6th edn. Oliver and Boyd, Edinburgh.

Table A1 Standard Normal distribution.

z	2-tailed *P*-value
0.0	1.000
0.1	0.920
0.2	0.841
0.3	0.764
0.4	0.689
0.5	0.617
0.6	0.549
0.7	0.484
0.8	0.424
0.9	0.368
1.0	0.317
1.1	0.271
1.2	0.230
1.3	0.194
1.4	0.162
1.5	0.134
1.6	0.110
1.7	0.089
1.8	0.072
1.9	0.057
2.0	0.046
2.1	0.036
2.2	0.028
2.3	0.021
2.4	0.016
2.5	0.012
2.6	0.009
2.7	0.007
2.8	0.005
2.9	0.004
3.0	0.003
3.1	0.002
3.2	0.001
3.3	0.001
3.4	0.001
3.5	0.000

Derived using Microsoft Excel Version 5.0.

Table A2 *t*-distribution.

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	6.314	12.706	63.656	636.58
2	2.920	4.303	9.925	31.600
3	2.353	3.182	5.841	12.924
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.869
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.408
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
11	1.796	2.201	3.106	4.437
12	1.782	2.179	3.055	4.318
13	1.771	2.160	3.012	4.221
14	1.761	2.145	2.977	4.140
15	1.753	2.131	2.947	4.073
16	1.746	2.120	2.921	4.015
17	1.740	2.110	2.898	3.965
18	1.734	2.101	2.878	3.922
19	1.729	2.093	2.861	3.883
20	1.725	2.086	2.845	3.850
21	1.721	2.080	2.831	3.819
22	1.717	2.074	2.819	3.792
23	1.714	2.069	2.807	3.768
24	1.711	2.064	2.797	3.745
25	1.708	2.060	2.787	3.725
26	1.706	2.056	2.779	3.707
27	1.703	2.052	2.771	3.689
28	1.701	2.048	2.763	3.674
29	1.699	2.045	2.756	3.660
30	1.697	2.042	2.750	3.646
40	1.684	2.021	2.704	3.551
50	1.676	2.009	2.678	3.496
100	1.660	1.984	2.626	3.390
200	1.653	1.972	2.601	3.340
5000	1.645	1.960	2.577	3.293

Derived using Microsoft Excel Version 5.0.

Table A3 Chi-squared distribution.

	Two-tailed *P*-value			
df	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.827
2	4.605	5.991	9.210	13.815
3	6.251	7.815	11.345	16.266
4	7.779	9.488	13.277	18.466
5	9.236	11.070	15.086	20.515
6	10.645	12.592	16.812	22.457
7	12.017	14.067	18.475	24.321
8	13.362	15.507	20.090	26.124
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.527
14	21.064	23.685	29.141	36.124
15	22.307	24.996	30.578	37.698
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.791
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.819
20	28.412	31.410	37.566	45.314
21	29.615	32.671	38.932	46.796
22	30.813	33.924	40.289	48.268
23	32.007	35.172	41.638	49.728
24	33.196	36.415	42.980	51.179
25	34.382	37.652	44.314	52.619
26	35.563	38.885	45.642	54.051
27	36.741	40.113	46.963	55.475
28	37.916	41.337	48.278	56.892
29	39.087	42.557	49.588	58.301
30	40.256	43.773	50.892	59.702
40	51.805	55.758	63.691	73.403
50	63.167	67.505	76.154	86.660
60	74.397	79.082	88.379	99.608
70	85.527	90.531	100.43	112.32
80	96.578	101.88	112.33	124.84
90	107.57	113.15	124.12	137.21
100	118.50	124.34	135.81	149.45

Derived using Microsoft Excel Version 5.0.

Table A4 Standard Normal distribution.

	Two-tailed *P*-value				
	0.50	0.10	0.05	0.01	0.001
Relevant CI	50%	90%	95%	99%	99.9%
z (i.e. CI multiplier)	0.67	1.64	1.96	2.58	3.29

Derived using Microsoft Excel Version 5.0.

Table A6 Sign test.

	r = number of 'positive differences' (see explanation)					
n′	0	1	2	3	4	5
4	0.125	0.624	1.000			
5	0.062	0.376	1.000			
6	0.032	0.218	0.688	1.000		
7	0.016	0.124	0.454	1.000		
8	0.008	0.070	0.290	0.726	1.000	
9	0.004	0.040	0.180	0.508	1.000	
10	0.001	0.022	0.110	0.344	0.754	1.000

Derived using Microsoft Excel Version 5.0.

Table A5 The *F*-distribution.

df of	2-tailed	1-tailed	Degrees of freedom (*df*) of the numerator												
denominator	*P*-value	*P*-value	1	2	3	4	5	6	7	8	9	10	15	25	500
1	**0.05**	**0.025**	**647.8**	**799.5**	**864.2**	**899.6**	**921.8**	**937.1**	**948.2**	**956.6**	**963.3**	**968.6**	**984.9**	**998.1**	**1017.0**
1	0.10	0.05	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	245.9	249.3	254.1
2	**0.05**	**0.025**	**38.51**	**39.00**	**39.17**	**39.25**	**39.30**	**39.33**	**39.36**	**39.37**	**39.39**	**39.40**	**39.43**	**39.46**	**39.50**
2	0.10	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.46	19.49
3	**0.05**	**0.025**	**17.44**	**16.04**	**15.44**	**15.10**	**14.88**	**14.73**	**14.62**	**14.54**	**14.47**	**14.42**	**14.25**	**14.12**	**13.91**
3	0.10	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.63	8.53
4	**0.05**	**0.025**	**12.22**	**10.65**	**9.98**	**9.60**	**9.36**	**9.20**	**9.07**	**8.98**	**8.90**	**8.84**	**8.66**	**8.50**	**8.27**
4	0.10	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.77	5.64
5	**0.05**	**0.025**	**10.01**	**8.43**	**7.76**	**7.39**	**7.15**	**6.98**	**6.85**	**6.76**	**6.68**	**6.62**	**6.43**	**6.27**	**6.03**
5	0.10	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.52	4.37
6	**0.05**	**0.025**	**8.81**	**7.26**	**6.60**	**6.23**	**5.99**	**5.82**	**5.70**	**5.60**	**5.52**	**5.46**	**5.27**	**5.11**	**4.86**
6	0.10	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.83	3.68
7	**0.05**	**0.025**	**8.07**	**6.54**	**5.89**	**5.52**	**5.29**	**5.12**	**4.99**	**4.90**	**4.82**	**4.76**	**4.57**	**4.40**	**4.16**
7	0.10	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.40	3.24
8	**0.05**	**0.025**	**7.57**	**6.06**	**5.42**	**5.05**	**4.82**	**4.65**	**4.53**	**4.43**	**4.36**	**4.30**	**4.10**	**3.94**	**3.68**
8	0.10	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.11	2.94
9	**0.05**	**0.025**	**7.21**	**5.71**	**5.08**	**4.72**	**4.48**	**4.32**	**4.20**	**4.10**	**4.03**	**3.96**	**3.77**	**3.60**	**3.35**
9	0.10	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.89	2.72
10	**0.05**	**0.025**	**6.94**	**5.46**	**4.83**	**4.47**	**4.24**	**4.07**	**3.95**	**3.85**	**3.78**	**3.72**	**3.52**	**3.35**	**3.09**
10	0.10	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.73	2.55
15	**0.05**	**0.025**	**6.20**	**4.77**	**4.15**	**3.80**	**3.58**	**3.41**	**3.29**	**3.20**	**3.12**	**3.06**	**2.86**	**2.69**	**2.41**
15	0.10	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.28	2.08
20	**0.05**	**0.025**	**5.87**	**4.46**	**3.86**	**3.51**	**3.29**	**3.13**	**3.01**	**2.91**	**2.84**	**2.77**	**2.57**	**2.40**	**2.10**
20	0.10	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.07	1.86
30	**0.05**	**0.025**	**5.57**	**4.18**	**3.59**	**3.25**	**3.03**	**2.87**	**2.75**	**2.65**	**2.57**	**2.51**	**2.31**	**2.12**	**1.81**
30	0.10	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.88	1.64
50	**0.05**	**0.025**	**5.34**	**3.97**	**3.39**	**3.05**	**2.83**	**2.67**	**2.55**	**2.46**	**2.38**	**2.32**	**2.11**	**1.92**	**1.57**
50	0.10	0.05	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.73	1.46
100	**0.05**	**0.025**	**5.18**	**3.83**	**3.25**	**2.92**	**2.70**	**2.54**	**2.42**	**2.32**	**2.24**	**2.18**	**1.97**	**1.77**	**1.38**
100	0.10	0.05	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.62	1.31
1000	**0.05**	**0.025**	**5.04**	**3.70**	**3.13**	**2.80**	**2.58**	**2.42**	**2.30**	**2.20**	**2.13**	**2.06**	**1.85**	**1.64**	**1.16**
1000	0.10	0.05	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.68	1.52	1.13

Derived using Microsoft Excel Version 5.0.

Table A7 Ranks for confidence intervals for the median.

	Approximate		
Sample size	90% CI	95% CI	99% CI
6	1,6	1,6	—
7	1,7	1,7	—
8	2,7	1,8	—
9	2,8	2,8	1,9
10	2,9	2,9	1,10
11	3,9	2,10	1,11
12	3,10	3,10	2,11
13	4,10	3,11	2,12
14	4,11	3,12	2,13
15	4,12	4,12	3,13
16	5,12	4,13	3,14
17	5,13	4,14	3,15
18	6,13	5,14	4,15
19	6,14	5,15	4,16
20	6,15	6,15	4,17
21	7,15	6,16	5,17
22	7,16	6,17	5,18
23	8,16	7,17	5,19
24	8,17	7,18	6,19
25	8,18	8,18	6,20
26	9,18	8,19	6,21
27	9,19	8,20	7,21
28	10,19	9,20	7,22
29	10,20	9,21	8,22
30	11,20	10,21	8,23
31	11,21	10,22	8,24
32	11,22	10,23	9,24
33	12,22	11,23	9,25
34	12,23	11,24	9,26
35	12,23	12,24	10,26
36	13,24	12,25	10,27
37	14,24	13,25	11,27
38	14,25	13,26	11,28
39	14,26	13,27	11,29
40	15,26	14,27	12,29
41	15,27	14,28	12,30
42	16,27	15,28	13,30
43	16,28	15,29	13,31
44	17,28	15,30	13,32
45	17,29	16,30	14,32
46	17,30	16,31	14,33
47	18,30	17,31	15,33
48	18,31	17,32	15,34
49	19,31	18,32	15,35
50	19,32	18,33	16,35

Derived using Microsoft Excel Version 5.0.

Table A8 Wilcoxon signed ranks test.

	Two-tailed *P*-value		
n'	0.05	0.01	0.001
6	0–21	—	—
7	2–26	—	—
8	3–33	0–36	—
9	5–40	1–44	—
10	8–47	3–52	—
11	10–56	5–61	0–66
12	13–65	7–71	1–77
13	17–74	9–82	2–89
14	21–84	12–93	4–101
15	25–95	15–105	6–114
16	29–107	19–117	9–127
17	34–119	23–130	11–142
18	40–131	27–144	14–157
19	46–144	32–158	18–172
20	52–158	37–173	21–189
21	58–173	42–189	26–205
22	66–187	48–205	30–223
23	73–203	54–222	35–241
24	81–219	61–239	40–260
25	89–236	68–257	45–280

Adapted with permission from Altman, D.G. (1991) *Practical Statistics for Medical Research*. Copyright CRC Press, Boca Raton.

