

Medical Statistics from Scratch

An Introduction for Health Professionals

Second Edition

David Bowers

Honorary Lecturer, School of Medicine, University of Leeds, UK



John Wiley & Sons, Ltd

Medical Statistics from Scratch

Second Edition

Medical Statistics from Scratch

An Introduction for Health Professionals

Second Edition

David Bowers

Honorary Lecturer, School of Medicine, University of Leeds, UK



John Wiley & Sons, Ltd

Copyright © 2008 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England
Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The Publisher is not associated with any product or vendor mentioned in this book.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, Ontario, L5R 4J3

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Bowers, David, 1938–

Medical statistics from scratch : an introduction for health professionals / David Bowers. — 2nd ed.
p. ; cm.

Includes bibliographical references and index.

ISBN 978-0-470-51301-9 (cloth : alk, paper)

1. Medical statistics. 2. Medicine—Research—Statistical methods. I. Title.

[DNLM: 1. Biometry. 2. Statistics—methods. WA 950 B786m 2007]

RA409.B669 2007

610.72'7—dc22

2007041619

British Library Cataloguing in Publication Data

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

ISBN 978-0-470-51301-9

Typeset in 10/12pt Minion by Aptara Inc., New Delhi, India

Printed and bound in Great Britain by Antony Rowe Ltd., Chippenham, Wilts

This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

This book is for Susanne

Contents

Preface to the 2nd Edition	xi
Preface to the 1st Edition	xiii
Introduction	xv
I Some Fundamental Stuff	1
1 First things first – the nature of data	3
Learning Objectives	3
Variables and data	3
The good, the bad, and the ugly – types of variable	4
Categorical variables	4
Metric variables	7
How can I tell what type of variable I am dealing with?	9
II Descriptive Statistics	15
2 Describing data with tables	17
Learning Objectives	17
What is descriptive statistics?	17
The frequency table	18
3 Describing data with charts	29
Learning Objectives	29
Picture it!	29
Charting nominal and ordinal data	30
Charting discrete metric data	34
Charting continuous metric data	35
Charting cumulative data	37
4 Describing data from its shape	43
Learning Objectives	43
The shape of things to come	43

5 Describing data with numeric summary values	51
Learning Objectives	51
Numbers R us	52
Summary measures of location	54
Summary measures of spread	57
Standard deviation and the Normal distribution	65
III Getting the Data	69
6 Doing it right first time – designing a study	71
Learning Objectives	71
Hey ho! Hey ho! It's off to work we go	72
Collecting the data – types of sample	74
Types of study	75
Confounding	81
Matching	81
Comparing cohort and case-control designs	83
Getting stuck in – experimental studies	83
IV From Little to Large – Statistical Inference	91
7 From samples to populations – making inferences	93
Learning Objectives	93
Statistical inference	93
8 Probability, risk and odds	97
Learning Objectives	97
Chance would be a fine thing – the idea of probability	98
Calculating probability	99
Probability and the Normal distribution	100
Risk	100
Odds	101
Why you can't calculate risk in a case-control study	102
The link between probability and odds	103
The risk ratio	104
The odds ratio	105
Number needed to treat (NNT)	106
V The Informed Guess – Confidence Interval Estimation	109
9 Estimating the value of a <i>single</i> population parameter – the idea of confidence intervals	111
Learning Objectives	111
Confidence interval estimation for a population mean	112
Confidence interval for a population proportion	116
Estimating a confidence interval for the median of a single population	117

10 Estimating the difference between two population parameters	119
Learning Objectives	119
What's the difference?	120
Estimating the difference between the means of two independent populations – using a method based on the two-sample <i>t</i> test	120
Estimating the difference between two matched population means – using a method based on the matched-pairs <i>t</i> test	125
Estimating the difference between two independent population proportions	126
Estimating the difference between two independent population medians – the Mann–Whitney rank-sums method	127
Estimating the difference between two matched population medians – Wilcoxon signed-ranks method	131
11 Estimating the <i>ratio</i> of two population parameters	133
Learning Objectives	133
Estimating ratios of means, risks and odds	133
VI Putting it to the Test	139
12 Testing hypotheses about the <i>difference</i> between two population parameters	141
Learning Objectives	141
The research question and the hypothesis test	142
A brief summary of a few of the commonest tests	144
Some examples of hypothesis tests from practice	146
Confidence intervals versus hypothesis testing	149
Nobody's perfect – types of error	149
The power of a test	151
Maximising power – calculating sample size	152
Rules of thumb	152
13 Testing hypotheses about the <i>ratio</i> of two population parameters	155
Learning Objectives	155
Testing the risk ratio	155
Testing the odds ratio	158
14 Testing hypotheses about the equality of population proportions: the chi-squared test	161
Learning Objectives	161
Of all the tests in all the world . . . the chi-squared (χ^2) test	162
VII Getting up Close	169
15 Measuring the association between two variables	171
Learning Objectives	171
Association	171
The correlation coefficient	175

16 Measuring agreement	181
Learning Objectives	181
To agree or not agree: that is the question	181
Cohen's kappa	182
Measuring agreement with ordinal data – weighted kappa	184
Measuring the agreement between two metric continuous variables	184
VIII Getting into a Relationship	187
17 Straight line models: linear regression	189
Learning Objectives	189
Health warning!	190
Relationship and association	190
The linear regression model	192
Model building and variable selection	200
18 Curvy models: logistic regression	213
Learning Objectives	213
A second health warning!	213
Binary dependent variables	214
The logistic regression model	215
IX Two More Chapters	225
19 Measuring survival	227
Learning Objectives	227
Introduction	227
Calculating survival probabilities and the proportion surviving: the Kaplan-Meier table	228
The Kaplan-Meier chart	230
Determining median survival time	231
Comparing survival with two groups	232
20 Systematic review and meta-analysis	239
Learning Objectives	239
Introduction	240
Systematic review	240
Publication and other biases	244
The funnel plot	244
Combining the studies	246
Appendix: Table of random numbers	251
Solutions to Exercises	253
References	273
Index	277

Preface to the 2nd Edition

This book is a ‘not-too-mathematical’ introduction to medical statistics. It should appeal to anyone training or working in the health care arena – whatever their particular discipline – who wants either a simple introduction to the subject, or a gentle reminder of stuff they might have forgotten. I have aimed the book at:

- Students doing a first degree or diploma in clinical and health care courses.
- Students doing post-graduate clinical and health care studies.
- Health care professionals doing professional and membership examinations.
- Health care professionals who want to brush up on some medical statistics generally, or who want a simple reminder of a particular topic.
- Anybody else who wants to know a bit of what medical statistics is about.

The most significant change in this second edition is the addition of two new chapters, one on measuring survival, and one on systematic review and meta-analysis. The ability to understand the principles of survival analysis is important, not least because of its popularity in clinical research, and consequently in the clinical literature. Similarly, the increasing importance of evidence-based clinical practice means that systematic review and meta-analysis also demand a place. In addition, I have taken the opportunity to correct and freshen the text in a few places, as well as adding a small number of new examples. My thanks to Lucy Sayer, my editor at John Wiley, for her enthusiastic support, to Liz Renwick and Robert Hambrook, and all the other wiley people, for their invaluable help and special thanks to my copy-editor Barbara Noble, for her truly excellent work and enthusiasm (of course, any remaining errors are mine).

I am happy to get any comments and criticisms from you. You can e-mail me at: slothist@hotmail.com.

Preface to the 1st Edition

This book is intended to be an introduction to medical statistics but one which is not too mathematical—in fact has the absolute minimum of maths. The exceptions however are Chapters 17 and 18, on linear and logistic regression. It's really impossible to provide material on these procedures without some maths, and I hesitated about including them at all. However they are such useful and widely used techniques, particularly logistic regression and its production of odds ratios, that I felt they must go in. Of course you don't *have* to read them. It should appeal to anyone training or working in the health care arena—whatever their particular discipline—who wants a simple, not-too-technical introduction to the subject. I have aimed the book at:

- students doing either a first degree or diploma in health care-related courses
- students doing postgraduate health care studies
- health care professionals doing professional and membership examinations
- health care professionals who want to brush up on some medical statistics generally, or who want a simple reminder of one particular topic
- anybody else who wants to know a bit of what medical statistics is about.

I intended originally to make this book an amalgam of two previous books of mine, *Statistics from Scratch for Health Care Professionals* and *Statistics Further from Scratch*. However, although it covers a lot of the same material as in those two books, this is in reality a completely new book, with a lot of extra stuff, particularly on linear and logistic regression. I am happy to get any comments and criticisms from you. You can e-mail me at: slothist@hotmail.com.

Introduction

Before the spread of personal computers, researchers had to do most things by hand (by which I mean with a calculator), and so most statistics books were full of equations and their derivations, with many pages of the necessary statistical tables. Analysing anything other than small samples could be time-consuming and error prone. You also needed to be reasonably good at maths. Of course, for the statistics specialist there is still a need for books that deal with statistical theory, and the often complex mathematics which underlies the subject.

However, now that there are computers in most offices and homes, and many professionals have some access to a computer statistics programme, there is room for books which focus more on an understanding of the principal ideas which underlie the statistical procedures, on knowing which approach is the most appropriate, and under what circumstances, and on the interpretation of the outputs from a statistics program.

I have thus tried to keep the technical stuff to a minimum. There are a few equations here and there (most in the last few chapters), but those I have provided are mainly for the purposes of doing some of the exercises. I have also assumed that readers will have a nodding acquaintance of either SPSS or Minitab. Short courses in these programs are now widely available to most clinical staff. I also provide a few examples of outputs from SPSS and Minitab, for the commonest applications, which I hope will help you make sense of any results you get. Both SPSS and Minitab have excellent *Help* facilities, which should answer most of the difficulties you may have.

Remember this is an introductory book. If you want to explore any of the methods I describe in more detail, you can always turn to one of the more comprehensive medical statistics books, such as Altman (1991), or Bland (1995).

I

Some Fundamental Stuff

1

First things first – the nature of data

Learning objectives

When you have finished this chapter, you should be able to:

- Explain the difference between nominal, ordinal, and metric discrete and metric continuous variables.
- Identify the type of a variable.
- Explain the non-numeric nature of ordinal data.

Variables and data

A *variable* is something whose value can *vary*. For example, *age*, *sex* and *blood type* are variables. *Data* are the values you get when you measure¹ a variable. For example, *32 years* (for the variable *age*), or *female* (for the variable *sex*). I have illustrated the idea in Table 1.1.

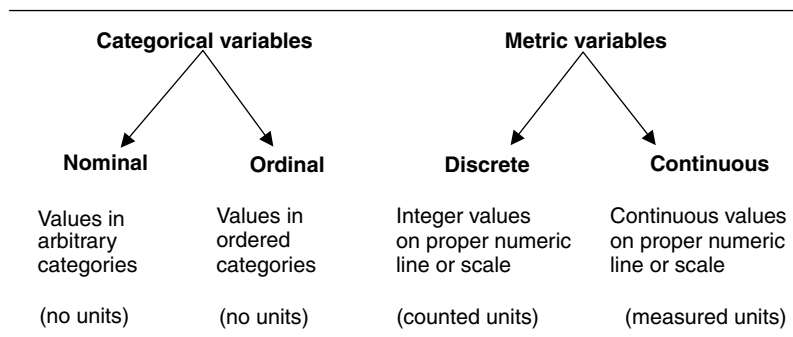
¹ I am using ‘measure’ in the broadest sense here. We wouldn’t measure the sex or the ethnicity of someone, for example. We would instead usually observe it or ask the person or get the value from a questionnaire. But we would measure their height or their blood pressure. More on this shortly.

Table 1.1 Variables and data

	Mrs Brown	Mr Patel	Ms Manda
Age	32	24	20
Sex	Female	Male	Female
Blood type	O	O	A

The good, the bad, and the ugly – types of variable

There are two major types of variable – *categorical* variables and *metric*² variables. Each of these can be further divided into two sub-types, as shown in Figure 1.1, which also summarises their main characteristics.

**Figure 1.1** Types of variable

Categorical variables

Nominal categorical variables

Consider the variable *blood type*. Let's assume for simplicity that there are only four different blood types: O, A, B, and A/B. Suppose we have a group of 100 patients. We can first determine the blood type of each and then allocate the result to one of the four blood type categories. We might end up with a table like Table 1.2.

² You will also see metric data referred to as *interval/ratio* data. The computer package SPSS uses the term 'scale' data.

Table 1.2 Blood types of 100 patients (fictitious data)

Blood type	Number of patients (or frequency)
O	65
A	15
B	12
A/B	8

By the way, a table like Table 1.2 is called a *frequency table*, or a *contingency table*. It shows how the number, or frequency, of the different blood types is *distributed* across the four categories. So 65 patients have a blood type O, 15 blood type A, and so on. We'll look at frequency tables in more detail in the next chapter.