Table A9(a) Wilcoxon rank sum test for a two-tailed $P = 0.05$.

	n_s (the number of observations in the smaller sample)											
n_L	4	5	6	7	8	9	10	11	12	13	14	15
4	10–26	16–34	23–43	31–53	40–64	49–77	60–90	72–104	85–119	99–135	114–152	130–170
5	11–29	17–38	24–48	33–58	42–70	52–83	63–97	75–112	89–127	103–144	118–162	134–181
6	12–32	18–42	26–52	34–64	44–76	55–89	66–104	79–119	92–136	107–153	122–172	139–191
7	13–35	20–45	27–57	36–69	46–82	57–96	69–111	82–127	96–144	111–162	127–181	144–201
8	14–38	21–49	29–61	38–74	49–87	60–102	72–118	85–135	100–152	115–171	131–191	149–211
9	14–42	22–53	31–65	40–79	51–93	62–109	75–125	89–142	104–160	119–180	136–200	154–221
10	15–45	23–57	32–70	42–84	53–99	65–115	78–132	92–150	107–169	124–188	141–209	159–231
11	16–48	24–61	34–74	44–89	55–105	68–121	81–139	96–157	111–177	128–197	145–219	164–241
12	17–51	26–64	35–79	46–94	58–110	71–127	84–146	99–165	115–185	132–206	150–228	169–251
13	18–54	27–68	37–83	48–99	60–116	73–134	88–152	103–172	119–193	136–215	155–237	174–261
14	19–57	28–72	38–88	50–104	62–122	76–140	91–159	106–180	123–201	141–223	160–246	179–271
15	20–60	29–76	40–92	52–109	65–127	79–146	94–166	110–187	127–209	145–232	164–256	184–281

Table A9(b) Wilcoxon rank sum test for a two-tailed $P = 0.01$.

	n_s (the number of observations in the smaller sample)											
n_L	4	5	6	7	8	9	10	11	12	13	14	15
4	—	—	21–45	28–56	37–67	46–80	57–93	68–108	81–123	94–140	109–157	125–175
5	—	15–40	22–50	29–62	38–74	48–87	59–101	71–116	84–132	98–149	112–168	128–187
6	10–34	16–44	23–55	31–67	40–80	50–94	61–109	73–125	87–141	101–159	116–178	132–198
7	10–38	16–49	24–60	32–73	42–86	52–101	64–116	76–133	90–150	104–169	120–188	136–209
8	11–48	17–53	25–65	34–78	43–93	54–108	66–124	79–141	93–159	108–178	123–199	140–220
9	11–45	18–57	26–70	35–84	45–99	56–115	68–132	82–149	96–168	111–188	127–209	144–231
10	12–48	19–61	27–75	37–89	47–105	58–122	71–139	84–158	99–177	115–197	131–219	149–241
11	12–52	20–65	28–80	38–95	49–111	61–128	73–147	87–166	102–186	118–207	135–229	153–252
12	13–55	21–69	30–84	40–100	51–117	63–135	76–154	90–174	105–195	122–216	139–239	157–263
13	13–59	22–73	31–89	41–106	53–123	65–142	79–161	93–182	109–203	125–226	143–249	162–273
14	14–62	22–78	32–94	43–111	54–130	67–149	81–169	96–190	112–212	129–235	147–259	166–284
15	15–65	23–82	33–99	44–117	56–136	69–156	84–176	99–198	115–221	133–244	151–269	171–294

Extracted with permission from Diem, K. (1970) *Documenta Geigy Scientific Tables*, 7th edn, Blackwell Publishing, Oxford.

Table A10 Pearson's correlation coefficient.

Sample size	Two-tailed *P*-value		
	0.05	0.01	0.001
5	0.878	0.959	0.991
6	0.811	0.917	0.974
7	0.755	0.875	0.951
8	0.707	0.834	0.925
9	0.666	0.798	0.898
10	0.632	0.765	0.872
11	0.602	0.735	0.847
12	0.576	0.708	0.823
13	0.553	0.684	0.801
14	0.532	0.661	0.780
15	0.514	0.641	0.760
16	0.497	0.623	0.742
17	0.482	0.606	0.725
18	0.468	0.590	0.708
19	0.456	0.575	0.693
20	0.444	0.561	0.679
21	0.433	0.549	0.665
22	0.423	0.537	0.652
23	0.413	0.526	0.640
24	0.404	0.515	0.629
25	0.396	0.505	0.618
26	0.388	0.496	0.607
27	0.381	0.487	0.597
28	0.374	0.479	0.588
29	0.367	0.471	0.579
30	0.361	0.463	0.570
35	0.334	0.430	0.532
40	0.312	0.403	0.501
45	0.294	0.380	0.474
50	0.279	0.361	0.451
55	0.266	0.345	0.432
60	0.254	0.330	0.414
70	0.235	0.306	0.385
80	0.220	0.286	0.361
90	0.207	0.270	0.341
100	0.217	0.283	0.357
150	0.160	0.210	0.266

Extracted with permission from Diem, K. (1970) *Documenta Geigy Scientific Tables*, 7th edn, Blackwell Publishing, Oxford.

Table A11 Spearman's correlation coefficient.

Sample size	Two tailed *P*-value		
	0.05	0.01	0.001
5	1.000		
6	0.886	1.000	
7	0.786	0.929	1.000
8	0.738	0.881	0.976
9	0.700	0.833	0.933
10	0.648	0.794	0.903

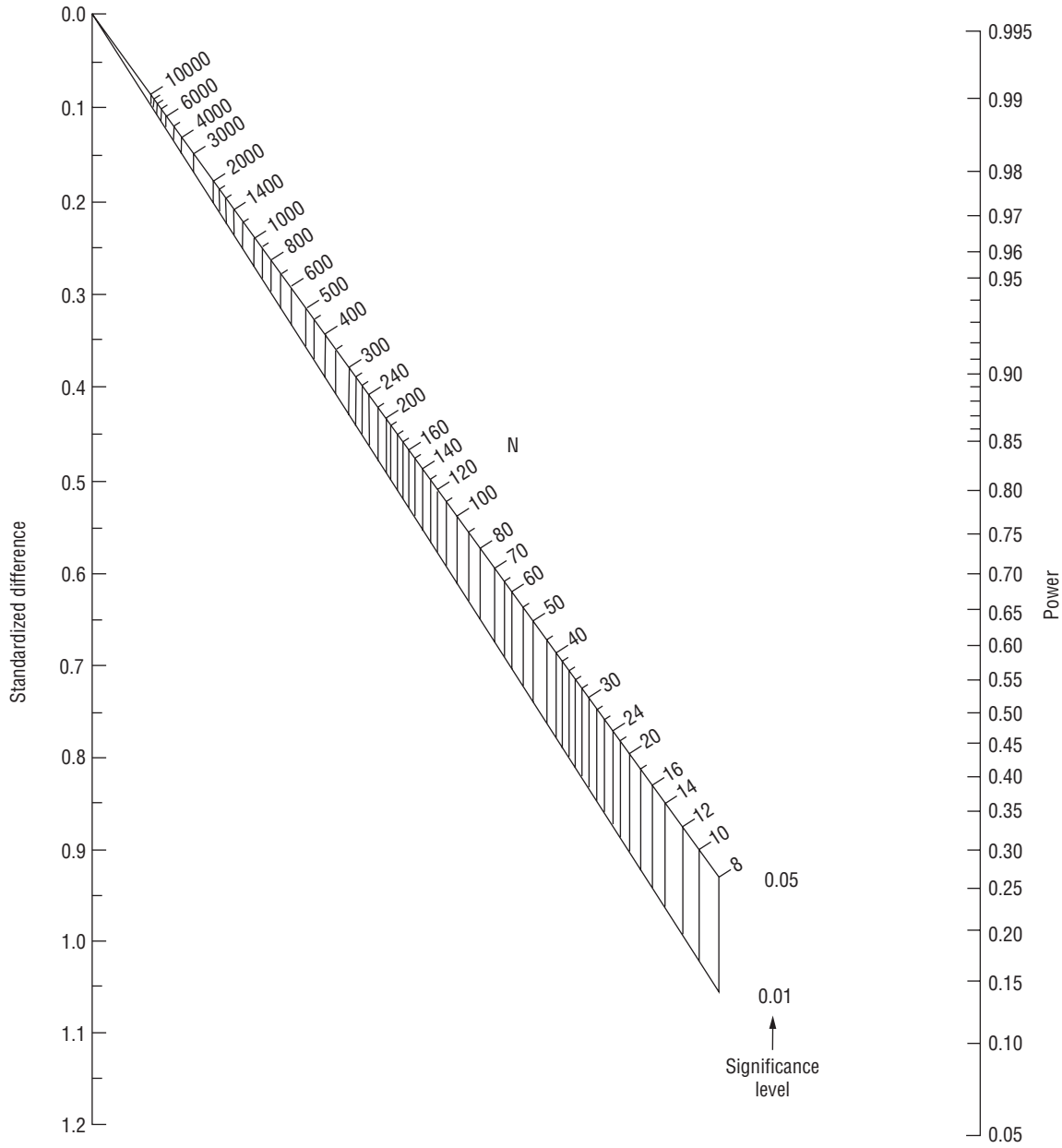
Adapted from Siegel, S. & Castellan, N.J. (1988) *Nonparametric Statistics for the Behavioural Sciences*, 2nd edn, McGraw-Hill, New York, and used with permission of McGraw-Hill Companies.

Table A12 Random numbers.

34814	68020	28998	51687	40088	35458	24708	01815	53776
99106	50899	07394	91071	22411	61643	64435	62552	64316
47185	31782	48894	68790	51852	36918	05737	90653	61123
81354	57296	39329	52263	43194	51624	42429	61367	41207
83467	85622	95778	05347	00445	51334	29445	99176	30091
27924	34167	57060	57535	32278	16949	04960	04116	91467
58319	88164	94130	07743	16917	15681	93572	99753	49117
49732	66702	72425	99117	49298	87265	14195	83391	19794
69594	26749	68743	39139	44495	11944	12970	56523	62411
30074	97517	97450	54251	51777	21073	03909	26519	39578
81147	57508	93479	87826	28965	74474	97468	80149	17834
74689	28933	59819	93052	61325	83145	44684	72958	91824
14802	25982	48024	15461	37570	44685	47386	09504	77831
68501	34194	85355	38411	46559	41694	99678	88268	86674
48734	92671	85252	85985	34228	91289	56331	14683	36493
84102	81699	97352	54509	93196	51204	43351	11818	41179
28432	32873	83834	09862	12720	64569	42218	26726	80866
91458	82524	75523	01276	19591	47473	90251	99103	72947
45435	30389	69732	81962	30243	96199	33546	39672	83760
23557	78437	44957	98728	65674	34701	83398	54102	65845
30395	91850	52004	04844	28848	19728	96571	13317	70859
69991	12755	97916	57639	43445	90463	85556	35469	19749
32980	43608	20592	72527	63583	46443	53929	87219	55198
59776	37035	53765	55196	68659	71429	25225	91942	51132
73714	79868	23880	92254	72984	07792	81306	24277	82366
61547	16575	68520	59869	67299	73565	77316	96682	18031
87737	01058	76012	76247	75616	51335	70364	78942	40564
98669	08334	40520	78389	56498	74336	02434	48599	67579
81535	46690	92814	44456	29227	48122	30522	13852	48436
05975	47110	32733	46929	98261	52193	83215	53192	83109

Derived using Microsoft Excel Version 5.0.

Appendix B: Altman's nomogram for sample size calculations (Chapter 36)



Extracted from: Altman, D.G. (1982) How large a sample? In: *Statistics in Practice* (eds S.M. Gore & D.G. Altman). BMA, London, with permission from Blackwell Publishing Ltd.

Appendix C: Typical computer output

Analysis of pocket depth data described in Topic 20, generated by SPSS

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
PDIFF	96	50.0%	96	50.0%	192	100.0%

Descriptives

			Statistic	Std. Error
PDDIFF	Mean		.1486	5.716E-02
	95% Confidence Interval for Mean	Lower Bound	3.511E-02	
		Upper Bound	.2621	
	5% Trimmed Mean		.1934	
	Median		.2242	
	Variance		.314	
	Std. Deviation		.5601	
	Minimum		-2.69	
	Maximum		2.48	
	Range		5.17	
	Interquartile Range		.3171	
	Skewness		-2.243	.246
	Kurtosis		15.146	.488

This is 0.05716

Table of summary measures for the differences ('before' minus 'after') in pocket depth

PDDIFF Stem-and-Leaf Plot

```
  Frequency stem & Leaf
    4.00    Extremes        (=<-.910
    1.00          -3   .   8
    4.00          -2   .   0016
    3.00          -1   .   0.29
    7.00          -0   .   0127789
   13.00           0   .   0112234588889
   11.00           1   .   3344566679
   21.00           2   .   001122334445667888999
   15.00           3   .   112233344557888
   11.00           4   .   00123347899
    1.00           5   .   0
    1.00           6   .   7
    1.00           7   .   9
    3.00    Extremes        (>=.84)

Stem width:      .10
Each leaf:       1 case(s)
```

Stem-and-leaf plot shows that the differences are approximately Normally distributed.

Topic 20

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	PDAVBEFO	2.5787	96	.4771	4.869E-02
	PDAVAFTE	2.4301	96	.3827	3.906E-02

Results of paired t-test show that $\bar{d}$=0.1486, s_d=0.5601, t=2.60 and P-value=0.011

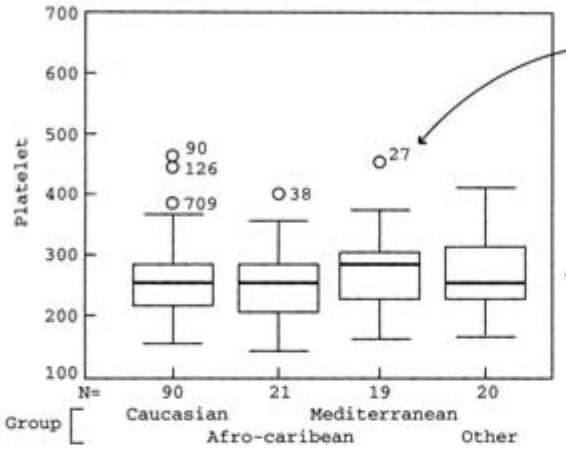
Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the difference Lower	Upper			
Pair 1	PDAVBEFO-PDAVAFTE	.1486	.5601	5.716E02	3.511E-02	.2621	2.600	95	.011

P-value

This is 0.05716

Analysis of platelet data described in Topic 22, generated by SPSS



Patient 27 is an outlier

Box-plots showing distribution of platelet counts in the four ethnic groups

Oneway

Report

Platelet

Group	Mean	N	Std. Deviation	Std. Error Of Mean
Caucasian	268.1000	90	77.0784	8.1248
Afro-caribbean	254.2857	21	67.5005	14.7298
Mediterranean	281.0526	19	71.0934	16.3099
Other	273.3000	20	63.4243	14.1821
Total	268.5000	150	73.0451	5.9641

Summary measures for each of the four groups

Platelet

Test of Homogeneity of Variances

Levene Statistic	df1	df2	Sig.
.041	3	146	.989

Results from Levine's test; the *P*-value of 0.989 indicates that there is no evidence that the variances are different in the four groups

Platelet

Anova

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7711.967	3	2570.656	.477	.699
Within Groups	787289.533	146	5392.394		
Total	795001.500	149			

The ANOVA table

P-value

Topic 22

Analysis of FEV1 data described in Topic 21, generated by SAS

The SAS System

OBS	GRP	FEV
1	Placebo	1.28571
2	Placebo	1.31250
3	Placebo	1.60000
4	Placebo	1.41250
5	Placebo	1.60000
49	Treated	1.60000
50	Treated	1.80000
51	Treated	1.94286
52	Treated	1.84286
53	Treated	1.90000

Print out of first five observations in each group

..... Treatment Group=Placebo .....

Univariate Procedure

Variable=FEV

Moments

N	48	Sum Wgts	48
Mean	1.536759	Sum	73.76441
Std Dev	0.245819	Variance	0.060427
Skewness	0.272608	Kurtosis	0.500457
USS	116.1981	CSS	2.840059
CV	15.99592	Std Mean	0.035481
T:Mean=0	43.31232	Pr>|T|	0.0001
Num ^=0	48	Num > 0	48
M (Sign)	24	Pr>=|M|	0.0001
Sgn Rank	588	Pr>=|S|	0.0001

Univariate summary statistics showing that the mean and median are fairly similar in the placebo group. Thus we believe that the values are approximately Normally distributed

Median

Quantiles (Def=5)

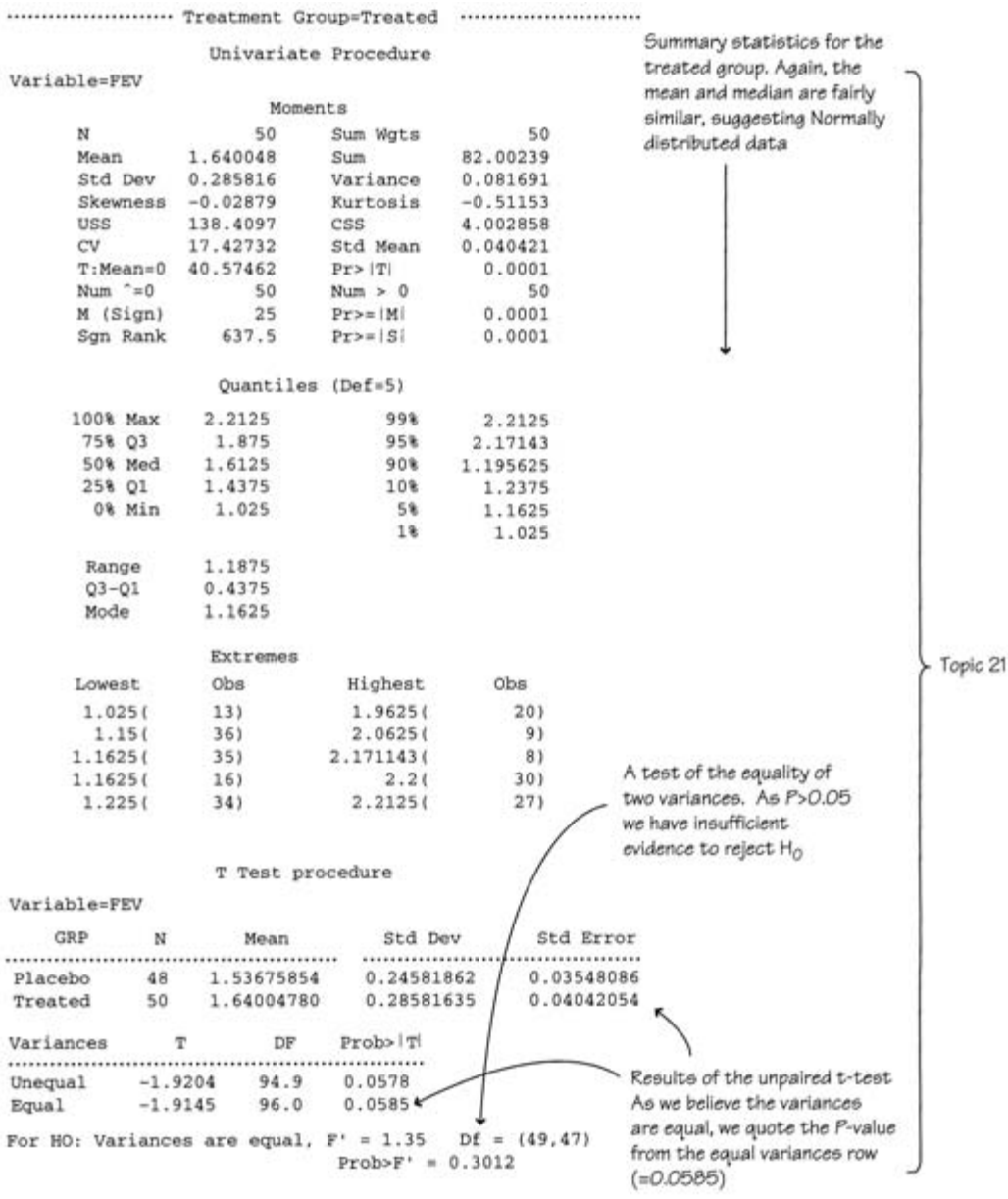
100% Max	2.1875	99%	2.1875
75% Q3	1.7	95%	1.91429
50% Med	1.551785	90%	1.85714
25% Q1	1.36905	10%	1.28571
0% Min	1	5%	1.12857
		1%	1
Range	1.1875		
Q3-Q1	0.33095		
Mode	1.3875		

Extremes

Lowest	Obs	Highest	Obs
1(	21)	1.85714(	47)
1.04(	33)	1.9(	26)
1.12857(	45)	1.91429(	46)
1.18571(	12)	2.1125(	27)
1.28571(	1)	2.1875(	20)

Topic 21

continued


Treatment Group=Treated
Univariate Procedure
Variable=FEV
Moments
N 50 Sum Wgts 50
Mean 1.640048 Sum 82.00239
Std Dev 0.285816 Variance 0.081691
Skewness -0.02879 Kurtosis -0.51153
USS 138.4097 CSS 4.002858
CV 17.42732 Std Mean 0.040421
T:Mean=0 40.57462 Pr>|T| 0.0001
Num ^=0 50 Num > 0 50
M (Sign) 25 Pr>=|M| 0.0001
Sgn Rank 637.5 Pr>=|S| 0.0001
Quantiles (Def=5)
100% Max 2.2125 99% 2.2125
75% Q3 1.875 95% 2.17143
50% Med 1.6125 90% 1.195625
25% Q1 1.4375 10% 1.2375
0% Min 1.025 5% 1.1625
1% 1.025
Range 1.1875
Q3-Q1 0.4375
Mode 1.1625
Extremes
Lowest Obs Highest Obs
1.025(13) 1.9625(20)
1.15(36) 2.0625(9)
1.1625(35) 2.171143(8)
1.1625(16) 2.2(30)
1.225(34) 2.2125(27)
T Test procedure
Variable=FEV
GRP N Mean Std Dev Std Error
Placebo 48 1.53675854 0.24581862 0.03548086
Treated 50 1.64004780 0.28581635 0.04042054
Variances T DF Prob>|T|
Unequal -1.9204 94.9 0.0578
Equal -1.9145 96.0 0.0585
For HO: Variances are equal, F' = 1.35 Df = (49,47)
Prob>F' = 0.3012
Summary statistics for the treated group. Again, the mean and median are fairly similar, suggesting Normally distributed data
A test of the equality of two variances. As P>0.05 we have insufficient evidence to reject H0
Results of the unpaired t-test As we believe the variances are equal, we quote the P-value from the equal variances row (=0.0585)
Topic 21

Analysis of anthropometric data described in Topics 26, 28 and 29, generated by SAS

OBS	SBP	Height	Weight	Sex
1	91.00	119.7	20.0	0
2	122.50	124.6	42.5	0
3	109.50	111.3	19.8	0
4	100.50	110.3	18.9	0
5	99.00	112.5	19.0	0
6	103.50	115.1	19.3	0
7	101.00	116.3	19.6	0
8	103.00	111.1	17.1	1
9	106.50	117.2	20.7	1
10	102.50	113.2	22.1	1

Print out of data from first 10 children

Correlation Analysis

4 'VAR' Variables: SBP Height Weight Age

Simple Statistics

Variable	N	Mean	Std Dev	Sum
SBP	100	104.414700	9.430933	10441
Height	100	120.054000	6.439986	12005
Weight	100	22.826000	4.223303	2282.600000
Age	100	6.696900	0.731717	669.690000

Simple Statistics

Variable	Minimum	Maximum
SBP	81.500000	128.850000
Height	107.1000000	136.800000
Weight	15.900000	42.500000
Age	5.130000	8.840000