The variable 'blood type' is a *nominal categorical* variable. Notice two things about this variable, which is typical of all nominal variables:

- The data do not have any units of measurement.³
- The ordering of the categories is completely *arbitrary*. In other words, the categories cannot be ordered in any meaningful way.⁴

In other words we could just as easily write the blood type categories as A/B, B, O, A or B, O, A, A/B, or B, A, A/B, O, or whatever. We can't say that being in any particular category is better, or shorter, or quicker, or longer, than being in any other category.

Exercise 1.1 Suggest a few other nominal variables.

Ordinal categorical variables

Let's now consider another variable some of you may be familiar with – the Glasgow Coma Scale, or GCS for short. As the name suggests, this scale measures the degree of brain injury following head trauma. A patient's Glasgow Coma Scale score is judged by their responsiveness, *as observed* by a clinician, in three areas: eye opening response, verbal response and motor response. The GCS score can vary from 3 (death or severe injury) to 15 (mild or no injury). In other words, there are 13 possible values or categories of brain injury.

Imagine that we determine the Glasgow Coma Scale scores of the last 90 patients admitted to an Emergency Department with head trauma, and we allocate the score of each patient to one of the 13 categories. The results might look like the frequency table shown in Table 1.3.

³ For example, cm, or seconds, or ccs, or kg, etc.

⁴ We are excluding trivial arrangements such as alphabetic.

Table 1.3 A frequency table showing the (hypothetical) distribution of 90 Glasgow Coma Scale scores

Glasgow Coma Scale score	Number of patients
3	8
4	1
5	6
6	5
7	5
8	7
9	6
10	8
11	8
12	10
13	12
14	9
15	5

The Glasgow Coma Scale is an *ordinal categorical* variable. Notice two things about this variable, which is typical of all ordinal variables:

- The data do not have any units of measurement (so the same as for nominal variables).
- The ordering of the categories is *not* arbitrary as it was with nominal variables. It *is* now possible to order the categories in a meaningful way.

In other words, we can say that a patient in the category ‘15’ has less brain injury than a patient in category ‘14’. Similarly, a patient in the category ‘14’ has less brain injury than a patient in category ‘13’, and so on.

However, there is one additional and very important feature of these scores, (or any other set of ordinal scores). Namely, the difference between any pair of adjacent scores is *not necessarily the same* as the difference between any other pair of adjacent scores.

For example, the difference in the degree of brain injury between Glasgow Coma Scale scores of 5 and 6, and scores of 6 and 7, is not necessarily the same. Nor can we say that a patient with a score of say 6 has *exactly* twice the degree of brain injury as a patient with a score of 12. The direct consequence of this is that ordinal data therefore *are not real numbers*. They cannot be placed on the number line.⁵ The reason is, of course, that the Glasgow Coma Scale data, and

⁵ The number line can be visualised as a horizontal line stretching from minus infinity on the left to plus infinity on the right. Any real number, whether negative or positive, decimal or integer (whole number), can be placed somewhere on this line.

the data of most other clinical scales, are *not properly measured* but *assessed* in some way, by the clinician working with the patient.⁶ This is a characteristic of all ordinal data.

Because ordinal data are not real numbers, it is not appropriate to apply any of the rules of basic arithmetic to this sort of data. You should not add, subtract, multiply or divide ordinal values. This limitation has marked implications for the sorts of analyses we can do with such data – as you will see later in this book.



Exercise 1.2 Suggest a few more scales with which you may be familiar from your clinical work.

Exercise 1.3 Explain why it wouldn't really make sense to calculate an average Glasgow Coma Scale for a group of head injury patients.

Metric variables

Continuous metric variables

Look at Table 1.4, which shows the weight in kg (rounded to two decimal places) of six individuals.

⁶ There are some scales that may involve *some* degree of proper measurement, but these will still produce ordinal values if even one part of the score is determined by a non-measured element.

Table 1.4 The weight of six patients

Patient	Weight (kg)
Ms V. Wood	68.25
Mr P. Green	80.63
Ms S. Lakin	75.00
Mrs B. Noble	71.21
Ms G. Taylor	73.44
Ms J. Taylor	76.98

The variable ‘weight’ is a *metric continuous* variable. With metric variables, proper measurement *is* possible. For example, if we want to know someone’s weight, we can use a weighing machine, we don’t have to look at the patient and make a guess (which would be approximate), or ask them how heavy they are (very unreliable). Similarly, if we want to know their diastolic blood pressure we can use a sphygmometer.⁷ Guessing, or asking, is not necessary.

Because they can be properly measured, these variables produce data that *are* real numbers, and so can be placed on the number line. Some common examples of metric continuous variables include: birthweight (g), blood pressure (mmHg), blood cholesterol ($\mu\text{g/ml}$), waiting time (minutes), body mass index (kg/m^2), peak expiry flow (l per min), and so on. Notice that all of these variables have units of measurement attached to them. This is a characteristic of all metric continuous variables.

In contrast to ordinal values, the difference between any pair of adjacent values is exactly the same. The difference between birthweights of 4000 g and 4001 g is the same as the difference between 4001 g and 4002 g, and so on. This property of real numbers is known as the *interval property* (and as we have seen, it’s not a property possessed by ordinal values). Moreover, a blood cholesterol score, for example, of 8.4 $\mu\text{g/ml}$ is exactly twice a blood cholesterol of 4.2 $\mu\text{g/ml}$. This property is known as the *ratio property* (again not shared by ordinal values).⁸ In summary:

- Metric continuous variables can be properly *measured* and have units of measurement.
- They produce data that are real numbers (located on the number line).

These properties are in marked contrast to the characteristics of nominal and ordinal variables.

Because metric data values are real numbers, you can apply all of the usual mathematical operations to them. This opens up a much wider range of analytical possibilities than is possible with either nominal or ordinal data – as you will see.

Exercise 1.4 Suggest a few continuous metric variables with which you are familiar. What is the difference between, and consequences of, assessing the value of something and measuring it?

⁷ We call the device we use to obtain the measured value, e.g. a weighing scale, or a sphygmometer, or tape measure, etc., a *measuring instrument*.

⁸ It is for these two reasons that metric data is also known as ‘interval/ratio’ data – but ‘metric’ data is shorter!

Table 1.5 The number of times that a group of children with asthma used their inhalers in the past 24 hours

Patient	Number of times inhaler used in past 24 hours
Tim	1
Jane	2
Susie	6
Barbara	6
Peter	7
Gill	8

Discrete metric variables

Consider the data in Table 1.5. This shows the number of times in the past 24 hours that each of six children with asthma used their inhalers.

Continuous metric data usually comes from *measuring*. Discrete metric data, such as that in Table 1.5, usually comes from *counting*. For example, number of deaths, number of pressure sores, number of angina attacks, and so on, are all discrete metric variables. The data produced are real numbers, and are invariably integer (i.e. whole number). They can be placed on the number line, and have the same interval and ratio properties as continuous metric data:

- Metric discrete variables can be properly *counted* and have units of measurement – ‘numbers of things’.
- They produce data which are real numbers located on the number line.

Exercise 1.5 Suggest a few discrete metric variables with which you are familiar.

Exercise 1.6 What is the difference between a continuous and a discrete metric variable? Somebody shows you a six-pack egg carton. List (a) the possible number of eggs that the carton could contain; (b) the number of possible values for the weight of the empty carton. What do you conclude?

How can I tell what type of variable I am dealing with?

The easiest way to tell whether data is metric is to check whether it has *units* attached to it, such as: g, mm, °C, $\mu\text{g}/\text{cm}^3$, *number of pressure sores*, *number of deaths*, and so on. If not, it may be ordinal or nominal – the former if the values can be put in any meaningful order. Figure 1.2 is an aid to variable type recognition.

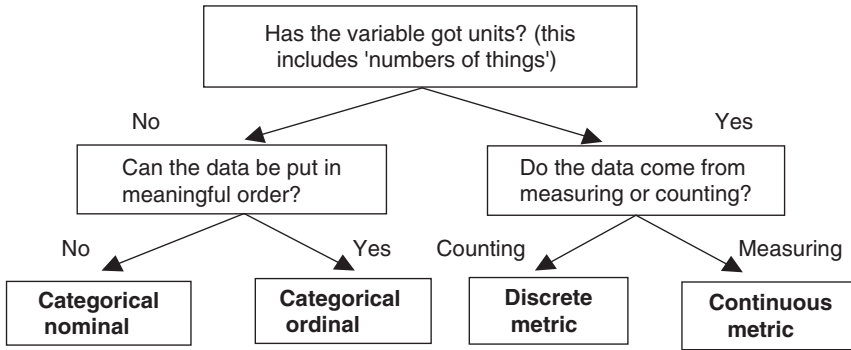


Figure 1.2 An algorithm to help identify variable type

Exercise 1.7 Four migraine patients are asked to assess the severity of their migraine pain one hour after the first symptoms of an attack, by marking a point on a horizontal line, 100 mm long. The line is marked 'No pain' at the left-hand end, and 'Worst possible pain' at the right-hand end. The distance of each patient's mark from the left-hand end is subsequently measured with a mm rule, and their scores are 25 mm, 44 mm, 68 mm and 85 mm. What sort of data is this? Can you calculate the average pain of these four patients? Note that this form of measurement (using a line and getting subjects to mark it) is known as a visual analogue scale (VAS).

Exercise 1.8 Table 1.6 contains the characteristics of cases and controls from a case-control study⁹ into stressful life events and breast cancer in women (Protheroe *et al.* 1999). Identify the type of each variable in the table.

Exercise 1.9 Table 1.7 is from a cross-section study to determine the incidence of pregnancy-related venous thromboembolic events and their relationship to selected risk factors, such as maternal age, parity, smoking, and so on (Lindqvist *et al.* 1999). Identify the type of each variable in the table.

Exercise 1.10 Table 1.8 is from a study to compare two lotions, Malathion and *d*-phenothrin, in the treatment of head lice (Chosidow *et al.* 1994). In 193 schoolchildren, 95 children were given Malathion and 98 *d*-phenothrin. Identify the type of each variable in the table.

At the end of each chapter you should look again at the learning objectives and satisfy yourself that you have achieved them.

⁹Don't worry about the different types of study, I will discuss them in detail in Chapter 6.

Table 1.6 Characteristics of cases and controls from a case-control study into stressful life events and breast cancer in women. Values are mean (SD) unless stated otherwise. Reproduced from *BMJ*, 319, 1027–30, courtesy of BMJ Publishing Group

Variable	Breast cancer group (n = 106)	Control group (n = 226)	P value
Age	61.6 (10.9)	51.0 (8.5)	0.000*
Social class† (%):			
I	10 (10)	20 (9)	
II	38 (36)	82 (36)	
III non-manual	28 (26)	72 (32)	0.094‡
III manual	13 (12)	24 (11)	
IV	11 (10)	21 (9)	
V	3 (3)	2 (1)	
VI	3 (3)	4 (2)	
No of children (%):			
0	15 (14)	31 (14)	
1	16 (15)	31 (13.7)	0.97
2	42 (40)	84 (37)	
≥3	32 (31)†	80 (35)	
Age at birth of first child	21.3 (5.6)	20.5 (4.3)	0.500*
Age at menarche	12.8 (1.4)	13.0 (1.6)	0.200*
Menopausal state (%):			
Premenopausal	14 (13)	66 (29)	
Perimenopausal	9 (9)	43 (19)	0.000§
Postmenopausal	83 (78)	117 (52)	
Age at menopause	47.7 (4.5)	45.6 (5.2)	0.001*
Lifetime use of oral contraceptives (%)	38	61	0.000‡
No of years taking oral contraceptives	3.0 (5.4)	4.2 (5.0)	0.065§
No of months breastfeeding	(n = 90)	(n = 195)	
	7.4 (9.9)	7.4 (12.1)	0.990*
Lifetime use of hormone replacement therapy (%)	29 (27)	78 (35)	0.193§
Mean years of hormone replacement therapy	1.6 (3.7)	1.9 (4.0)	0.460*
Family history of ovarian cancer (%)	8 (8)	10 (4)	0.241§
History of benign breast disease (%)	15 (15)	105 (47)	0.000§
Family history of breast cancer¶ (%)	16 (15)	35 (16)	0.997§
Units of alcohol/week (%):			
0	38 (36)	59 (26)	
0–4	26 (25)	71 (31)	0.927‡
5–9	20 (19)	52 (23)	
≥10	22 (21)	44 (20)	
No of cigarettes/day:			
0	83 (78.3)	170 (75.2)	
1–9	8 (7.6)	14 (6.2)	0.383‡
≥10	15 (14.2)	42 (18.6)	
Body mass index (kg/m ²)	26.8 (5.5)	24.8 (4.2)	0.001*

*Two sample t test.

†Data for one case missing.

‡ χ^2 test for trend.§ χ^2 test.

¶No data for one control.

Table 1.7 Patient characteristics from a cross-section study of thrombotic risk during pregnancy. Reproduced with permission from Elsevier (*Obstetrics and Gynaecology*, 1999, Vol. **94**, pages 595–599).