Summary statistics for each variable

Pearson Correlation Coefficients/Prob>|R| under Ho:Rho=0
/N=100

	SBP	Height	Weight	Age
SBP	1.00000	0.33066	0.51774	0.16373
	0.0	0.0008	0.0001	0.1036
Height	0.33066	1.00000	0.69151	0.64486
	0.0008	0.0	0.0001	0.0001
Weight	0.51774	0.69151	1.00000	0.38935
	0.0001	0.0001	0.0	0.0001
Age	0.16373	0.64486	0.38935	1.00000
	0.1036	0.0001	0.0001	0.0

Pearson's correlation coefficient between SBP and age

Associated P-value

Spearman Correlation Coefficients/Prob>|R| under Ho:Rho=0
/N=100

	SBP	Height	Weight	Age
SBP	1.00000	0.31519	0.45453	0.14778
	0.0	0.0014	0.0001	0.1423
Height	0.31519	1.00000	0.82298	0.61491
	0.0014	0.0	0.0001	0.0001
Weight	0.45453	0.82298	1.00000	0.51260
	0.0001	0.0001	0.0	0.0001
Age	0.14778	0.61491	0.51260	1.00000
	0.1423	0.0001	0.0001	0.0

Spearman's correlation coefficient between height and age

P-value

Topic 26

```
Model:MODEL1
Dependent Variable:SBP

                    Analysis of Variance

Source          DF         Sum of          Mean      F Value      Prob>F
                          Squares        Square

Model            1      962.71441     962.71441       12.030      0.0008
Error           98     7842.59208      80.02645
C Total         99     8805.30649

     Root MSE            8.94575      R-square        0.1093
     Dep Mean          104.41470      Adj R-sq        0.1002
         C.V.            8.56752

                  Parameter Estimates

                    Parameter     Standard      T for HO:
Variable   DF       Estimate        Error     Parameter=0

Intercep    1     46.281684   16.78450788          2.757
Height      1      0.484224    0.13960927          3.468

Variable DF       Prob>|T|

Intercep    1      0.0070
Height      1      0.0008
```

Intercept, a

Slope, b

Anova table

Results from simple linear regression of SBP (systolic blood pressure) on height Topic 28

```
Model:MODEL1
Dependent Variable:SBP

                    Analysis of Variance

Source          DF         Sum of          Mean      F Value      Prob>F
                          Squares        Square

Model            3     2804.04514     934.68171       14.952      0.0001
Error           96     6001.26135      62.51314
C Total         99     8805.30649

Root MSE                    7.90653      R-square        0.3184
Dep Mean                  104.41470      Adj R-sq        0.2972
C.V.                        7.57223

                  Parameter Estimates

 Variable   DF    Parameter     Standard      T for HO:
                   Estimate        Error     Parameter=0

Intercep    1    79.439541   17.11822110          4.641
Height      1    -0.031023    0.17170250         -0.181
Weight      1     1.179495    0.26139400          4.512
Sex         1     4.229540    1.61054848          2.626

Variable    DF    Prob>|T|

Intercep    1      0.0001
Height      1      0.8570
Weight      1      0.0001
Sex         1      0.0101
```

Estimated partial regression coefficients

Results from multiple linear regression of SBP on height, weight and gender Topic 29

Analysis of HHV-8 data described in Topics 23, 24 and 30, generated by STATA

```
.List hhv8 gonnorrho syphilis hsv2 hiv age in 1/10

        hhv8     gonorrho     syphilis   hsv2    hiv      age
 1.  negative    history      0          0       0        28
 2.  negative    history      0          0       0        40
 3.  negative    history      0          0       0        26
 4.  negative    history      0          1       0        42
 5.  positive    history      0          0       0        30
 6.  negative    nohistory    0          0       0        33
 7.  negative    history      0          1       0        27
 8.  positive    history      0          0       0        32
 9.  negative    history      1          0       0        35
10.  positive    history      0          0       0        35
```

Print out of data from first 10 men

```
.  Tabulate gonorrho hhv8, chi row col

           |      hhv8
  gonorrhoe|  negative |  positive |     Total
-----------+-----------+-----------+----------
   History |       192 |        36 |       228
   Row %   |     84.21 |     15.79 |    100.00
   Column % |    86.88 |     72.00 |     84.13
-----------+-----------+-----------+----------
No history |        29 |        14 |        43
           |     67.44 |     32.56 |    100.00
           |     13.12 |     28.00 |     15.87
-----------+-----------+-----------+----------
     Total |       221 |        50 |       271
           |     81.55 |     18.45 |    100.00
           |    100.00 |    100.00 |    100.00

          Pearson chi2(1) = 6.7609   Pr = 0.009
```

Contingency table; Row marginal total; Observed frequency; Column marginal total; Overall total; Topic 24

```
. Logit hhv8 gonorrho syphilis hsv2 hiv age, or tab

Interation 0:   Log Likelihood = -122.86506
Interation 1:   Log Likelihood = -111.87072
Interation 2:   Log Likelihood = -110.58712
Interation 3:   Log Likelihood = -110.56596
Interation 4:   Log Likelihood = -110.56595

Logit Estimates                           Number of obs =     260
                                          chi2 (5)      =   24.60
                                          Prob > chi2   =  0.0002
Logit Likelihood = -110.56595             Pseudo R2     =  0.1001

------------------------------------------------------------------------------
        hhv8 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   gonorrho |    .5093263   .4363219     1.167   0.243    -.345849    1.364502
   syphilis |    1.192442   .7110707     1.677   0.094    -.201231    2.586115
       hsv2 |    .7910041   .3871114     2.043   0.041     .0322798   1.549728
        hiv |    1.635669   .6028147     2.713   0.007     .4541736   2.817164
        age |    .0061609   .0204152     0.302   0.763    -.0338521    .046174
      _cons |   -2.224164   .6511603    -3.416   0.001    -3.500415   -.9479135
------------------------------------------------------------------------------

------------------------------------------------------------------------------
        hhv8 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   gonorrho |     1.66417   .7261137     1.167   0.243     .7076193   3.913772
   syphilis |    3.295118   2.343062     1.677   0.094     .8177235   13.27808
       hsv2 |     2.20561   .8538167     2.043   0.041     1.032806   4.710191
        hiv |    5.132889   3.094181     2.713   0.007     1.574871   16.72934
        age |     1.00618   .0205413     0.302   0.763     .9667145   1.047257
------------------------------------------------------------------------------

     Comparison of outcomes and probabilites

     Outcome |   Pr < .5  |  Pr > = .5 |     Total
-------------+------------+------------+----------
     Failure |        208 |          5 |       213
     Success |         38 |          9 |        47
-------------+------------+------------+----------
       Total |        246 |         14 |       260
```

Chi-square for covariates and its P-value; P-value; CI for odds ratio; Predicted outcome <0.5 = 0 (No) ≥0.5 = 1 (Yes); Observed outcome Failure = 0 (No) Success = 1 (Yes); Classification table

Results from multiple logistic regression Topic 30

Analysis of virological failure data described in Chapters 31–33, generated by SAS

Time since initial response (<1 yr = 1, 1 – 2 yrs = 2, >2 yrs = 3)

Virological failure (No = 0, Yes = 1)

Length of follow-up (days)

Female = 0 Male = 1

OBS	PATIENT	PERIOD	EVENT	PDAYS	SEX	BASECD8	TRTSTATUS
1	1	1	0	365.25	0	665	1
2	1	2	0	48.75	0	665	1
3	2	1	0	365.25	1	2053	1
4	2	2	0	365.25	1	2053	1
5	2	3	0	592.50	1	2053	1
7	4	1	0	30.00	0	327	1
8	5	1	0	365.25	1	931	1
9	5	2	0	365.25	1	931	1
10	5	3	0	732.50	1	931	1
13	6	1	0	166.00	1	1754	1
14	7	1	0	84.00	1	665	1
15	8	1	0	365.25	1	297	1
16	8	2	0	152.75	1	297	1
17	9	1	0	142.00	1	455	1
18	10	1	0	230.00	0	736	1

Treatment status (Previously received treatment = 0, No previous treatment = 1)

Print out of data from first 10 patients (each patient has a row of data for each time period)

The GENMOD Procedure
Model Information

Data Set	WORK.APPENDIX_POISSON
Distribution	Poisson
Link Function	Log
Dependent Variable	EVENT
Offset Variable	LTIME = Log(PDAYS)
Observations Used	988

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	984	393.1203	0.3995
Scaled Deviance	984	393.1203	0.3995
Pearson Chi-Square	984	7574.2725	7.6974
Scaled Pearson X2	984	7574.2725	7.6974
Log Likelihood		-257.5601	

Analysis Of Parameter Estimates

Baseline CD8 count divided by 100

Wald test statistics

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-1.1698	0.3228	-1.8024	-0.5372	13.14	0.0003
TRTSTATUS	1	-0.6096	0.2583	-1.1159	-0.1033	5.57	0.0183
BASECD8_100	1	-0.0587	0.0268	-0.1112	-0.0063	4.82	0.0281
SEX	1	-0.4923	0.2660	-1.0136	0.0290	3.43	0.0642
Scale	0	1.0000	0.0000	1.0000	1.0000		

Scale parameter used to adjust for extra-Poisson dispersion

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
TRTSTATUS	1	5.40	0.0201
BASECD8_100	1	5.46	0.0194
SEX	1	3.27	0.0707

P-value for significance of each variable in model

Model 1 excluding 2 dummies for time since initial response. Chapter 32

Model Information

Data Set	WORK.APPENDIX_POISSON
Distribution	Poisson
Link Function	Log
Dependent Variable	EVENT
Offset Variable	LTIME
Observations Used	988

LRS or deviance gives $P > 0.99$ for evaluating goodness of fit

Class Level Information

Class	Levels	Values
PERIOD3	3	1 2 3

Degrees of freedom

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	982	387.5904	0.3947
Scaled Deviance	982	387.5904	0.3947
Pearson Chi-Square	982	5890.6342	5.9986
Scaled Pearson X2	982	5890.6342	5.9986
Log Likelihood		-254.7952	

This is substantially <1, indicating underdispersion

Analysis Of Parameter Estimates

Estimates of model parameters shown in Table 31.1.—Relative rates obtained by antilogging estimates

CI for model coefficients

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-1.2855	0.3400	-1.9518	-0.6192	14.30	0.0002
TRTSTATUS		1	-0.5871	0.2587	-1.0942	-0.0800	5.15	0.0233
BASECD8_100		1	-0.0558	0.0267	-0.1083	-0.0034	4.36	0.0369
SEX		1	-0.4868	0.2664	-1.0089	0.0353	3.34	0.0676
PERIOD	1	0	0.0000	0.0000	0.0000	0.0000	.	.
PERIOD	2	1	0.4256	0.2702	-0.1039	0.9552	2.48	0.1152
PERIOD	3	1	-0.5835	0.4825	-1.5292	0.3622	1.46	0.2265
Scale		0	1.0000	0.0000	1.0000	1.0000		

Zeros in this row indicate that Period 1 is reference category

LR Statistics For Type 3 Analysis

Test statistic = difference in deviances of 2 models = 393.1203 − 387.5904

Source	DF	Chi-Square	Pr > ChiSq
TRTSTATUS	1	5.00	0.0253
BASECD8_100	1	4.91	0.0267
SEX	1	3.19	0.0742
PERIOD	2	5.53	0.0630

P-value for test of difference in deviancies from models with and without dummy variables for time since initial response

Model 2 including 2 dummies for time since initial response and CD8 count as a numerical variable. Chapters 31 and 32

Degrees of freedom = difference in number of parameters in Models 1 and 2

Model Information

Data Set	WORK.APPENDIX_POISSON
Distribution	Poisson
Link Function	Log
Dependent Variable	EVENT
Offset Variable	LTIME
Observations Used	988

Class Level Information

Class	Levels	Values
PERIOD	3	1 2 3

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	983	392.5001	0.3993
Scaled Deviance	983	392.5001	0.3993
Pearson Chi-Square	983	5580.2152	5.6767
Scaled Pearson X2	983	5580.2152	5.6767
Log Likelihood		-257.2501	

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-1.7549	0.2713	-2.2866	-1.2232	41.85	<.0001
TRTSTATUS		1	-0.6290	0.2577	-1.1340	-0.1240	5.96	0.0146
SEX		1	-0.5444	0.2649	-1.0637	-0.0252	4.22	0.0399
PERIOD	1	0	0.0000	0.0000	0.0000	0.0000	.	.
PERIOD	2	1	0.4191	0.2701	-0.1103	0.9485	2.41	0.1207
PERIOD	3	1	-0.6481	0.4814	-1.5918	0.2955	1.81	0.1782
Scale		0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
TRTSTATUS	1	5.77	0.0163
SEX	1	4.00	0.0455
PERIOD	2	6.08	0.0478

Model excluding baseline CD8 count. Chapter 33

Model Information

Data Set	WORK.APPENDIX_POISSON
Distribution	Poisson
Link Function	Log
Dependent Variable	EVENT
Offset Variable	LTIME
Observations Used	988

Class Level Information

Class	Levels	Values
PERIOD	3	1 2 3
CD8GRP	5	1 2 3 4 5

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	979	387.1458	0.3955
Scaled Deviance	979	387.1458	0.3955
Pearson Chi-Square	979	5852.1596	5.9777
Scaled Pearson X2	979	5852.1596	5.9777
Log Likelihood		-254.5729	

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-1.2451	0.6116	-2.4439	-0.0463	4.14	0.0418
TRTSTATUS		1	-0.5580	0.2600	-1.0677	-0.0483	4.60	0.0319
SEX		1	-0.4971	0.2675	-1.0214	0.0272	3.45	0.0631
PERIOD	1	0	0.0000	0.0000	0.0000	0.0000	.	.
PERIOD	2	1	0.4550	0.2715	-0.0771	0.9871	2.81	0.0937
PERIOD	3	1	-0.5386	0.4849	-1.4890	0.4119	1.23	0.2667
CD8GRP	1	1	-0.2150	0.6221	-1.4343	1.0044	0.12	0.7297
CD8GRP	2	1	-0.3646	0.7648	-1.8636	1.1345	0.23	0.6336
CD8GRP	3	0	0.0000	0.0000	0.0000	0.0000	.	.
CD8GRP	4	1	-0.3270	1.1595	-2.5996	1.9455	0.08	0.7779
CD8GRP	5	1	-0.8264	0.6057	-2.0136	0.3608	1.86	0.1725
Scale		0	1.0000	0.0000	1.0000	1.0000		

Model including baseline CD8 count as a series of dummy variables. Chapter 33

Parameter estimates for dummy variables for baseline CD8 count where category 3 (≥25, <1100) is reference category

LR Statistics For Type 3 Analysis

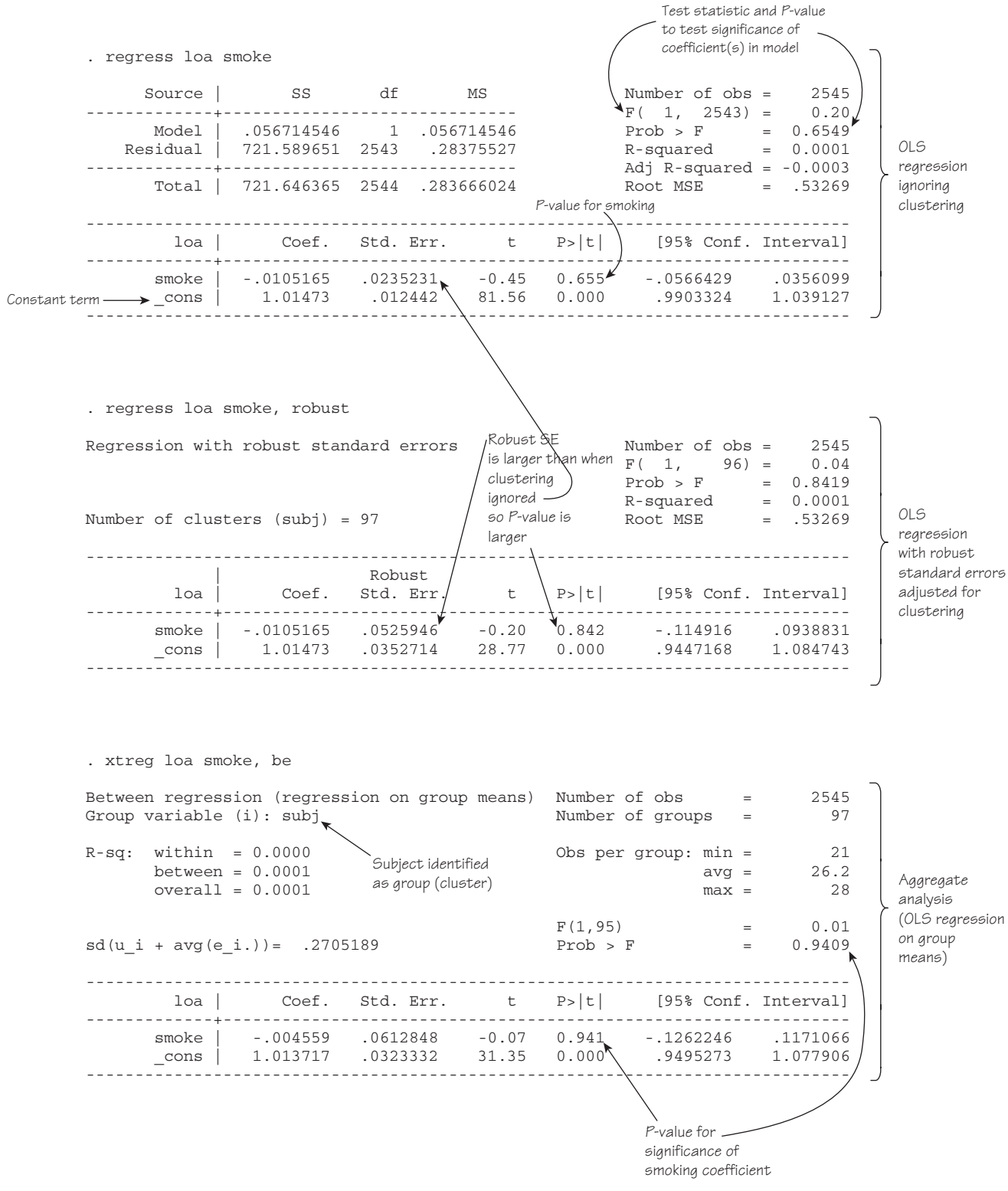
Source	DF	Chi-Square	Pr > ChiSq
TRTSTATUS	1	4.48	0.0342
SEX	1	3.30	0.0695
PERIOD	2	5.54	0.0628
CD8GRP	4	5.35	0.2528

Number of additional variables in larger model

Test statistic = 392.5001 – 387.1458

P-value for test of significance of baseline CD8 count when incorporated as a categorical variable

Analysis of periodontal data used in Chapter 42, generated by Stata



```
. iis subj
. xtreg loa smoke, pa robust corr(exchangeable)

Iteration 1: tolerance = .00516018
Iteration 2: tolerance = 2.204e-07

GEE population-averaged model                   Number of obs      =      2545
Group variable:                      subj       Number of groups   =        97
Link:                              identity     Obs per group: min =        21
Family:                            Gaussian                  avg =      26.2
Correlation:                   exchangeable                  max =        28
                                                Wald chi2(1)       =      0.01
Scale parameter:                   .2835381     Prob > chi2        =    0.9198

                                 (standard errors adjusted for clustering on subj)
------------------------------------------------------------------------------
             |               Semi-robust
         loa |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |  -.0053018   .0526501    -0.10   0.920    -.1084941    .0978905
       _cons |   1.013841   .0347063    29.21   0.000     .9458185    1.081865
------------------------------------------------------------------------------
```

Correlation structure identified as exchangeable

Model chi-squared to test significance of coefficient in model

P-value for model chi-squared

GEE with robust standard errors and exchangeable correlation structure

Wald test statistic

```
. xtreg loa smoke, mle

Fitting constant-only model:
Iteration 0:   log likelihood = -1785.7026
Iteration 1:   log likelihood = -1785.7004

Fitting full model:
Iteration 0:   log likelihood = -1785.7027
Iteration 1:   log likelihood = -1785.6966
Iteration 2:   log likelihood = -1785.6966

Random-effects ML regression                    Number of obs      =      2545
Group variable (i): subj                        Number of groups   =        97

Random effects u_i ~ Gaussian                   Obs per group: min =        21
                                                               avg =      26.2
                                                               max =        28
                                                LR chi2(1)         =      0.01
Log likelihood  = -1785.6966                    Prob > chi2        =    0.9302

------------------------------------------------------------------------------
         loa |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |  -.0053168   .0607203    -0.09   0.930    -.1243265    .1136928
       _cons |   1.013844    .032046    31.64   0.000      .951035    1.076653
-------------+----------------------------------------------------------------
     /sigma_u |   .2519226   .0204583    12.31   0.000     .2118251    .2920201
     /sigma_e |   .4684954   .0066952    69.98   0.000     .4553731    .4816176
-------------+----------------------------------------------------------------
         rho |   .2242953   .0288039                      .1719879    .2846119
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)=  443.21 Prob>=chibar2 = 0.000
```

difference = −0.0038 so
− 2 log likelihood ratio
= 2 × 0.0038
= 0.0076
≈ 0.01

final iteration provides stable estimates

Degrees of freedom

LRS = −2 log likelihood ratio

P-value

Random effects model

σ_c → /sigma_u

σ → /sigma_e

intracluster correlation coefficient

$$= \frac{0.2519226^2}{0.2519226^2 + 0.4684954^2}$$

Appendix D: Glossary of terms

2 × 2 table: A contingency table of frequencies with two rows and two columns

–2 log likelihood: See likelihood ratio statistic

Accuracy: Refers to the way in which an observed value of a quantity agrees with the true value

All subsets model selection: See model selection

Allocation bias: A systematic distortion of the data resulting from the way in which individuals are assigned to treatment groups

Alternative hypothesis: The hypothesis about the effect of interest that disagrees with the null hypothesis and is true if the null hypothesis is false

Altman's nomogram: A diagram that relates the sample size of a statistical test to the power, significance level and the standardized difference

Analysis of covariance: A special form of analysis of variance that compares values of a dependent variable between groups of individuals after adjusting for the effect of one or more explanatory variables

Analysis of variance (ANOVA): A general term for analyses that compare means of groups of observations by splitting the total variance of a variable into its component parts, each attributed to a particular factor

ANOVA : See analysis of variance

Arithmetic mean: A measure of location obtained by dividing the sum of the observations by the number of observations. Often called the mean

ASCII or text file format: Data are available on the computer as rows of text

Automatic model selection: A method of selecting variables to be included in a mathematical model, e.g. forwards, backwards, stepwise, all subsets

Average: A general term for a measure of location

Backwards selection: See model selection

Bar or column chart: A diagram that illustrates the distribution of a categorical or discrete variable by showing a separate horizontal or vertical bar for each 'category', its length being proportional to the (relative) frequency in that 'category'

Bartlett's test: Used to compare variances

Bayes theorem: The posterior probability of an event/hypothesis is proportional to the product of its prior probability and the likelihood

Bayesian approach to inference: Uses not only current information (e.g. from a trial) but also an individual's previous belief (often subjective) about a hypothesis to evaluate the posterior belief in the hypothesis

Bias: A systematic difference between the results obtained from a study and the true state of affairs

Bimodal distribution: Data whose distribution has two 'peaks'

Binary variable: A categorical variable with two categories. Also called a dichotomous variable

Binomial distribution: A discrete probability distribution of a binary random variable; useful for inferences about proportions