	Thrombosis cases (<i>n</i> = 608)	Controls (<i>n</i> = 114,940)	OR	95% CI
Maternal age (y) (classification 1)				
≤19	26 (4.3)	2817 (2.5)	1.9	1.3, 2.9
20–24	125 (20.6)	23,006 (20.0)	1.1	0.9, 1.4
25–29	216 (35.5)	44,763 (38.9)	1.0	Reference
30–34	151 (24.8)	30,135 (26.2)	1.0	0.8, 1.3
≥35	90 (14.8)	14,219 (12.4)	1.3	1.0, 1.7
Maternal age (y) (classification 2)				
≤19	26 (4.3)	2817 (2.5)	1.8	1.2, 2.7
20–34	492 (80.9)	97,904 (85.2)	1.0	Reference
≥35	90 (14.8)	14,219 (12.4)	1.3	1.0, 1.6
Parity				
Para 0	304 (50.0)	47,425 (41.3)	1.8	1.5, 2.2
Para 1	142 (23.4)	40,734 (35.4)	1.0	Reference
Para 2	93 (15.3)	18,113 (15.8)	1.5	1.1, 1.9
≥Para 3	69 (11.3)	8429 (7.3)	2.4	1.8, 3.1
Missing data	0 (0)	239 (0.2)		
No. of cigarettes daily				
0	423 (69.6)	87,408 (76.0)	1.0	Reference
1–9	80 (13.2)	14,295 (12.4)	1.2	0.9, 1.5
≥10	57 (9.4)	8177 (7.1)	1.4	1.1, 1.9
Missing data	48 (7.9)	5060 (4.4)		
Multiple pregnancy				
No	593 (97.5)	113,330 (98.6)	1.0	Reference
Yes	15 (2.5)	1610 (1.4)	1.8	1.1, 3.0
Preeclampsia				
No	562 (92.4)	111,788 (97.3)	1.0	Reference
Yes	46 (7.6)	3152 (2.7)	2.9	2.1, 3.9
Cesarean delivery				
No	420 (69.1)	102,181 (88.9)	1.0	Reference
Yes	188 (30.9)	12,759 (11.1)	3.6	3.0, 4.3

OR = odds ratio; CI = confidence interval.
Data presented as *n* (%).

Table 1.8 Basic characteristics of two groups of children in a study to compare two lotions in the treatment of head lice. One group (95 children) were given Malathion lotion, the second group (98 children), *d*-phenothrin. Reprinted courtesy of Elsevier (*The Lancet*, 1994, **344**, 1724–26)

Characteristic	Malathion (<i>n</i> = 95)	<i>d</i> -phenothrin (<i>n</i> = 98)
Age at randomisation (yr)	8.6 (1.6)	8.9 (1.6)
Sex—no of children (%)		
Male	31 (33)	41 (42)
Female	64 (67)	57 (58)
Home no (mean)		
Number of rooms	3.3 (1.2)	3.3 (1.8)
Length of hair—no of children (%)*		
Long	37 (39)	20 (21)
Mid-long	23 (24)	33 (34)
Short	35 (37)	44 (46)
Colour of hair—no of children (%)		
Blond	15 (16)	18 (18)
Brown	49 (52)	55 (56)
Red	4 (4)	4 (4)
Dark	27 (28)	21 (22)
Texture of hair—no of children (%)		
Straight	67 (71)	69 (70)
Curly	19 (20)	25 (26)
Frizzy/kinky	9 (9)	4 (4)
Pruritus—no of children (%)	54 (57)	65 (66)
Excoriations—no of children (%)	25 (26)	39 (40)
Evaluation of infestation		
Live lice-no of children (%)		
0	18 (19)	24 (24)
+	45 (47)	35 (36)
++	9 (9)	15 (15)
+++	12 (13)	15 (15)
++++	11 (12)	9 (9)
Viable nits-no of children (%)*		
0	19 (20)	8 (8)
+	32 (34)	41 (45)
++	22 (23)	24 (25)
+++	18 (19)	20 (21)
++++	4 (4)	4 (4)

The 2 groups were similar at baseline except for a significant difference for the length of hair ($p = 0.02$; chi-square).
*One value missing in the *d*-phenothrin group.

Baseline characteristics of the *P Humanus capitis*-infested schoolchildren assigned to receive malathion or *d*-phenothrin lotion*

II

Descriptive Statistics

2

Describing data with tables

Learning objectives

When you have finished this chapter you should be able to:

- Explain what a frequency distribution is.
- Construct a frequency table from raw data.
- Construct relative frequency, cumulative frequency and relative cumulative frequency tables.
- Construct grouped frequency tables.
- Construct a cross-tabulation table.
- Explain what a contingency table is.
- Rank data.

What is descriptive statistics?

The next four chapters of the book are about the processes of *descriptive statistics*. What does this mean? When we first collect data for some project, it will usually be in a 'raw' form. That is, not organised in any way, making it difficult to see what's going on. Descriptive statistics is a series of procedures designed to illuminate the data, so that its principal characteristics and

main features are revealed. This may mean sorting the data by size; perhaps putting it into a table, maybe presenting it in an appropriate chart, or summarising it numerically; and so on.



An important consideration in this process is the type of variable concerned. The data from some variables are best described with a table, some with a chart, some, perhaps, with both. With other variables, a numeric summary is more appropriate. In this chapter, I am going to focus on putting the data into an appropriate table. In subsequent chapters, I will look at the use of charts and of numeric summaries.

The frequency table

We'll begin with another look the *frequency table*, which you first encountered in the previous chapter. Let's start with an example using nominal data.

Nominal variables - organising the data into non-ordered categories

In Table 1.8 we had data from the nit lotion study comparing two types of treatment for nits, Malathion or *d-phenothrin*, using a sample of 95 children, and for each child information was collected on nine variables (Chosidow *et al.* 1994). The raw data thus consisted of 95 questionnaires, each containing data on the nine variables, one being the child's hair colour blonde, brown, red and dark.

The resulting *frequency table* for the four colour categories is shown in Table 2.1. As you know, the ordering of nominal categories is arbitrary, and in this example they are shown by the number of children in each – largest first. Notice that total frequency ($n = 95$), is shown at the top of the frequency column. This is helpful to any reader and is good practice. Table 2.1 tells us how the hair colour of each of the 95 children is *distributed* across the four colour categories. In other words, Table 2.1 describes the *frequency distribution* of the variable 'hair colour'.

Table 2.1 Frequency table showing the distribution of hair colour of each of 95 children in a study of Malathion versus *d*-phenothrin for the treatment of nits

Category (hair colour)	Frequency (number of children) $n = 95$
Brown	49
Dark	27
Blonde	15
Red	4

Relative frequency

Often of more use than the actual *number* of subjects in each category are the *percentages*. Tables with this information are called *relative* or *percentage* frequency tables. The third column of Table 2.2 shows the percentage of children in each hair-colour category.

Table 2.2 Relative frequency table, showing the *percentage* of children in each hair-colour category

Category (hair colour)	Frequency (number of children) $n = 95$	Relative frequency (% of children in each category)
Brown	49	51.6
Dark	27	28.4
Blonde	15	15.8
Red	4	4.2

$$(49/95) \times 100 = 51.6$$

Exercise 2.1 Table 2.3 shows the frequency distribution for cause of blunt injury to limbs in 75 patients (Rainer *et al.* 2000). Calculate a column of relative frequencies. What percentage of patients had crush injuries?

Table 2.3 Frequency table showing causes of blunt injury to limbs in 75 patients. Reproduced from *BMJ*, **321**, 1247–51, courtesy of BMJ Publishing Group

Cause of injury	Frequency (number of patients) $n = 75$
Falls	46
Crush	20
Motor vehicle crash	6
Other	3

Table 2.4 The frequency distributions for the ordinal variable 'level of satisfaction', with nursing care by 475 psychiatric in-patients. Reproduced from *Brit J Nursing*, 3, 16–17, courtesy of MA Healthcare Limited

Satisfaction with nursing care	Frequency (number of patients) <i>n</i> = 475
Very satisfied	121
Satisfied	161
Neutral	90
Dissatisfied	51
Very dissatisfied	52

Ordinal variables – organising the data into ordered categories

When the variable in question is ordinal, we can allocate the data into ordered categories. As an example, Table 2.4 shows the frequency distribution for the variable, *level of satisfaction*, with their psychiatric nursing care, by 475 psychiatric in-patients (Rodgers and Pilgim 1991). The variable has five categories as shown.

'Level of satisfaction' is clearly an ordinal variable. 'Satisfaction' cannot be properly measured, and has no units. But the categories can be meaningfully ordered, as they have been here. The frequency values indicate that more than half of the patients were happy with their psychiatric nursing care, 282 patients (121 + 161), out of 475. Much smaller numbers expressed dissatisfaction.

Exercise 2.2 Calculate the relative frequencies for the frequency data in Table 2.4. What percentage of patients were 'very dissatisfied' with their care?

Continuous metric variables – organising the data by value

Organising raw metric *continuous* data into a frequency table is usually impractical, because there are such a large number of possible values. Indeed, there may well be no value that occurs more than once. This means that the corresponding frequency table is likely to have a large, and thus unhelpful, number of rows. Not of much help in uncovering any pattern in the data. The most useful approach with metric continuous data is to *group* them first, and then construct a frequency distribution of the grouped data. Let's see how this works.

Grouping metric continuous data

As an illustration, consider the data in the first two columns of Table 2.5, which shows the birthweight (g) of 30 infants. Birthweight is a metric continuous variable, although it is shown

Table 2.5 Raw data showing a number of characteristics associated with 30 infants, including birthweight (g)

Infant I/D ($n = 30$)	Birthweight (g)	Apgar score ^a	Sex	Mother smoked during pregnancy	Mother's parity
1	3710	8	M	no	1
2	3650	7	F	no	1
3	4490	8	M	no	0
4	3421	6	F	yes	1
5	3399	6	F	no	2
6	4094	9	M	no	3
7	4006	8	M	no	0
8	3287	5	F	yes	5
9	3594	7	F	no	2
10	4206	9	M	no	4
11	3508	7	F	no	0
12	4010	8	M	no	2
13	3896	8	M	no	0
14	3800	8	F	no	0
15	2860	4	M	no	6
16	3798	8	F	no	2
17	3666	7	F	no	0
18	4200	9	M	yes	2
19	3615	7	M	no	1
20	3193	4	F	yes	1
21	2994	5	F	yes	1
22	3266	5	M	yes	1
23	3400	6	F	no	0
24	4090	8	M	no	3
25	3303	6	F	yes	0
26	3447	6	M	yes	1
27	3388	6	F	yes	1
28	3613	7	M	no	1
29	3541	7	M	no	1
30	3886	8	M	yes	1

^aThe Apgar Scale is a measure of the well-being of new-born infants. It can vary between 0 and 10 (low scores bad).

here to the nearest integer value, greater precision not being necessary. Among the 30 infants there are *none* with the same birthweight, and a frequency table with 30 rows and a frequency of 1 in every row would add very little to what you already know from the raw data (apart from telling you what the minimum and maximum birthweights are). One solution is to *group* the data into (if possible) groups of equal width, to produce a *grouped frequency distribution*. This is only worthwhile, however, if you have enough data values, the 30 here is barely enough, but in practice there will, hopefully, be more.

The resulting grouped frequency table for birthweight is shown in Table 2.6. This gives us a much better idea of the data's main features than did the raw data. For example, you can now

Table 2.6 Grouped frequency distribution for birthweight of 30 infants (data in Table 2.5)

Birthweight (g)	No of infants (frequency) $n = 30$
2700–2999	2
3000–3299	3
3300–3599	9
3600–3899	9
3900–4199	4
4200–4499	3

see that most of the infants had a birthweight around the middle of the range of values, about 3600g, with progressively fewer values above and below this.

Exercise 2.3 The data in Table 2.7 is from a study to ascertain the extent of variation in the case-mix of adult admissions to intensive care units (ICUs) in Britain and Ireland, and its impact on outcomes (Rowan 1993). The table records the percentage mortality in 26 intensive care units. Construct a grouped frequency table of percentage mortality. What do you observe?

Table 2.7 Percentage mortality in 26 intensive care units. Reproduced from *BMJ*, 1992, **307**, 972–981, by permission of BMJ Publishing Group

ICU	1	2	3	4	5	6	7	8	9	10	11	12	13
% mortality	15.2	31.3	14.9	16.3	19.3	18.2	20.2	12.8	14.7	29.4	21.1	20.4	13.6
ICU	14	15	16	17	18	19	20	21	22	23	24	25	26
% mortality	22.4	14.0	14.3	22.8	26.7	18.9	13.7	17.7	27.2	19.3	16.1	13.5	11.2

Open-ended groups

One problem arises when one or two values are a long way from the general mass of the data, either much lower or much higher. These values are called *outliers*. Their presence can mean having lots of empty or near-empty rows at one or both ends of the frequency table. For example, one infant with a birthweight of 6050 g would mean having five empty cells before this value appears. One favoured solution is to use *open-ended* groups. If you define a new last group as ≥ 5000 g, you can record a frequency of 1 in this row,¹ and thus incorporate all of the intervening empty groups into one. As an example, the grouped age distribution at the top of Table 1.7 on p. 12 uses open-ended groups at both ends, i.e. ≤ 19 y, and ≥ 35 y.