Blinding: When the patients, clinicians and the assessors of response to treatment in a clinical trial are unaware of the treatment allocation (double-blind), or when the patient is aware of the treatment received but the assessor of response is not (single blind). Also called masking

Block: A homogeneous group of experimental units that share similar characteristics. Also called a stratum

Bonferroni correction (adjustment): A *post hoc* adjustment to the *P*-value to take account of the number of tests performed in multiple hypothesis testing

Bootstrapping: Simulation process used to derive a confidence interval for a parameter. It involves estimating the parameter from each of many random samples obtained by sampling with replacement from the original sample; the CI is derived by considering the variability of the distribution of these estimates

Box (box-and-whisker) plot: A diagram illustrating the distribution of a variable; it indicates the median, upper and lower quartiles, and, often, the maximum and minimum values

British Standards Institution repeatability coefficient: The maximum difference that is likely to occur between two repeated measurements

C statistic: Measures the area under a ROC curve and may be used to compare diagnostic tests for the same condition

Carry-over effect: The residual effect of the previous treatment in a cross-over trial

Case: An individual with the disease under investigation in a case–control study

Case–control study: Groups of individuals with the disease (the cases) and without the disease (the controls) are identified, and exposures to risk factors in these groups are compared

Categorical (qualitative) variable: Each individual belongs to one of a number of distinct categories of the variable

Cell of a contingency table: The designation of a particular row and a particular column of the table

Censored data: Occur in survival analysis because there is incomplete information on outcome. See right- and left-censored data

Chi-squared (χ^2) distribution: A right skewed continuous distribution characterized by its degrees of freedom; useful for analysing categorical data

Chi-squared test: Used on frequency data. It tests the null hypothesis that there is no association between the factors that define a contingency table. Also used to test differences in proportions

CI: See confidence interval

Clinical cohort: A group of patients with the same clinical condition whose outcomes are observed over time

Clinical heterogeneity: Exists when the trials included in a meta-analysis have differences in the patient population, definition of variables, etc., which create problems of non-compatibility

Clinical trial: Any form of planned experiment on humans that is used to evaluate a new treatment on a clinical outcome

Cluster randomized trial: A study in which groups (clusters) of individuals are randomized to different 'treatments' so that every individual within a particular cluster receives the same treatment

Cluster randomization: Groups of individuals, rather than separate individuals, are randomly (by chance) allocated to treatments

Cochrane Collaboration: An international network of clinicians, methodologists and consumers who continuously update systematic reviews and make them available to others

Coefficient of variation: The standard deviation divided by the mean (often expressed as a percentage)

Cohen's kappa (κ): A measure of agreement between two sets of categorical measurements on the same individuals. If $\kappa = 1$ there is perfect agreement; if $\kappa = 0$, there is no better than chance agreement

Cohort study: A group of individuals, all without the outcome of interest (e.g. disease), is followed (usually prospectively) to study the effect on future outcomes of exposure to a risk factor

Collinearity: Pairs of explanatory variables in a regression analysis are very highly correlated, i.e. with correlation coefficients very close to ± 1

Complete randomized design: Experimental units assigned randomly to treatment groups

Conditional logistic regression: A form of logistic regression used when individuals in a study are matched

Conditional probability: The probability of an event, *given* that another event has occurred

Confidence interval (CI) for a parameter: The range of values within which we are (usually) 95% confident that the true population parameter lies. Strictly, after repeated sampling, 95% of the estimates of the parameter lie in the interval

Confidence limits: The upper and lower values of a confidence interval

Confounding: When one or more explanatory variables are related to the outcome and each other so that it is difficult to assess the independent effect of each one on the outcome variable.

CONSORT statement: Facilitates critical appraisal and interpretation of RCTs by providing guidance, in the form of a checklist and flowchart, to authors about how to report their trials

Contingency table: A (usually) two-way table in which the entries are frequencies

Continuity correction: A correction applied to a test statistic to adjust for the approximation of a discrete distribution by a continuous distribution

Continuous probability distribution: The random variable defining the distribution is continuous

Continuous variable: A numerical variable in which there is no limitation on the values that the variable can take other than that restricted by the degree of accuracy of the measuring technique

Control group: A term used in comparative studies, e.g. clinical trials, to denote a comparison group. See also positive and negative controls

Control: An individual without the disease under investigation in a case–control study, or not receiving the new treatment in a clinical trial

Convenience sample: A group of individuals believed to be representative of the population from which it is selected, but chosen because it is close at hand rather than being randomly selected

Correlation coefficient (Pearson's): A quantitative measure, ranging from -1 to $+1$, of the extent to which points in a scatter diagram conform to a straight line. See also Spearman's rank correlation coefficient

Covariate: See independent variable

Cox proportional hazards regression model: See proportional hazards regression model

Cross-over design: Each individual receives more than one treatment under investigation, one after the other in random order

Cross-sectional studies: Those that are carried out at a single point in time

Cumulative frequency: The number of individuals who have values below and including the specified value of a variable

Data: Observations on one or more variables

Deciles: Those values that divide the ordered observations into 10 equal parts

Degrees of freedom (*df*) of a statistic: the sample size minus the number of parameters that have to be estimated to calculate the statistic—they indicate the extent to which the observations are 'free' to vary

Dependent variable: A variable (usually denoted by y) that is predicted by the explanatory variable in regression analysis. Also called the response or outcome variable

Deviance: See likelihood ratio statistic

***df*:** See degrees of freedom

Diagnostic test: Used to aid or make a diagnosis of a particular condition

Dichotomous variable: See binary variable

Discrete probability distribution: The random variable defining the distribution takes discrete values

Discrete variable: A numerical variable that can only take integer values

Discriminant analysis: A method, similar to logistic regression, which can be used to identify factors that are significantly associated with a binary response

Distribution-free tests: See non-parametric tests

Dot plot: A diagram in which each observation on a variable is represented by one dot on a horizontal (or vertical) line

Double-blind: See blinding

Dummy variables: The $k-1$ binary variables that are created from a nominal or ordinal categorical variable with $k > 2$ categories, affording a comparison of each of the $k-1$ categories with a reference category in a regression analysis. Also called indicator variables

Effect modification: See interaction

Effect of interest: The value of the response variable that reflects the comparison of interest, e.g. the difference in means

Empirical distribution: The observed distribution of a variable

Epidemiological studies: Observational studies that assess the relationship between risk factors and disease

Equivalence trial: Used to show that two treatments are clinically equivalent

Error variation: See residual variation

Estimate: A quantity obtained from a sample that is used to represent a population parameter

Evidence-based medicine (EBM): The use of current best evidence in making decisions about the care of individual patients

Exchangeable model: Assumes the estimation procedure is not affected if two observations within a cluster are interchanged

Extra-Binomial variation: Occurs when the residual variance is greater (overdispersion) or less (underdispersion) than that expected in a Binomial model

Extra-Poisson variation: Occurs when the residual variance is greater (overdispersion) or less (underdispersion) than that expected in a Poisson model

Expected frequency: The frequency that is expected under the null hypothesis

Experimental study: The investigator intervenes in some way to affect the outcome

Experimental unit: The smallest group of individuals who can be regarded as independent for analysis purposes

Explanatory variable: A variable (usually denoted by x) that is used to predict the dependent variable in a regression analysis. Also called the independent or predictor variable or a covariate

Factorial experiment: Allows the simultaneous analysis of a number of factors of interest

Fagan's nomogram: A diagram relating the pre-test probability of a diagnostic test result to the likelihood and the post-test probability. It is usually used to convert the former into the latter

False negative: An individual who has the disease but is diagnosed as disease-free

False positive: An individual who is free of the disease but is diagnosed as having the disease

***F*-distribution:** A right skewed continuous distribution characterized by the degrees of freedom of the numerator and denominator of the ratio that defines it; useful for comparing two variances, and more than two means using the analysis of variance

Fisher's exact test: A test that evaluates exact probabilities (i.e. does not rely on approximations to the Chi-squared distribution) in a contingency table (usually a 2×2 table), used when the expected frequencies are small

Fitted value: The predicted value of the response variable in a regression analysis corresponding to the particular value(s) of the explanatory variable(s)

Fixed effect: One where the levels of the factor make up the entire population of interest (e.g. the factor 'treatment' whose levels are drug, surgery and radiotherapy). It contrasts with a random effect where the levels represent only a sample from the population (e.g. the factor 'patient' whose levels are the 20 patients in a RCT)

Fixed effect model: Contains only fixed effects; used in a meta-analysis when there is no evidence of statistical heterogeneity

Follow-up: The time that an individual is in a study, from entry until s/he experiences the outcome (e.g. develops the disease) or leaves the study or until the conclusion of the study

Forest plot: A diagram used in a meta-analysis showing the estimated effect in each trial and their average (with confidence intervals)

Forwards selection: See model selection

Free format data: Each variable in the computer file is separated from the next by some delimiter, often a space or comma

Frequency distribution: Shows the frequency of occurrence of each possible observation, class of observations, or category, as appropriate

Frequency: The number of times an event occurs

Frequentist probability: Proportion of times an event would occur if we were to repeat the experiment a large number of times

***F*-test:** See variance ratio test

Gaussian distribution: See Normal distribution

GEE: See generalized estimating equation

Generalized estimating equation (GEE): Used in a two-level hierarchical structure to estimate parameters and their standard errors to take into account the clustering of the data without referring to a parametric model for the random effects; sometimes referred to as population averaged or marginal.

Generalized linear model (GLM): A regression model which is expressed in a general form via a link function which relates the mean value of the dependent variable (with a known probability distribution such as Normal, Binomial or Poisson) to a linear function of covariates.

Geometric mean: A measure of location for data whose distribution is skewed to the right; it is the antilog of the arithmetic mean of the log data

GLM: See generalized linear model

Gold-standard test: Provides a definitive diagnosis of a particular condition

Goodness of fit: A measure of the extent to which the values obtained from a model agree with the observed data

Hazard ratio: See relative hazard

Hazard: The instantaneous risk of reaching the endpoint in survival analysis

Healthy entrant effect: By choosing disease-free individuals to participate in a study, the response of interest (typically, mortality) is lower at the start of the study than would be expected in the general population

Heterogeneity of variance: Unequal variances

Hierarchical model: See multilevel model

Histogram: A diagram that illustrates the (relative) frequency distribution of a continuous variable by using connected bars. The bar's area is proportional to the (relative) frequency in the range specified by the boundaries of the bar

Historical controls: Individuals who are not assigned to a treatment group at the start of the study, but who received treatment some time in the past, and are used as a comparison group

Homoscedasticity: Equal variances; also described as homogeneity of variance

Hypothesis test: The process of using a sample to assess how much evidence there is against a null hypothesis about the population. Also called a significance test

I^2: An index which can be used to quantify the impact of statistical heterogeneity between the studies in a meta-analysis

ICC: See intraclass correlation coefficient

Incidence: The number of new cases of a disease in a defined period divided by the number of individuals susceptible at the start or mid-point of the period

Incidence rate: The number of new cases of a disease in a defined period divided by the person-years of follow-up of individuals susceptible at the start of the period

Incidence rate ratio: A relative rate defined as the ratio of two incidence rates

Incident cases: Patients who have just been diagnosed

Independent samples: Each unit in every sample is unrelated to the units in the other samples

Independent variable: See explanatory variable

Indicator variables: See dummy variables

Inference: The process of drawing conclusions about the population using sample data

Influential point: An observation, which, if omitted from a regression analysis, will lead to a change in one or more of the parameter estimates

Intention-to-treat analysis: All patients in the clinical trial are analysed in the groups to which they were originally assigned

Interaction: Occurs between two explanatory variables in a regression analysis when the effect of one of the variables on the depen-

dent variable varies according to the level of the other. In the context of ANOVA, an interaction exists between two factors when the difference between the levels of one factor is different for two or more levels of the second factor. Also called effect modification

Intercept: The value of the dependent variable in a regression equation when the value(s) of the explanatory variable(s) is (are) zero

Interdecile range: The difference between the 10th and 90th percentiles; it contains the central 80% of the ordered observations

Interim analyses: Pre-planned analyses at intermediate stages of a study

Intermediate variable: A variable that lies on the causal pathway between an explanatory variable and an outcome of interest

Interpolate: Estimate the required value that lies between two known values

Interquartile range: The difference between the 25th and 75th percentiles; it contains the central 50% of the ordered observations

Interval estimate: A range of values within which we believe the population parameter lies.

Intraclass correlation coefficient (ICC): In a two-level structure, it expresses the variation between clusters as a proportion of the total variation; it represents the correlation between any two randomly chosen level 1 units in one randomly chosen cluster

Jackknifing: A method of estimating parameters and confidence intervals; each of n individuals is successively removed from the sample, the parameters are estimated from the remaining $n - 1$ individuals, and finally the estimates of each parameter are averaged

Kaplan–Meier plot: A survival curve in which the survival probability is plotted against the time from baseline. It is used when exact times to reach the endpoint are known

Kolmogorov–Smirnov test: Determines whether data are Normally distributed

Kruskal–Wallis test: A non-parametric alternative to the one-way ANOVA; used to compare the distributions of more than two independent groups of observations

Left-censored data: Come from patients in whom follow-up did not begin until after the baseline date

Lehr's formulae: Can be used to calculate the optimal sample sizes required for some hypothesis tests when the power is specified as 80% or 90% and the significance level as 0.05

Level one unit: The 'individual' at the lowest level of a hierarchical structure; a group of level one units (e.g. patients) comprise a cluster of individuals which is nested within a level two unit (e.g. ward)

Level two unit: The 'individual' at the second lowest level in a hierarchical structure; each level two unit (e.g. ward) comprises a cluster of level one units (e.g. patients)

Level: A particular category of a qualitative variable or factor

Levene's test: Tests the null hypothesis that two or more variances are equal

Leverage: a measure of the extent to which the value of the explanatory variable(s) for an individual differs from the mean of the explanatory variable(s) in a regression analysis

Lifetable approach to survival analysis: A way of determining survival probabilities when the time to reach the end-point is only known to within a particular time interval

Likelihood: The probability of the data, given the model. In the context of a diagnostic test, it describes the plausibility of the observed test result if the disease is present (or absent)

Likelihood ratio (LR): A ratio of two likelihoods; for diagnostic tests, the LR is the ratio of the chances of getting a particular test result in those having and not having the disease

Likelihood ratio statistic (LRS): Equal to −2 times the ratio of the log likelihood of a saturated model to that of the model of interest. It is used to assess adequacy of fit and may be called the deviance or, commonly, −2 log likelihood. The difference in the LRS in two nested models can be used to compare the models

Likelihood ratio test: Uses the likelihood ratio statistic to compare the fit of two regression models or to test the significance of one or a set of parameters in a regression model

Limits of agreement: In an assessment of repeatability, it is the range of values between which we expect 95% of the differences between repeated measurements in the population to lie

Linear regression line: The straight line that is defined by an algebraic expression linking two variables

Linear relationship: Implies a straight line relationship between two variables

Link function: In a generalized linear model, it is a transformation of the mean value of the dependent variable which is modelled as a linear combination of the covariates

Logistic regression coefficient: The partial regression coefficient in a logistic regression equation

Logistic regression: A form of generalized linear model used to relate one or more explanatory variables to the logit of the expected proportion of individuals with a particular outcome when the response is binary

Logit (logistic) transformation: A transformation applied to a proportion or probability, p, such that $\text{logit}(p) = \ln\{p/(1-p)\} = \ln(\text{odds})$

Lognormal distribution: A right skewed probability distribution of a random variable whose logarithm follows the Normal distribution

Log-rank test: A non-parametric approach to comparing two survival curves

Longitudinal study: Follows individuals over a period of time

LRS: See likelihood ratio statistic

Main outcome variable: That which relates to the major objective of the study

Mann–Whitney *U* test: See Wilcoxon rank sum test

Marginal model: See generalized estimating equation

Marginal total in a contingency table: The sum of the frequencies in a given row (or column) of the table

Masking: See blinding

Matching: A process of selecting individuals who are similar with respect to variables that may influence the response of interest

Maximum likelihood estimation (MLE): An iterative process of estimation of a parameter which maximizes the likelihood

McNemar's test: Compares proportions in two related groups using a Chi-squared test statistic

Mean: See arithmetic mean

Median: A measure of location that is the middle value of the ordered observations

Meta-analysis (overview): A quantitative systematic review that combines the results of relevant studies to produce, and investigate, an estimate of the overall effect of interest

Method of least squares: A method of estimating the parameters in a regression analysis, based on minimizing the sum of the squared residuals

Mixed model: Some of the parameters in the model have random effects and others have fixed effects

MLE: See maximum likelihood estimation

Mode: The value of a single variable that occurs most frequently in a data set

Model Chi-squared test: Usually refers to a hypothesis test in a regression analysis that tests the null hypothesis that all the parameters associated with the covariates are zero; it is based on the difference in two likelihood ratio statistics

Model sensitivity: The extent to which estimates in a regression model are affected by one or more individuals in the data set or mis-specification of the model

Model: Describes, in algebraic terms, the relationship between two or more variables

Mortality rate: The death rate

Multilevel model: Used for the analysis of hierarchical data in which level one units (e.g. patients) are nested within level two units (e.g. wards) which may be nested within level three units (e.g. hospitals), etc. Also called hierarchical model

Multinomial logistic regression: A form of logistic regression used when the nominal outcome variable has more than two categories. Also called polychotomous logistic regression

Multiple linear regression: A linear regression model in which there is a single numerical dependent variable and two or more explanatory variables

Multivariable regression model: Any regression model that has a single outcome variable and two or more explanatory variables

Multivariate regression model: Has two or more outcome variables and two or more explanatory variables

Mutually exclusive categories: Each individual can belong to only one category

Negative controls: Those patients in a randomized controlled trial (RCT) who do not receive active treatment

Negative predictive value: The proportion of individuals with a negative test result who do not have the disease

Nested models: Two regression models, the larger of which includes the covariates in the smaller model, plus additional covariate(s)

Nominal variable: A categorical variable whose categories have no natural ordering

Non-inferiority trial: Used to demonstrate that a given treatment is clinically not inferior to another

Non-parametric tests: Hypothesis tests that do not make assumptions about the distribution of the data. Sometimes called distribution-free tests or rank methods

Normal (Gaussian) distribution: A continuous probability distribution that is bell-shaped and symmetrical; its parameters are the mean and variance

Normal plot: A diagram for assessing, visually, the Normality of data; a straight line on the Normal plot implies Normality

Normal range: See reference interval

Null hypothesis, H_0: The statement that assumes no effect in the population

Number of patients needed to treat (NNT): The number of patients we need to treat with the experimental rather than the control treatment to prevent one of them developing the 'bad' outcome

Numerical (quantitative) variable: A variable that takes either discrete or continuous values

Observational study: The investigator does nothing to affect the outcome

Odds ratio: The ratio of two odds (e.g. the odds of disease in individuals exposed and unexposed to a factor). Often taken as an estimate of the relative risk in a case–control study

Odds: The ratio of the probabilities of two complimentary events, typically the probability of having a disease divided by the probability of not having the disease

Offset: An explanatory variable whose regression coefficient is fixed at unity in a generalized linear model. It is the log of the total person-years (or months/days, etc) of follow-up in a Poisson model when the dependent variable is defined as the number of events occurring instead of a rate

One-sample *t*-test: Investigates whether the mean of a variable differs from some hypothesized value

One-tailed test: The alternative hypothesis specifies the direction of the effect of interest

One-way analysis of variance: A particular form of ANOVA used to compare the means of more than two independent groups of observations

On-treatment analysis: Patients in a clinical trial are only included in the analysis if they complete a full course of the treatment to which they were randomly assigned

Ordinal logistic regression: A form of logistic regression used when the ordinal outcome variable has more than two categories

Ordinal variable: A categorical variable whose categories are ordered in some way

Outlier: An observation that is distinct from the main body of the data and is incompatible with the rest of the data

Overdispersion: Occurs when the residual variance is greater than that expected by the defined regression model (e.g. Binomial or Poisson)

Overfitted model: A model containing too many variables, e.g. more than 1/10th of the number of individuals in a multiple linear regression model

Overview: See meta-analysis

Paired observations: Relate to responses from matched individuals or the same individual in two different circumstances

Paired *t*-test: Tests the null hypothesis that the mean of a set of differences of paired observations is equal to zero

Parallel trial: Each patient receives only one treatment

Parameter: A summary measure (e.g. the mean, proportion) that characterizes a probability distribution. Its value relates to the population

Parametric test: Hypothesis test that makes certain distributional assumptions about the data

Partial regression coefficients: The parameters, other than the intercept, which describe a multivariable regression model

Pearson's correlation coefficient: See correlation coefficient

Percentage point: The percentile of a distribution; it indicates the proportion of the distribution that lies to its right (i.e. in the right hand tail), to its left (i.e. in the left-hand tail), or in both the right- and left-hand tails

Percentiles: Those values that divide the ordered observations into 100 equal parts

Person-years of follow-up: The sum, over all individuals, of the number of years that each individual is followed-up in a study.

Pie chart: A diagram showing the frequency distribution of a categorical or discrete variable. A circular 'pie' is split into sections,

one for each 'category'; the area of each section is proportional to the frequency in that category

Pilot study: Small scale preliminary investigation

Placebo: An inert 'treatment', identical in appearance to the active treatment that is compared to the active treatment in a negatively controlled clinical trial to assess the therapeutic effect of the active treatment by separating from it the effect of receiving treatment; also used to accommodate blinding

Point estimate: A single value, obtained from a sample, which estimates a population parameter

Point prevalence: The number of individuals with a disease (or percentage of those susceptible) at a particular point in time

Poisson distribution: A discrete probability distribution of a random variable representing the number of events occurring randomly and independently at a fixed average rate

Poisson regression model: A form of generalized linear model used to relate one or more explanatory variables to the log of the expected rate of an event (e.g. of disease) when the follow-up of the individuals varies but the rate is assumed constant over the study period.

Polynomial regression: A non-linear (e.g. quadratic, cubic, quartic) relationship between a dependent variable and one or more explanatory variables

Population averaged model: See genereralized estimating equation

Population: The entire group of individuals in whom we are interested

Positive controls: Those patients in a RCT who receive some form of active treatment as a basis of comparison for the novel treatment

Positive predictive value: The proportion of individuals with a positive diagnostic test result who have the disease

Posterior probability: An individual's belief, based on prior belief and new information (e.g. a test result), that an event will occur

***Post-hoc* comparison adjustments:** Are made to adjust the *P*-values when multiple comparisons are performed, e.g. Bonferroni

Post-test probability: The posterior probability, determined from previous information and the diagnostic test result, that an individual has a disease

Power: The probability of rejecting the null hypothesis when it is false

Precision: A measure of sampling error. Refers to how well repeated observations agree with one another

Predictor variable: See independent variable

Pre-test probability: The prior probability, evaluated before a diagnostic test result is available, that an individual has a disease

Prevalence: The number (proportion) of individuals with a disease at a given point in time (point prevalence) or within a defined interval (period prevalence)

Prevalent cases: Patients who have the disease at a given point in time or within a defined interval but who were diagnosed at a previous time.