¹ \geq means greater than or equal to; \leq means less than or equal to.

Table 2.8 Frequency table for discrete metric data showing number of times that inhaler used in past 24 hours by 53 children with asthma

Number of times inhaler used in past 24 hours	Frequency (number of children) $n = 53$
0	6
1	16
2	12
3	8
4	5
≥ 5	6

Frequency tables with discrete metric variables

Constructing frequency tables for metric *discrete* data is often less of a problem than with continuous metric data, because the number of possible values which the variable can take is often limited (although, if necessary, the data can be grouped in just the same way). As an example, Table 2.8 is a frequency table showing the number of times in the past 24 hours that 53 asthmatic children used their inhaler. We can easily see that most used their inhaler once or twice. Notice the open-ended row showing that six children had used their inhaler five or more times.

Exercise 2.4 The data below are the *parity* (the number of previous live births) of 40 women chosen at random from the 332 women in the stress and breast cancer study referred to in Table 1.6. (a) Construct frequency and relative frequency tables for this parity data. (b) Describe briefly what is revealed about the principal features of parity in these women.

4 0 2 3 2 2 3 3 0 3 1 2 8 3 4 2 1 2 2 2 2 2 3 2
2 3 0 3 2 4 0 1 3 5 1 1 0 3 2 1

Cumulative frequency

The data in Table 2.9 shows the frequency distribution of Glasgow Coma Scale score (GCS) for the last 154 patients admitted to an emergency department with head injury following a road traffic accident (RTA).

Suppose you are asked, ‘How many patients had a GCS score of 7 or less?’. You could answer this question by looking at Table 2.9 and adding up all of the values in the first five rows. But, if questions like this are likely to come up frequently, it may pay to calculate the *cumulative frequencies*. To do this we successively add, or *cumulate*, the frequency values one by one, starting at the top of the column. The results are shown in the third column of Table 2.10.

Table 2.9 The Glasgow Coma Scale scores of 154 road traffic accident patients

GCS score	Frequency (number of patients) <i>n</i> = 154
3	10
4	5
5	6
6	2
7	12
8	15
9	18
10	14
11	15
12	21
13	13
14	17
15	6

The cumulative frequency for each category tells us how many subjects there are in that category, *and* in all the lesser-valued categories in the table. For example, 35 of the total of 154 patients had a GCS score of 7 *or less*.

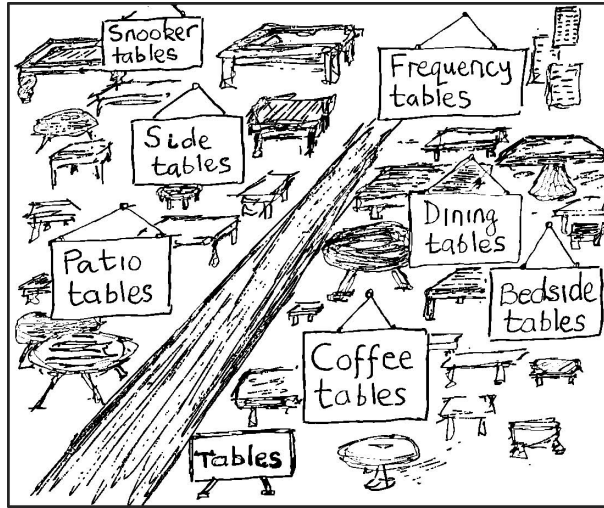
A cumulative frequency table provides us with a somewhat different view of the data. Moreover it allows us to draw a useful chart, as you will see in Chapter 3. Note that although you can legitimately calculate cumulative frequencies for both metric and ordinal data, it makes no sense to do so for nominal data, because of the arbitrary category order.

Exercise 2.5 (a) Add relative and cumulative relative frequency columns to Table 2.10. (b) What percentage of subjects had a GCS score of 10 or less?

Table 2.10 The Glasgow Coma Scale scores of Table 2.9 showing the cumulative frequency values

GCS score	Frequency (number of patients)	Cumulative frequency (cumulative number of patients)
3	10	10
4	5	15
5	6	21
6	2	23
7	12	35
8	15	50
9	18	68
10	14	82
11	15	97
12	21	118
13	13	131
14	17	148
15	6	154

Cumulative frequency is found by adding successive frequencies, i.e.
 $10 + 5 = 15$
 $15 + 6 = 21$,
 and so on, ...



Cross-tabulation

Each of the frequency tables above provides us with a description of the frequency distribution of a *single* variable. Sometimes, however, you will want to examine the association between *two* variables, within a *single* group of individuals. You can do this by putting the data into a table of *cross-tabulations*, where the rows represent the categories of one variable, and the columns represent the categories of a second variable. These tables can provide some insights into *sub-group* structures.²

To illustrate the idea, let's return to the 30 infants whose data is recorded in Table 2.5. Suppose you are particularly interested in a possible association between infants whose Apgar score is less than 7 (since this is an indicator for potential problems in the infant's well-being), and whether during pregnancy the mother smoked or not. Notice that we have only one group here, the 30 infants, but two sub-groups, those with an Apgar score of less than 7, and those with a score of 7 or more.

We have two nominal variables each with two categories, and we will thus need a cross-tab table with two rows and two columns, giving us four *cells* in total. We then need to go through the raw data in Table 2.5 and count the number of infants to be allocated to each cell. The final result is shown in Table 2.11.³

Obviously Table 2.11 is much more informative than the raw data in Table 2.5. You can see immediately that 11 out of 30 babies had Apgar scores <7, and of these 11 babies, the number with mothers who smoked (8) is almost nearly three times as large as those with non-smoking

² A 'sub-group' is a smaller identifiable group within the overall group, such as male infants and female infants, among all infants.

³ We tend to refer to cross-tabulation tables like Table 2.12 as *contingency tables* rather than frequency tables (although they are the same thing). A contingency table represents the *frequency* values for *one* group of individuals, but separated into *sub-groups*, as here for the smoking and non-smoking mothers.

Table 2.11 A cross-tabulation of the variables ‘Mother smoked during pregnancy? (Y/N)’ and ‘Apgar score <7? (Y/N)’, for 30 newborn infants (see Table 2.5)

		Apgar < 7	
		Yes	No
Mother smoked?	Yes	8	2
	No	3	17

Table 2.12 The same cross-tabulation as Table 2.11, but with values expressed as percentages of the *column* totals

		Apgar < 7 (%)	
		Yes	No
Mother smoked?	Yes	72.7	10.5
	No	27.3	89.5

mothers (3). More helpful would be a cross-tabulation with *percentage* values, like that in Table 2.12, which shows the data in Table 2.11 expressed as percentages of the *column* totals.⁴

You can see that 72.7 per cent of infants with low Apgar scores had mothers who had smoked, compared to only 27.3 per cent with mothers who hadn’t. These results might provoke you into thinking that maybe there’s a link of some sort between these two variables. Note that when appropriate you can also express the cross-tabulation with values as percentages of the *row* totals.

Exercise 2.6 The diagnosis (breast lump benign = 0; breast lump malignant = 1), for the same 40 women (in the same order), as in Exercise 2.4, is shown below. (a) Cross-tabulate *diagnosis* against *parity* (with categories, ‘two or fewer children’, and ‘more than two children’). (b) Repeat expressing the values as percentages. (c) Does the cross-tabulation suggest any possible association between diagnosis and parity?

0 0 0 0 0 1 0 0 0 0 0 1 1 1 0 0 1 0 0 1 0 0 0 0
 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0

Exercise 2.7 Using data from Table 1.6, the life stress and breast cancer study, construct a suitable 2-by-2 table, in percentage terms, with the columns being *cases* (breast cancer), and *controls* (no breast cancer), and the rows *lifetime use of oral contraceptives, OCP (yes or no)*. Comment on any patterns you can see in the table. Is this a contingency table? Explain your answer.

⁴ Note that tables with percentage values are not contingency tables.

Ranking data

As you will see later in the book, some statistical techniques require the data to be *ranked*, before any analysis takes place. Ranking means first arranging the data by size, and then giving the largest value a rank of 1, the second largest value a rank of 2, and so on.⁵ Any values which are the same, i.e. which are *tied*, are given the average rank. For example, the seven values: 2, 3, 5, 5, 5, 6, 8, could be ranked as: 1, 2, 4 = , 4 = , 4 = , 6, 7, because the three 5 values have the original ranks of 3, 4, 5, the average of which is 4. SPSS and Minitab will both rank data for you if necessary.

⁵ Or you could give the smallest a rank of 1, the next smallest a rank of 2, and so on.

3

Describing data with charts

Learning objectives

When you have finished this chapter you should be able to:

- Choose the most appropriate chart for a given data type.
- Draw pie charts; and simple, clustered and stacked, bar charts.
- Draw histograms.
- Draw step charts and ogives.
- Draw time series charts.
- Interpret and explain what a chart reveals.

Picture it!

In terms of describing data, of seeing ‘what’s going on’, an appropriate chart is almost always a good idea. What ‘appropriate’ means depends primarily on the *type* of data, as well as on what particular features of it you want to explore. In addition, if you are writing a report, a chart will always give you an ‘impact’ factor. Finally, a chart can often be used to illustrate or explain a complex situation for which a form of words or a table might be clumsy, lengthy or otherwise

Children receiving Malathion - % by hair colour

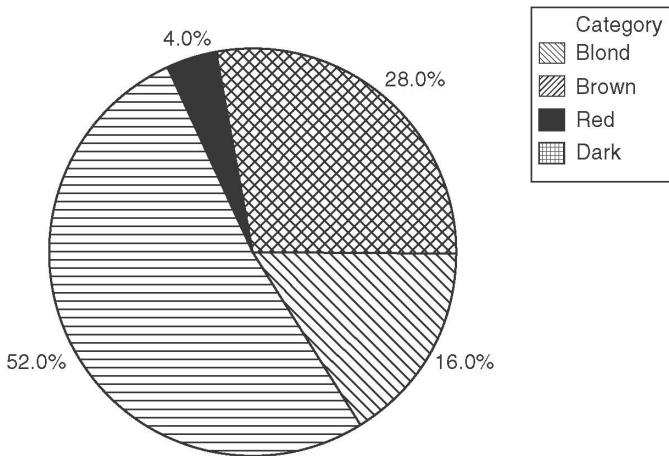


Figure 3.1 Pie chart: children receiving Malathion in nit lotion study, percentage by hair colour. Data in Table 2.1

inadequate. In this chapter I am going to examine some of the commonest charts available for describing data, and indicate which charts are appropriate for each type of data.

Charting nominal and ordinal data

The pie chart

You will all know what a pie chart is, so just a few comments here. Each segment (slice) of a pie chart should be proportional to the frequency of the category it represents. For example, Figure 3.1 is a pie chart of hair colour for the children receiving Malathion in the nit lotion study in Table 2.1. I have chosen to display the percentage values, which are often more helpful. A disadvantage of a pie chart is that it can only represent *one* variable (in Figure 3.1, hair colour). You will therefore need a separate pie chart for each variable you want to chart. Moreover a pie chart can lose clarity if it is used to represent more than four or five categories.

Exercise 3.1 The two pie charts in Figure 3.2 are from a study to investigate the types of stroke in patients with asymptomatic internal-carotid-artery stenosis (Inzitari *et al.* 2000). They show the types (in percentages) of disabling and non-disabling ipsilateral strokes, among two categories of patients: those with < 60 per cent stenosis, and those with 60–99 per cent stenosis. What is the most common type of stroke in each of the two categories of stenosis? What is the second most common type?

Exercise 3.2 Sketch a pie chart for the patient satisfaction data in Table 2.4.