Primary endpoint: The outcome that most accurately reflects the benefit of a new therapy in a clinical trial

Prior probability: An individual's belief, based on subjective views and/or retrospective observations, that an event will occur

Probability density function: The equation that defines a probability distribution

Probability distribution: A theoretical distribution that is described by a mathematical model. It shows the probabilities of all possible values of a random variable

Probability: Measures the chance of an event occurring. It ranges from 0 to 1. See also conditional, prior and posterior probability

Prognostic index: Assesses the likelihood that an individual has a disease. Also called a risk score

Proportion: The ratio of the number of events of interest to the total number in the sample or population

Proportional hazards regression model (Cox): Used in survival analysis to study the simultaneous effect of a number of explanatory variables on survival

Prospective study: Individuals are followed forward from some point in time

Protocol deviations: The patients who enter a clinical trial but do not fulfill the protocol criteria

Protocol: A full written description of all aspects of a clinical trial

Publication bias: A tendency for journals to publish only papers that contain statistically significant results

***P*-value:** The probability of obtaining our results, or something more extreme, if the null hypothesis is true

Qualitative variable: See categorical variable

Quantitative variable: See numerical variable

Quartiles: Those values that divide the ordered observations into four equal parts

Quota sampling: Non-random sampling in which the investigator chooses sample members to fulfil a specified 'quota'

R^2: The proportion of the total variation in the dependent variable in a simple or multiple regression analysis that is explained by the model. It is a subjective measure of goodness of fit

Random effect: The effect of a factor whose levels are assumed to represent a random sample from the population

Random effects model: A model for a hierarchical data structure, such as a two-level structure with level 1 units nested in level 2 units, in which a random effect is the source of error attributable to the level 2 units

Random intercepts model: A random effects hierarchical model which assumes, for the two-level structure, that the linear relationship between the mean value of the dependent variable and a single covariate for every level two unit has the same slope for all level two units and an intercept that varies randomly about the mean intercept

Random sampling: Every possible sample of a given size in the population has an equal probability of being chosen

Random slopes model: A random effects hierarchical model which assumes, for the two-level structure, that the linear relationship between the mean value of the dependent variable and a single covariate for each level two unit has a slope that varies randomly about the mean slope and an intercept that varies randomly about the mean intercept

Random variable: A quantity that can take any one of a set of mutually exclusive values with a given probability

Random variation: Variability that cannot be attributed to any explained sources

Randomization: Patients are allocated to treatment groups in a random (based on chance) manner. May be *stratified* (controlling for the effect of important factors) or *blocked* (ensuring approximately equally sized treatment groups)

Randomized controlled trial (RCT): A comparative clinical trial in which there is random allocation of patients to treatments

Range: The difference between the smallest and largest observations

Rank correlation coefficient: See Spearman's rank correlation coefficient

Rank methods: See non-parametric tests

Rate: The number of events occurring expressed as a proportion of the total follow-up time of all individuals in the study

RCT: See randomized controlled trial

Recall bias: A systematic distortion of the data resulting from the way in which individuals remember past events

Receiver operating characteristic (ROC) curve: A two-way plot of the sensitivity against one minus the specificity for different cut-off values for a continuous variable in a diagnostic test; used to select the optimal cut-off value or to compare tests

Reference interval: The range of values (usually the central 95%) of a variable that are typically seen in healthy individuals. Also called the normal or reference range

Regression coefficients: The parameters (i.e. the slope and intercept in simple regression) that describe a regression equation

Regression to the mean: A phenomenon whereby a subset of extreme results is followed by results that are less extreme on average e.g. tall fathers having shorter (but still tall) sons

Relative frequency: The frequency expressed as a percentage or proportion of the total frequency

Relative hazard: The ratio of two hazards, interpreted in a similar way to the relative risk. Also called the hazard ratio

Relative rate: The ratio of two rates (often the rate of disease in those exposed to a factor divided by the disease rate in those unexposed to the factor).

Relative risk (RR): The ratio of two risks, usually the risk of a disease in a group of individuals exposed to some factor divided by the risk in unexposed individuals

Reliability: A general term which encompasses repeatability, reproducibility and agreement

Repeatability: The extent to which repeated measurements by the same observer in identical conditions agree

Repeated measures ANOVA: A special form of analysis of variance used when a numerical variable is measured in each member of a group of individuals more than once (e.g. on different occasions)

Repeated measures: The variable of interest is measured on the same individual in more than one set of circumstances (e.g. on different occasions)

Replication: The individual has more than one measurement of the variable on a given occasion

Reproducibility: The extent to which the same results can be obtained in different circumstances, e.g. by two methods of measurement, or by two observers

Residual variation: The variance of a variable that remains after the variability attributable to factors of interest has been removed. It is the variance unexplained by the model, and is the residual mean square in an ANOVA table. Also called the error or unexplained variation

Residual: The difference between the observed and fitted values of the dependent variable in a regression analysis

Response variable: See dependent variable

Retrospective studies: Individuals are selected, and factors that have occurred in their past are studied

Right-censored data: Come from patients who were known not to have reached the endpoint of interest when they were last under follow-up

Risk factor: A determinant that affects the incidence of a particular outcome, e.g. a disease

Risk of disease: The probability of developing the disease in the stated time period; it is estimated by the number of new cases of disease in the period divided by the number of individuals disease-free at the start of the period

Risk score: See prognostic index

Robust standard error: Based on the variability in the data rather than on that assumed by the regression model**:** more robust to violations of the underlying assumptions of the regression model than estimates from OLS

Robust: A test is robust to violations of its assumptions if its *P*-value and the power are not appreciably affected by the violations

RR: See relative risk

Sample: A subgroup of the population

Sampling distribution of the mean: The distribution of the sample means obtained after taking repeated samples of a fixed size from the population

Sampling distribution of the proportion: The distribution of the sample proportions obtained after taking repeated samples of a fixed size from the population

Sampling error: The difference, attributed to taking only a sample of values, between a population parameter and its sample estimate

Sampling frame: A list of all the individuals in the population

Saturated model: One in which the number of variables equals or is greater than the number of individuals

Scatter diagram: The two-dimensional plot of one variable against another, with each pair of observations marked by a point

Screening: A process to ascertain which individuals in an apparently healthy population are likely to have (or, sometimes, not have) the disease of interest

SD: See standard deviation

Secondary endpoints: The outcomes in a clinical trial that are not of primary importance

Selection bias: A systematic distortion of the data resulting from the way in which individuals are included in a sample

SEM: See standard error of mean

Sensitivity: The proportion of individuals with the disease who are correctly diagnosed by the test

Shapiro–Wilk test: Determines whether data are Normally distributed

Shrinkage: process used in estimation of parameters in a random effects model to bring each cluster's estimate of the effect of interest closer to the mean effect from all the clusters

Sign test: A non-parametric test that investigates whether differences tend to be positive (or negative); whether observations tend to be greater (or less) than the median; whether the proportion of observations with a characteristic is greater (or less) than one half

Significance level: The probability, chosen at the outset of an investigation, which will lead us to reject the null hypothesis if our *P*-value lies below it. It is often chosen as 0.05

Significance test: See hypothesis test

Simple linear regression: The straight line relationship between a single dependent variable and a single explanatory variable

Single-blind: See blinding

Skewed distribution: The distribution of the data is asymmetrical; it has a long tail to the right with a few high values (*positively* skewed) or a long tail to the left with a few low values (*negatively* skewed)

Slope: The gradient of the regression line, showing the mean change in the dependent variable for a unit change in the explanatory variable

SND: See Standardized Normal Deviate

Spearman's rank correlation coefficient: A non-parametric alternative to the Pearson correlation coefficient; it provides a measure of association between two variables

Specificity: The proportion of individuals without the disease who are correctly identified by a diagnostic test

Standard deviation (SD): A measure of spread equal to the square root of the variance

Standard error of the mean (SEM): A measure of precision of the sample mean. It is the standard deviation of the sampling distribution of the mean

Standard error of the proportion: A measure of precision of the sample proportion. It is the standard deviation of the sampling distribution of the proportion

Standard Normal distribution: A particular Normal distribution with a mean of zero and a variance of one

Standardized difference: A ratio, used in Altman's nomogram and Lehr's formulae, which expresses the clinically important treatment difference as a multiple of the standard deviation

Standardized Normal Deviate (SND): A random variable whose distribution is Normal with zero mean and unit variance

Statistic: The sample estimate of a population parameter

Statistical heterogeneity: Is present in a meta-analysis when there is considerable variation between the separate estimates of the effect of interest

Statistically significant: The result of a hypothesis test is statistically significant at a particular level (say 1%) if we have sufficient evidence to reject the null hypothesis at that level (i.e. when $P < 0.01$)

Statistics: Encompasses the methods of collecting, summarizing, analysing and drawing conclusions from data

Stem-and-leaf plot: A mixture of a diagram and a table used to illustrate the distribution of data. It is similar to a histogram, and is effectively the data values displayed in increasing order of size

Stepwise selection: See model selection

Stratum: A subgroup of individuals; usually, the individuals within a stratum share similar characteristics. Sometimes called a block

Student's *t*-distribution: See *t*-distribution

Subjective probability: Personal degree of belief that an event will occur

Superiority trial: Used to demonstrate that two or more treatments are clinically different

Survival analysis: Examines the time taken for an individual to reach an endpoint of interest (e.g. death) when some data are censored

Symmetrical distribution: The data are centred around some midpoint, and the shape of the distribution to the left of the midpoint is a mirror image of that to the right of it

Systematic allocation: Patients in a clinical trial are allocated treatments in a systematized, non-random manner

Systematic review: A formalized and stringent approach to combining the results from all relevant studies of similar investigations of the same health condition

Systematic sampling: The sample is selected from the population using some systematic method rather than that based on chance

***t*-distribution:** Also called Student's *t*-distribution. A continuous distribution, whose shape is similar to the Normal distribution, characterized by its degrees of freedom. It is particularly useful for inferences about the mean

Test statistic: A quantity, derived from sample data, used to test a statistical hypothesis; its value is compared with a known probability distribution to obtain a *P*-value

Time-dependent variable: An explanatory variable in a regression analysis (e.g. in Poisson regression or Cox survival analysis) that takes different values for a given individual at various times in the study

Training sample: The first sample used to generate the model (e.g. in logistic regression or discriminant analysis). The results are authenticated by a second (validation) sample

Transformed data: Obtained by taking the same mathematical transformation (e.g. log) of each observation

Treatment effect: The effect of interest (e.g. the difference between means or the relative risk) that affords a treatment comparison

Trend: Values of the variable show a tendency to increase or decrease progressively over time

Two-sample *t*-test: See unpaired *t*-test

Two-tailed test: The direction of the effect of interest is not specified in the alternative hypothesis

Type I error: Rejection of the null hypothesis when it is true

Type II error: Non-rejection of the null hypothesis when it is false

Unbiased: Free from bias

Underdispersion: Occurs when the residual variance is less than that expected by the defined regression model (e.g. Binomial or Poisson)

Unexplained variation: See residual variation

Uniform distribution: Has no 'peaks' because each value is equally likely

Unimodal distribution: Has a single 'peak'

Univariable regression model: Has one outcome variable and one explanatory variable

Unpaired (two-sample) *t*-test: Tests the null hypothesis that two means from independent groups are equal

Validation sample: A second sample, used to authenticate the results from the training sample

Validity: Closeness to the truth

Variable: Any quantity that varies

Variance ratio (*F*-) test: Used to compare two variances by comparing their ratio to the *F*-distribution

Variance: A measure of spread equal to the square of the standard deviation

Wald test statistic: Used to test the significance of a parameter in a regression model; it follows the standard Normal distribution

Washout period: The interval between the end of one treatment period and the start of the second treatment period in a cross-over trial. It allows the residual effects of the first treatment to dissipate

Weighted kappa: A refinement of Cohen's kappa, measuring agreement, which takes into account the extent to which two sets of paired ordinal categorical measurements disagree

Weighted mean: A modification of the arithmetic mean, obtained by attaching weights to each value of the variable in the data set

Wilcoxon rank sum (two-sample) test: A non-parametric test comparing the distributions of two independent groups of observations. It is equivalent to the Mann–Whitney *U* test.

Wilcoxon signed ranks test: A non-parametric test comparing paired observations

Index

Page numbers in **bold** represent tables, those in *italics* represent figures.