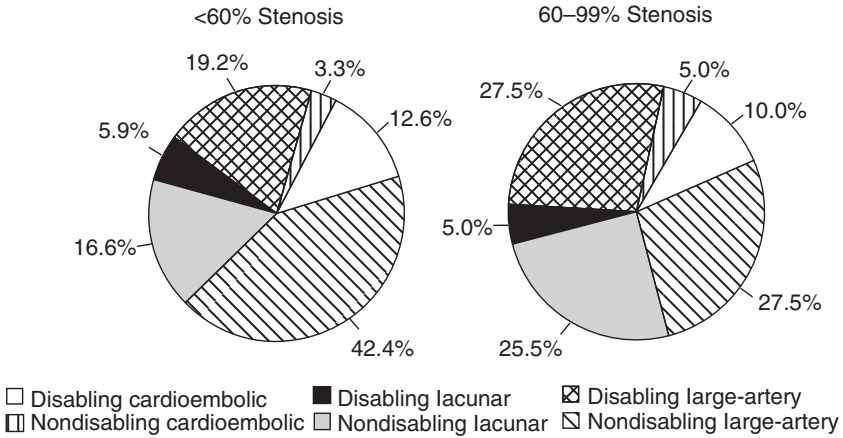


Figure 3.2 Pie charts showing the types (by percentages) of disabling and non-disabling ipsilateral strokes, among two categories of patients, those with < 60 per cent stenosis, and those with 60-99 per cent stenosis. Reproduced from *NEJM*, **342**, 1693-9, by permission of New England Journal of Medicine

The simple bar chart

An alternative to the pie chart for nominal data is the *bar chart*. This is a chart with frequency on the vertical axis and category on the horizontal axis. The *simple bar chart* is appropriate if only one variable is to be shown. Figure 3.3 is a simple bar chart of hair colour for the group of children receiving Malathion in the nit lotion study. Note that the bars should all be the same *width*, and there should be (equal) spaces between bars. These spaces emphasise the categorical nature of the data.

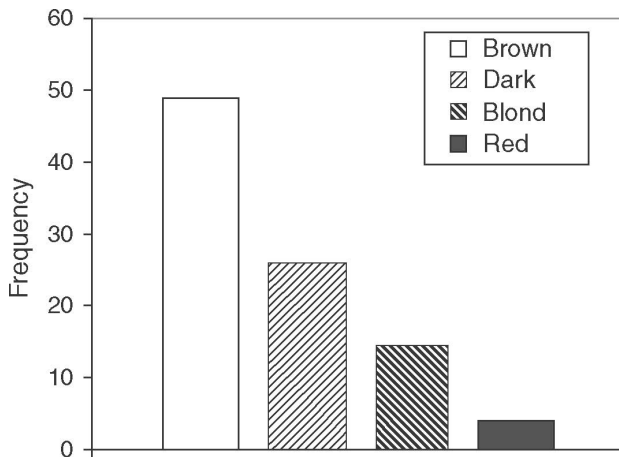


Figure 3.3 Simple bar chart of hair colour of children receiving Malathion in nit lotion study (data in Table 2.1)

Exercise 3.3 Use the data in Table 1.8 to sketch a simple bar chart, showing the hair colour of the children receiving *d*-phenothrin.

Exercise 3.4 Draw a simple bar chart for the patient satisfaction data in Table 2.4. In Exercise 3.2, you drew a pie chart for this data. Which chart do you think works best? Why?

The clustered bar chart

If you have more than one group you can use the *clustered* bar chart. Suppose you also know the *sex* of the children receiving Malathion in the above example. This gives us two sub-groups, boys and girls, with the data shown in Table 3.1.

There are two ways of presenting a clustered bar chart. Figure 3.4 shows one possibility, with hair colour categories on the horizontal axis. This arrangement is helpful if you want to compare the relative sizes of the groups *within each category* (e.g. redheaded boys versus redheaded girls).

Table 3.1 Frequency distribution of hair colour by sex of Malathion children in nit lotion study

Hair colour	Frequency	
	Boys	Girls
Blonde	4	11
Brown	29	20
Red	1	3
Dark	14	13

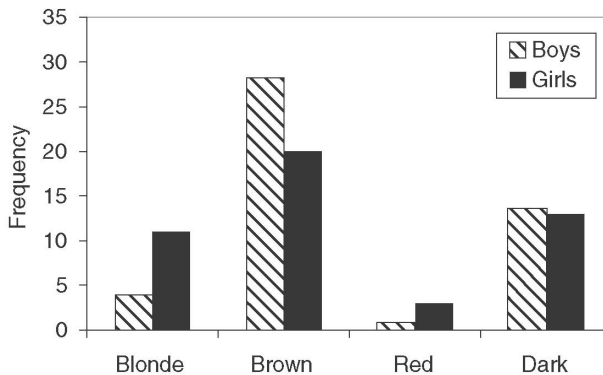


Figure 3.4 Clustered bar chart of hair colour by sex for children in Table 3.1

Alternatively, the chart could have been drawn with the categories *boys* and *girls*, on the horizontal axis. This format would be more useful if you wanted to compare category sizes *within each group*. For example, red haired girls compared to dark haired girls. Which chart is more appropriate depends on what aspect of the data you want to examine.

Exercise 3.5 Use the data in Table 3.1 to sketch a clustered percentage bar chart showing the hair colour of children receiving Malathion and *d*-phenothrin. There are two possible formats. Explain why you chose the one you did.

An example from practice

The clustered bar chart in Figure 3.5 is from a study describing the development of the APACHE II scale, used to assess risk of death, and used mainly in ICUs (Knaus *et al.* 1985). APACHE II has a range of 0 (least risk of death) to 71 (greatest risk). Data was available on two groups of patients, one group admitted to ICU for medical emergencies, the second admitted directly to ICU following surgery. The bar chart shows the percentage death rate (vertical axis), against

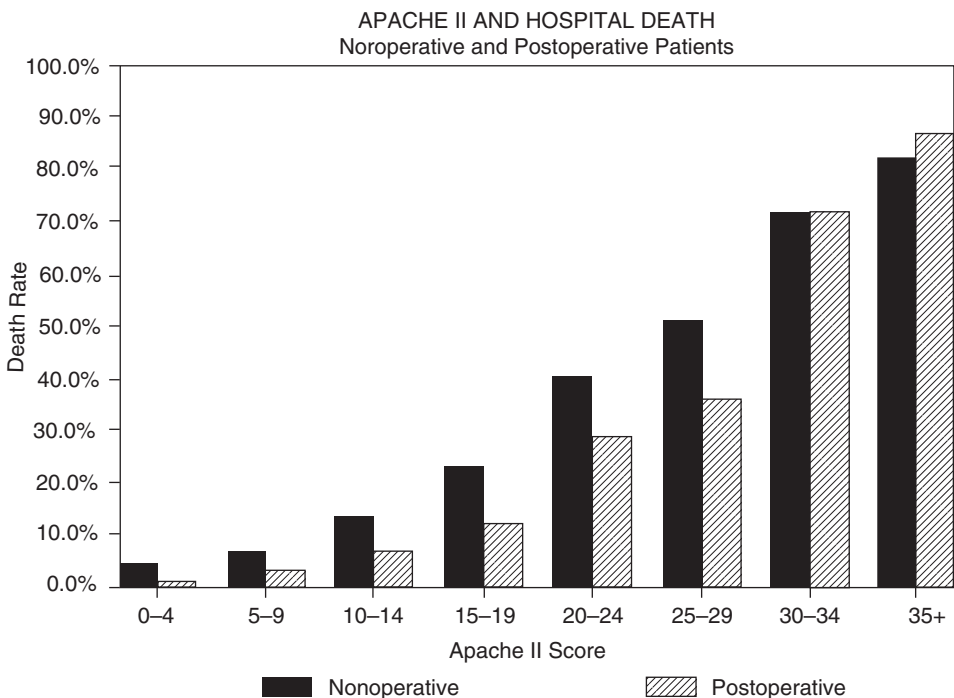


Figure 3.5 Clustered bar chart of APACHE II scores. Data on two groups of patients, one group admitted to ICU for medical emergencies, the second admitted directly to ICU following surgery. The vertical axis is death rate (per cent). Reproduced from *Critical Care Medicine*, **13**, 818-29, courtesy of Lippincott Williams Wilkins

bands of the APACHE II score. Quite clearly, for those less severely ill, percentage mortality among the medical emergency group is noticeably higher than among the post-operative group. For those patients classified as the most severely ill (scores of 35+), the situation is reversed.

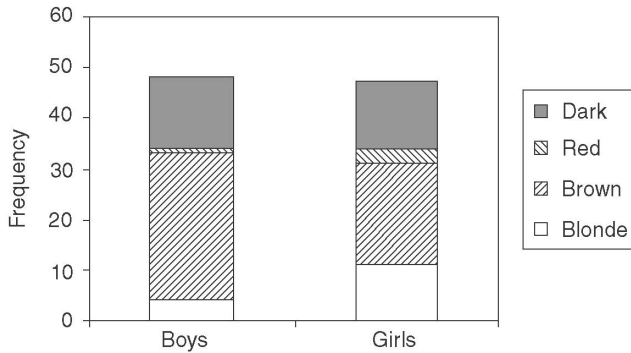


Figure 3.6 A stacked bar chart of hair colour by sex

The stacked bar chart

Figure 3.6 shows a *stacked* bar chart for the same hair colour and sex data shown in Table 3.1. Instead of appearing side by side, as in the clustered bar chart of Figure 3.5, the bars are now stacked on top of each other.¹ Stacked bar charts are appropriate if you want to compare the *total* number of subjects in each group (total number of boys and girls for example), but not so good if you want to compare category sizes *between* groups, e.g. redheaded girls with redheaded boys.

Exercise 3.6 Draw a stacked bar chart showing the same data as in Figure 3.6, but grouped by hair colour (i.e. hair colour on the horizontal axis).

Charting discrete metric data

We can use bar charts to graph discrete metric data in the same way as with ordinal data.²

¹ We could, alternatively, have used four columns for the four colour categories, with two groups per column (boys and girls). As with the clustered bar chart, the most appropriate arrangement depends on what aspects of the data you want to compare.

² In theory we should represent the discrete metric values with vertical lines and not bars, since they are 'point' values, but most common computer statistics packages don't offer this facility.

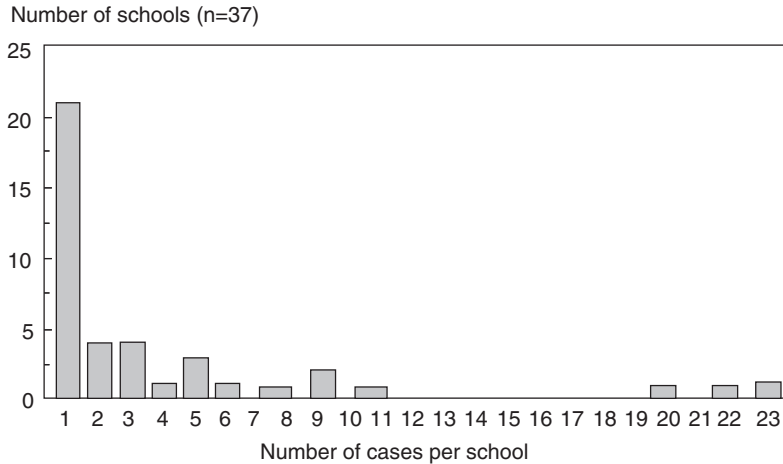


Figure 3.7 Bar chart used to represent discrete metric data on numbers of measles cases in 37 schools. Reproduced from *Amer. J. Epid.*, **146**, 881–2, courtesy of OUP

An example from practice

Figure 3.7 is an example of a bar chart used to present numbers of measles cases (discrete metric data), in 37 schools in Kentucky in a school year (Prevots *et al.* 1997).

Exercise 3.7 What does Figure 3.7 tell you about the distribution of measles cases in these 37 schools?

Charting continuous metric data

The histogram

A continuous metric variable can take a very large number of values, so it is usually impractical to plot them without first grouping the values. The *grouped* data is plotted using a *frequency histogram*, which has frequency plotted on the vertical axis and group size on the horizontal axis.

A histogram looks like a bar chart but without any gaps between adjacent bars. This emphasises the continuous nature of the underlying variable. If the groups in the frequency table are all of the same width, then the bars in the histogram will also all be of the same width.³ Figure 3.8 shows a histogram of the grouped birthweight data in Table 2.6.

One limitation of the histogram is that it can represent only one variable at a time (like the pie chart), and this can make comparisons between two histograms difficult, because, if you try to plot more than one histogram on the same axes, invariably parts of one chart will overlap the other.

³ But if one group is twice as wide as the others then the frequency must be halved, etc.

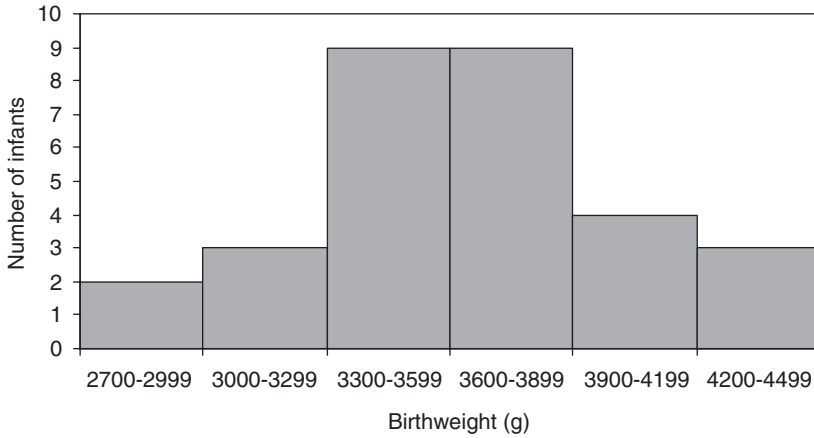


Figure 3.8 Histogram of the grouped birthweight data in Table 2.6

Exercise 3.8 The histogram in Figure 3.9 is from the British Regional Heart Study and shows the serum potassium levels (mmol/l) of 7262 men aged 40–59 *not* receiving treatment for hypertension (Wannamethee *et al.* 1997). Comment on what the histogram reveals about serum potassium levels in this sample of 7262 British men.

Exercise 3.9 The grouped age data in Table 3.2 is from a study to identify predictive factors for suicide, and shows the age distribution by sex of 974 subjects who attempted suicide unsuccessfully, and those among them who were later successful (Nordentoft *et al.* 1993). Sketch separate histograms of percentage age for the *male* attempters and for the later succeeders. Comment on what the charts show.

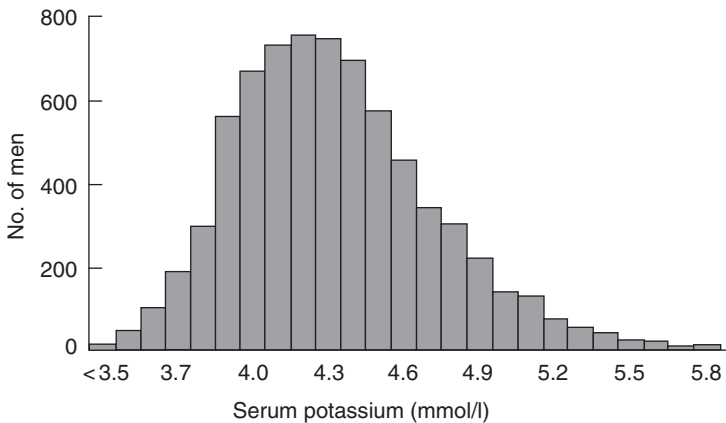


Figure 3.9 Histogram of the serum potassium levels of 7262 British men aged 40–59 years. Reproduced from *Amer. J. Epid.*, **145**, 598–607, courtesy of OUP

Table 3.2 Grouped age data from a follow-up cohort study to identify predictive factors for suicide. Reproduced from *BMJ*, 1993, **306**, 1637–1641, by permission of BMJ Publishing Group

	No (%) attempting suicide		No (%) later successful	
	Men (n = 412)	Women (n = 562)	Men (n = 48)	Women (n = 55)
Age (years)				
15–24	57 (13.8)	80 (14.2)	3 (6.3)	3 (5.5)
25–34	131 (31.8)	132 (23.5)	10 (20.8)	12 (21.8)
35–44	103 (25.0)	146 (26.0)	16 (33.3)	16 (29.1)
45–54	62 (15.0)	90 (16.0)	11 (22.9)	9 (16.4)
55–64	38 (9.2)	58 (10.3)	4 (8.3)	4 (7.3)
65–74	18 (4.4)	43 (7.7)	3 (6.3)	8 (14.5)
75–84	1 (0.2)	11 (2.0)	0	2 (3.6)
>85	2 (0.5)	2 (0.4)	1 (2.1)	1 (1.8)
Living alone	96 (23.3)	85 (15.1)	17 (35.4)	14 (25.5)
Employed	139 (33.7)	185 (32.9)	14 (29.2)	13 (23.6)

Charting cumulative data

The step chart

You can chart *cumulative* ordinal data or cumulative discrete metric data (data for both types of variables are integers) with a *step chart*. In a step chart the total height of each step above the horizontal axis represents the cumulative frequency, up to and including that category or value. The height of each individual step is the frequency of the corresponding category or value.

An example from practice

Figure 3.10 is a step chart of the cumulative rate of suicide (number per 1000 of the population), in 152 Swedish municipalities, taken from a study into the use of calcium channel blockers

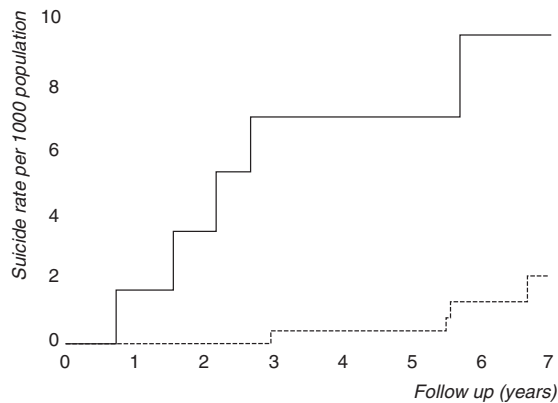


Figure 3.10 A step chart of the cumulative rate of suicide (number per 1000 of the population) in 152 Swedish municipalities. 617 users (continuous line) and 2780 non-users (dotted line). Reproduced from *BMJ*, **316**, 741–5, courtesy of BMJ Publishing Group

(prescribed for hypertension) and the risk of suicide (Lindberg *et al.* 1998). So for example, in year 4 the suicide rate per 1000 of the population was $(7 - 5.2) = 1.8$ (the approximate height of the step). And over the course of the first four years, the suicide rate had risen to seven per thousand. You can produce step charts for numeric ordinal data, such as cumulative Apgar scores in exactly the same way, although not, as far as I am aware, with Word or Excel, or with SPSS or Minitab.

Table 3.3 Cumulative and relative cumulative frequency for the grouped birthweight from the data in Table 2.6

Birthweight (g)	No of infants (frequency)	Cumulative frequency	% cumulative frequency
2700–2999	2	2	6.67
3000–3299	3	5	16.67
3300–3599	9	14	46.67
3600–3899	9	23	76.67
3900–4199	4	27	90.00
4200–4499	3	30	100.00

Exercise 3.10 Draw a step chart for the percentage cumulative Apgar scores in Table 3.3.

The cumulative frequency curve or ogive

With *continuous* metric data, there is assumed to be a smooth *continuum* of values, so you can chart cumulative frequency with a correspondingly smooth curve, known as a *cumulative frequency curve*, or *ogive*.⁴ If you add columns for cumulative and relative cumulative frequency to the grouped birthweight data in Table 2.6, you get Table 3.3.

If you want to draw an ogive by hand, you plot, for each group or class, the group cumulative frequency value against the *lower* limit of the next *higher* group. So, for example, 16.67 is plotted against 3300, 46.67 against 3600, and so on. The points should be joined with a smooth curve.⁵ The result is shown in Figure 3.11. Notice that I have put a percentage cumulative frequency of zero in the imaginary group 2400–2699 g. This enables me to close the ogive at the left-hand end.

The ogive can be very useful if you want to estimate the cumulative frequency for any value on the horizontal axis, which is not one of the original group values. For example, suppose you want to know what percentage of infants had a birthweight of 3650g or less. By drawing a line vertically upwards from a value of 3750 g on the horizontal axis to the ogive, and then horizontally to the vertical axis, you can see that about 63 per cent of the infants weighed 3750 g or less. You can of course ask such questions in reverse, for example, what birthweight marks the lowest 50 per cent of birthweights? This time you would start with a value of 50 per cent

⁴ The 'g' in ogive is pronounced as the j in 'jive'.

⁵ Unfortunately, I couldn't find a program that would allow me to join the points with a smooth curve.

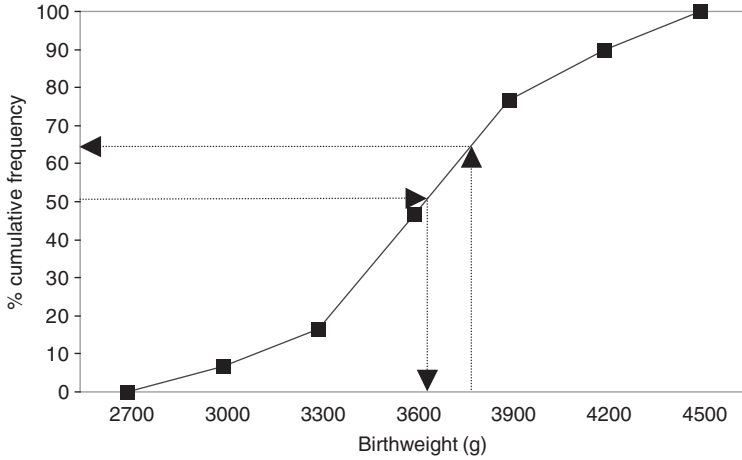


Figure 3.11 The relative cumulative frequency curve (or ogive) for the percentage cumulative birthweight data in Table 3.3

on the vertical axis, move right to the ogive, then down to the value of about 3700 g on the horizontal axis.

An example from practice

Figure 3.12 shows two per cent ogives for total cholesterol concentration in two groups taken from a study into the effectiveness of health checks conducted by nurses in primary care (Imperial Cancer Fund OXCHECK Study Group 1995)

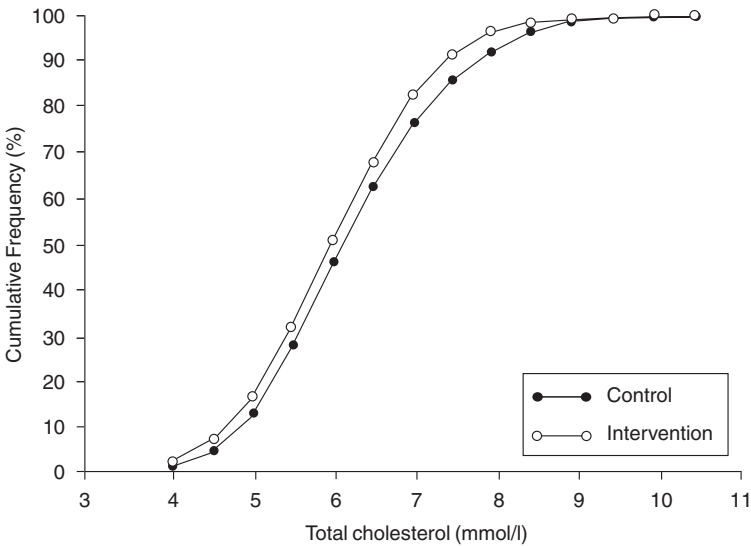


Figure 3.12 Percentage cumulative frequency curves for total cholesterol concentration in two groups. Reproduced from *BMJ*, 310, 1099–104, courtesy of BMJ

Exercise 3.11 (a) Comment on what Figure 3.12 reveals about the cholesterol levels in the two groups. (b) Sketch percentage cumulative frequency curves for the age of the male suicide attempters and later succeeders, shown in Table 3.2. For each of the two groups, half of the subjects are older than what age?



Charting time-based data – the time series chart

If the data you have collected are from measurements made at regular intervals of time (minutes, weeks, years, etc.), you can present the data with a *time series chart*. Usually these charts are used with metric data, but may also be appropriate for ordinal data. Time is always plotted on the horizontal axis, and data values on the vertical axis.

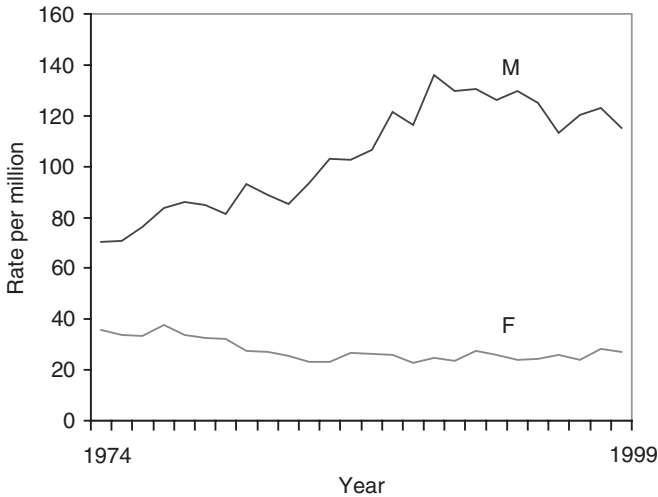


Figure 3.13 Suicide rates for males and females aged 15–29 years in England and Wales

Table 3.4 Choosing an appropriate chart

Data type	Pie chart	Bar chart	Histogram (if grouped)	Step chart	Ogive
Nominal	yes	yes	no	no	no
Ordinal	no	yes	no	yes (cumulative)	no
Metric discrete	no	yes	yes	yes (cumulative)	yes (cumulative)
Metric continuous	no	no	yes	no	yes (cumulative)

An example from practice

Figure 3.13 shows the suicide rates (number of suicides per one million of population), for males and females aged 15–29 years in England and Wales, between 1974 and 1999. The contrasting patterns in the male/female rates are noticeable, more perhaps in this chart form than if shown in a table.

There is one other useful chart, the *boxplot*, but that will have to wait until we meet some new ideas in the next two chapters. Meanwhile Table 3.4 may help you to decide on the most appropriate chart for any given set of data.

4

Describing data from its shape

Learning objectives

When you have finished this chapter you should be able to:

- Explain what is meant by the 'shape' of a frequency distribution.
- Sketch and explain: negatively skewed, symmetric and positively skewed distributions.
- Sketch and explain a bimodal distribution.
- Describe the approximate shape of a frequency distribution from a frequency table or chart.
- Sketch and describe a Normal distribution.

The shape of things to come

I have said previously that the choice of the most appropriate procedures for summarising and analysing data will depend on the type of variable involved. Variable type is the most important consideration. In addition, however, the way the data are distributed – the *shape of the distribution*, can also be influential. By 'shape' I mean:

- Are the values fairly evenly spread throughout their possible range? This is a *uniform* distribution.

- Are most of the values concentrated towards the bottom of the range, with progressively fewer values towards the top of the range? This is a *right or positively skewed* distribution. . .
- . . . or towards the top of the range, with progressively fewer values towards the bottom of the range? This is a *left or negatively skewed* distribution.
- Do most of the values clump together around *one* particular value, with progressively fewer values both below and above this value? This is a *symmetric* or *mound-shaped* distribution.
- Do most of the values clump around *two* or more particular values? This is a *bimodal* or multimodal distribution.

One simple way to assess the shape of a frequency distribution is to plot a bar chart, or a histogram. Here are some examples of the shapes described above.

Negative skew¹

Figure 4.1 shows age distribution of 2454 patients with acute pulmonary embolism and is drawn from 52 hospitals in seven countries (Goldhaber *et al.* 1999). You can see that most values lie towards the top end of the range, with progressively fewer lower values. This distribution is *negatively skewed*.

Exercise 4.1 In Figure 4.1, which age group has: (a) the highest number of patients? (b) the lowest number?

Positive skew

The histogram in Figure 4.2 shows serum E_2 levels from a study of hormone replacement therapy for osteoporosis prevention (Rodgers and Miller 1999). This distribution has most of its values in the lower end of the range with progressively fewer towards the upper end. There is a single high valued *outlier*. This distribution is *positively skewed*.

Exercise 4.2 In Figure 4.2, if the outlier was removed, would the distribution be less or more skewed?

¹ *Skewness* is the primary measure used to describe the asymmetry of frequency distributions, and many computer programs will calculate a skewness *coefficient* for you. This can vary between -1 (strong negative skew), and $+1$ (strong positive skew). Values of zero or close to it, indicate lower levels of skew, but do *not* necessarily mean that the distribution is symmetric.

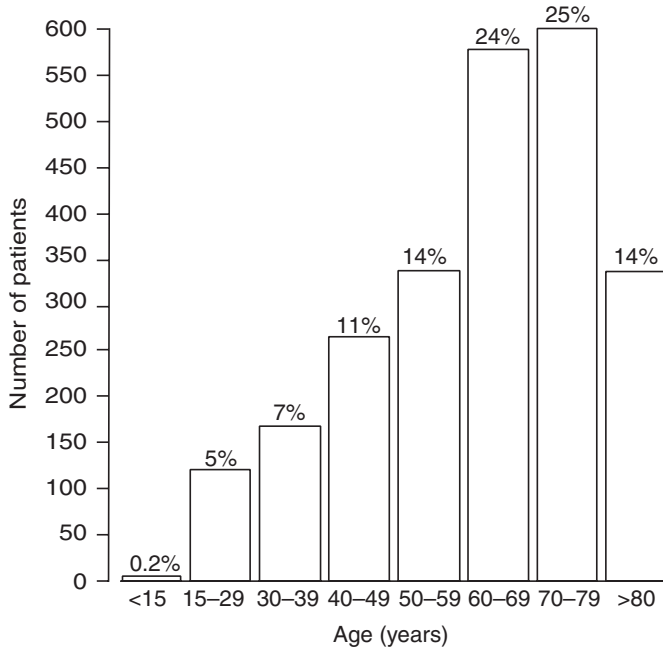


Figure 4.1 An example of negative skew. The age distribution of 2454 patients with acute pulmonary embolism. Reproduced with permission from Elsevier (*The Lancet*, 1999, Vol No. 353, pp. 1386-9)

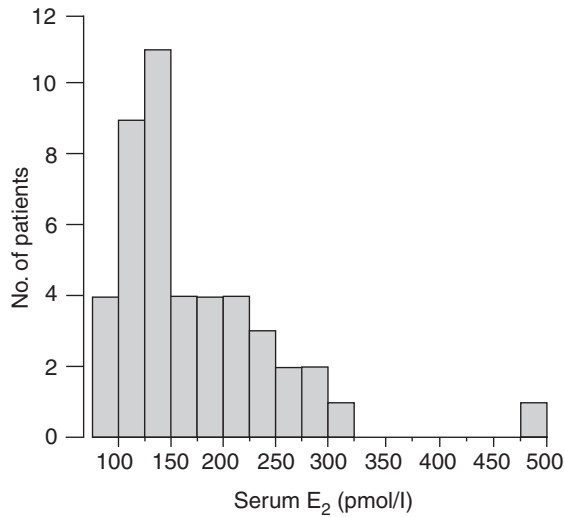


Figure 4.2 An example of positive skew. Serum E2 levels in 45 patients in a study of HRT for the prevention of osteoporosis. Reproduced with permission of the *British Journal of General Practice* (1997, Vol. 47, pages 161-165)

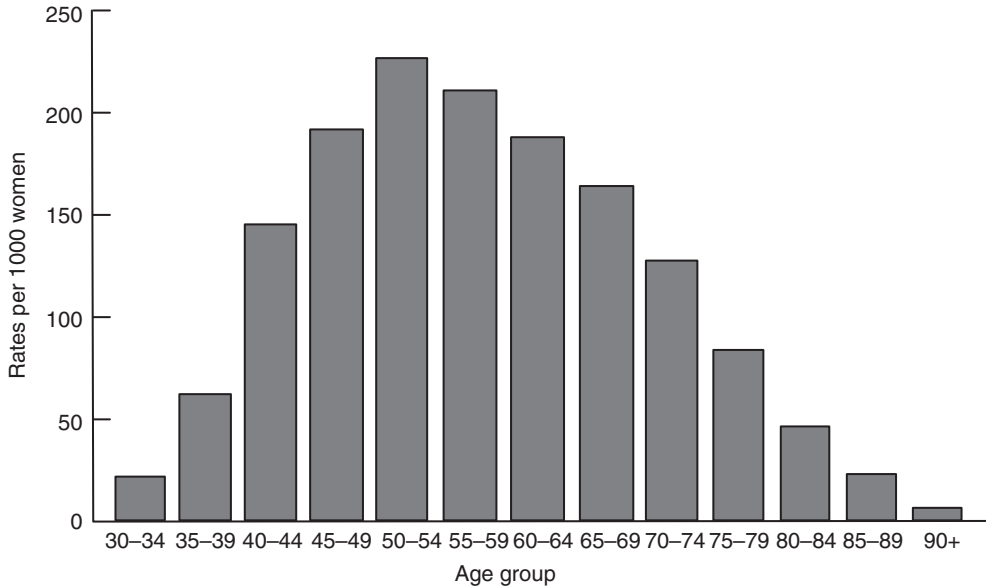


Figure 4.3 Histogram of mammography utilisations rate (per 1000 women), by broad age group, in 33 health districts in Ontario. Reproduced from *J. Epid. Comm. Health*, **51**, 378–82, courtesy of BMJ Publishing Group

Symmetric or mound-shaped distributions

The bar chart in Figure 4.3 is from a study into the use of the mammography service by women in the 33 health districts of Ontario, from mid-1990 to end-1991 (Goel *et al.* 1997). It shows the variation in the utilisation rates² by women for a number of age groups. You can see that the distribution is reasonably symmetric and mound shaped, and has only one peak.

Exercise 4.3 (a) What sort of skew is exhibited by the Apache scores in Figure 3.5? (b) The simple bar chart in Figure 4.4 is from a study describing the development of a new scale to measure psychiatric anxiety, called the Psychiatric Symptom Frequency scale (PSF) (Lindelov *et al.*). Describe the shape of the distribution of PSF in terms of symmetry, skewness, etc. Does this chart tell the whole story?

Exercise 4.4 Comment on the shapes of the age distributions shown in Table 3.2, for male and female suicide attempters, and later succeeders (you may also want to look at the histograms you drew in Exercise 3.9).

² The utilisation rate is the number of consultations per 1000 women.

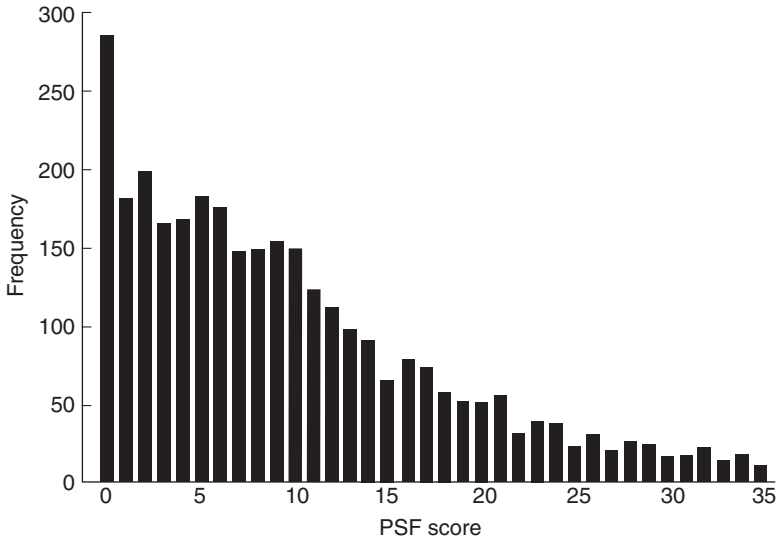


Figure 4.4 Simple bar chart showing the lowest 95 per cent of values of the Psychiatric Symptom Frequency scale. Reproduced from *J. Epid. Comm. Health*, **51**, 549–57, courtesy of BMJ Publishing Group

Bimodal distributions

A bimodal distribution is one with two distinct humps. These are less common than the shapes described above, and are sometimes the result of two separate distributions, which have not been disentangled. Figure 4.5 shows a hypothetical bimodal distribution of systolic blood pressure. The upper peak could be due to a sub-group of hypertensive patients, but whose presence in the group has not been separately identified.

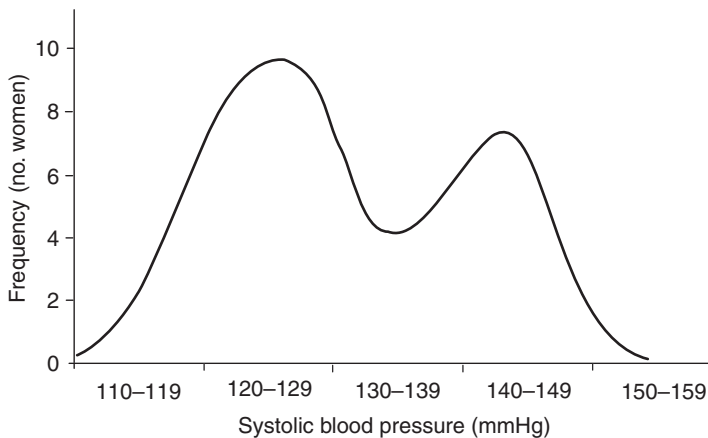
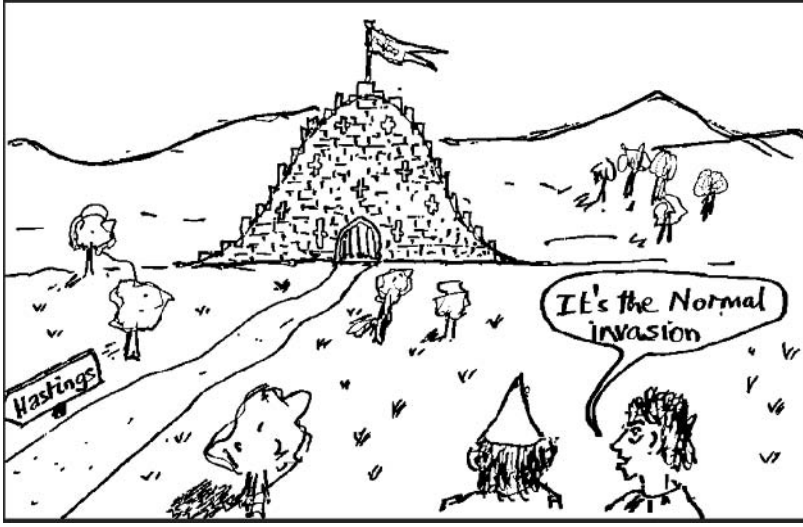


Figure 4.5 A bimodal frequency distribution

Normal-ness

There is one particular symmetric bell-shaped distribution, known as the *Normal distribution*, which has a special place in the heart of statisticians.³ Many human clinical features are distributed Normally, and the Normal distribution has a very important role to play in what is to come later in this book.



An example from practice

Figure 4.6 shows a histogram for the distribution of the cord platelet count ($10^9/l$), in 4382 Finnish infants, from a study of the prevalence and causes of thrombocytopenia⁴ in full-term infants (Sainio *et al.* 2000). You can see, even without the help of the Normal curve superimposed upon it, that the distribution has a very regular bell-shaped symmetric distribution – in fact is pretty well as Normal as it gets with real data.

Although the Normal distribution is one of the most important in a health context, you may also encounter the *binomial* and *Poisson* distributions. As an example of the former, suppose you need to choose a sample of 20 patients from a very large list of patients, which contains *equal* numbers of males and females. The chance of choosing a male patient is thus 1 in 2. Provided that the probability of picking a male patient each time remains fixed at 1 in 2, the binomial equation will tell you the probability of getting any given number of males (or females), in your 20 selected patients. For example, the probability of getting eight males in a sample of 20 patients is 0.1201 – about 12 chances in a 100.

³ Note the capitalised, 'N', to distinguish this statistical usage from that of the word 'normal' meaning usual, ordinary, etc.

⁴ Thrombocytopenia is deemed to exist when the cord platelet count is less than $150 \times 10^9/l$. It is a risk factor for intraventricular haemorrhage and contributes to the high neurological morbidity in infants affected.

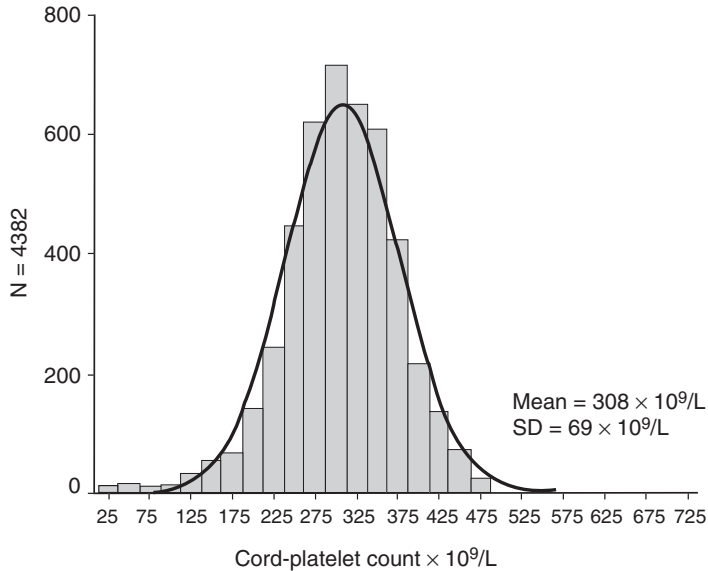


Figure 4.6 A Normal frequency curve superimposed on a histogram of cord platelet count ($10^9/l$) in 4382 infants. Reproduced from *Obstetrics and Gynecology*, **95**, 441–4, courtesy of Lippincott Williams Wilkins

The Poisson distribution is appropriate for calculating chance or probability when events occur in a seemingly random and unpredictable fashion. It describes the probability of a given number of events occurring in a fixed period of time. For example, suppose that the average number of children with burns arriving at an Emergency Department in any given 24-hour period is 12. Then the Poisson equation indicates that the probability of one child with burns arriving in the next hour is 30 in 100, the probability of two is about 7 in a 100.

To sum up so far. You have seen that you can describe the principal features of a set of data using tables and charts. A description of the shape of the distribution is also an important part of the picture. In the next chapter you will meet a way of describing data using *numeric* summary values.

5

Describing data with numeric summary values

Learning objectives

When you have finished this chapter, you should be able to:

- Explain what prevalence and incidence are.
- Explain what a summary measure of location is, and show that you understand the meaning of, and the difference between, the mode, the median and the mean.
- Be able to calculate the mode, median and mean for a set of values.
- Demonstrate that you understand the role of data type and distributional shape in choosing the most appropriate measure of location.
- Explain what a percentile is, and calculate any given percentile value.
- Explain what a summary measure of spread is, and show that you understand the difference between, and can calculate, the range, the interquartile range and the standard deviation.
- Show that you can estimate percentile values from an ogive.
- Demonstrate that you understand the role of data type and distributional shape in choosing the most appropriate measure of spread.

- Draw a boxplot and explain how it works.
- Show that you understand the area properties of the Normal distribution, and how these relate to standard deviation.

Numbers R us

As you saw in the previous two chapters, we can ‘describe’ a mass of raw data by charting it, or arranging it in table form. In addition, we can examine its shape. These procedures will help us to make some sense of what initially might be a confusing picture, and hopefully to see patterns in the data. As you are about to see, however, it is often more useful to summarise the data *numerically*. There are two principal features of a set of data that can be summarised with a single numeric value:

- First, a value around which the data has a tendency to congregate or cluster. This is called a *summary measure of location*.¹
- Second, a value which measures the degree to which the data are, or are not, spread out, called a *summary measure of spread or dispersion*.

With these two summary values you can then compare different sets of data *quantitatively*. Before I discuss these two measures, however, I want to look first at a number of simpler numeric summary measures.

Numbers, percentages and proportions

When you present the results of an investigation, you will almost certainly need to give the *numbers* of the subjects involved; and perhaps also provide values for *percentages*. In Table 1.6, the authors give the percentage of subjects who are in each ‘social class’ category. For example, 26 per cent, i.e. $(28/106) \times 100$, and 32 per cent, i.e. $(72/226) \times 100$, of the cases and controls respectively, are in the category, ‘III non-manual’. As in this example, it is usually categorical data that are summarised with a value for percentage or proportion.

Exercise 5.1 The data in Table 5.1 are taken from a study of duration of breast feeding and arterial distensibility leading to cardiovascular disease (Leeson *et al.* 2001). The table describes the basic characteristics of two groups, 149 subjects who were bottle-fed as infants, and 182 who were breast-fed. Using the values in the first row of the table in Table 3.2, calculate both the proportion and the percentage of men, among those subjects who were: (a) breastfed; (b) bottle-fed.

¹ Also known as measures of central tendency.

Table 5.1 Basic characteristics of two groups of individuals, breast-fed and bottle fed, from a study of duration of breast feeding and arterial distensibility leading to cardiovascular disease. Reproduced from *BMJ*, **322**, 643–7, courtesy of BMJ Publishing Group

Variable	Breast fed	Bottle fed	P value for difference between groups
No of participants (men/women)	149 (67/82)	182 (93/89)	—
Age (years)	23 (20 to 28)	23 (20 to 27)	0.07
Height (cm)	170 (10)	168 (9)	0.03
Weight (kg)	70.4 (14.5)	68.7 (13.1)	0.28
Body mass index (kg/m ²)	24.2 (4.1)	24.3 (3.7)	0.83
Length of breast feeding (months)	3.33 (0 to 18)	—	—
Resting arterial diameter (mm)	3.32 (0.59)	3.28 (0.59)	0.45
Distensibility coefficient (mm/Hg ⁻¹)	0.133 (0.07)	0.140 (0.08)	0.38
Cholesterol (mmol/l)	4.43 (0.99)	4.61 (1.01)	0.11
LDL cholesterol (mmol/l)	2.71 (0.88)	2.90 (0.93)	0.07
HDL cholesterol (mmol/l)	1.18 (0.25)	1.18 (0.31)	0.96
Systolic blood pressure (mm Hg)	128 (14)	128 (14)	0.93
Diastolic blood pressure (mm Hg)	70 (9)	71 (8)	0.31
Smoking history (No (%)):			
Smokers	49 (33)	64 (35)	
Former smokers	25 (17)	22 (12)	0.78
Non-smokers	75 (50)	96 (53)	
No (%) in social class:			
I	12 (8)	13 (7)	
II	36 (24)	33 (18)	
IIINM	51 (34)	62 (34)	
IIIM	24 (16)	36 (20)	0.19
IV	22 (15)	33 (18)	
V	4 (3)	5 (3)	

LDL = Low density lipoprotein, HDL = High density lipoprotein.

Prevalence and the incidence rate

If appropriate we can also summarise data by providing a value for the *prevalence* or the *incidence rate* of some condition. The *point prevalence* of a disease is the number of *existing* cases in some population at a given time. In practice, the *period prevalence* is more often used. We might typically report it as, 'the prevalence of genital chlamydia in single women in England in 1996 was 3.1 per cent'. The prevalence figure will include existing cases, i.e. those who contracted the disease before 1996, and still had it, *as well as* those first getting the disease in 1996. The *incidence* or inception rate of a disease is the number of *new* cases occurring per 1000, or per 10 000, of the population,² during some period, usually 12 months.

²Or whatever base is arithmetically appropriate.

Exercise 5.2 (a) When a group of 890 women was tested for genital chlamydia with a ligase chain reaction test, 23 of the women had a positive response. Assuming the test is always 100 per cent efficient, what is the prevalence of genital chlamydia among women in this group? (b) Suppose in a certain city that there were 10 000 live births in 2002. Ten of the infants died of sudden infant death syndrome. What is the incidence rate for sudden infant death syndrome in this city?

Summary measures of location

A summary measure of location is a value around which most of the data values tend to congregate or centre. I am going to discuss three measures of location: the mode; the median; and the mean. As you will see, the choice of the most appropriate measure depends crucially on the type of data involved. I will summarise which measure(s) you can most appropriately use with which type of data, later in the chapter

The mode

The *mode* is that category or value in the data that has the highest frequency (i.e. occurs the most often). In this sense, the mode is a measure of *common-ness* or *typical-ness*. As an example, the modal Apgar score in Table 2.5 is 8, this being the category with the highest frequency (of 9 infants), i.e. is the most commonly occurring. The mode is not particularly useful with metric continuous data where no two values may be the same. The other shortcoming of this measure is that there may be more than one mode in a set of data.

Exercise 5.3 Determine the modal category for: (a) Social class for both cases and controls, in the stress and breast cancer study shown in Table 1.6. (b) The level of satisfaction with nursing care, from the data in Table 2.4. (c) The PSF score in Figure 4.4.

Exercise 5.4 What is the modal cause of injury in Table 2.3?

The median

If we arrange the data in ascending order of size, the *median* is the middle value. Thus, half of the values will be equal to or less than the median value, and half equal to or above it. The median is thus a measure of *central-ness*. As an example of the calculation of the median, suppose you had the following data on age (in ascending order of years), for five individuals: 30 31 32 33 35. The middle value is 32, so the median age for these five people is 32 years. If you have an *even* number of values, the median is the average of the two values either side of the 'middle'.

An advantage of the median is that it is not much affected by skewness in the distribution, or by the presence of outliers. However, it discards a lot of information, because it ignores most of the values, apart from those in the centre of the distribution.

There is another, quite easy way, of determining the value of the median, which will also come in useful a bit later on. If you have n values arranged in ascending order, then:

$$\text{the median} = \frac{1}{2}(n + 1)^{\text{th}} \text{ value.}$$

So, for example, if the ages of six people are: 30 31 32 33 35 36, then $n = 6$, therefore:

$$\frac{1}{2}(n + 1) = \frac{1}{2} \times (6 + 1) = \frac{1}{2} \times 7 = 3.5.$$

Therefore the median is the 3.5th value. That is, it is the value half way between the 3rd value of 32, and the 4th value of 33, or 32.5 years, which is the same result as before.

Exercise 5.5 (a) Determine the median percentage mortality of the 26 ICUs in Table 2.7 (see also Exercise 2.3). (b) From the data in Table 3.2, determine which age group contains the median age for (i) men, and (ii) women, both for those attempting suicide, and for later successful suicides.

The mean

The mean, or the *arithmetic mean* to give it its full name, is more commonly known as the average. One advantage of the mean over the median is that it uses all of the information in the data set. However, it is affected by skewness in the distribution, and by the presence of outliers in the data. This may, on occasion, produce a mean that is not very representative of the general mass of the data. Moreover, it cannot be used with ordinal data (recall from Chapter 1 that ordinal data are not real numbers, so they cannot be added or divided).

Exercise 5.6 Comment on the likely relative sizes of the mean and median in the distributions of (a) serum potassium and (b) serum E_2 , shown in the histograms in Figure 3.9 and Figure 4.2.

Exercise 5.7 Determine the mean percentage mortality in the 26 ICUs in Table 2.7, and compare with the median value you determined in Exercise 5.5(a).

Exercise 5.8 The histogram of red blood cell thioguanine nucleotide concentration (RBCTNC), in $\text{pmol}/8 \times 10^8$ red blood cells, in 49 children, shown in Figure 5.1, is from a study into the potential causes of high incidence of secondary brain tumours in children after radiotherapy (Relling *et al.* 1999). (a) Using the information in the figure, calculate median and mean RBCTNC for the 49 children. (b) Remove the two outlier values of 3300, and re-calculate the mean and median. Compare and comment on the two sets of results.

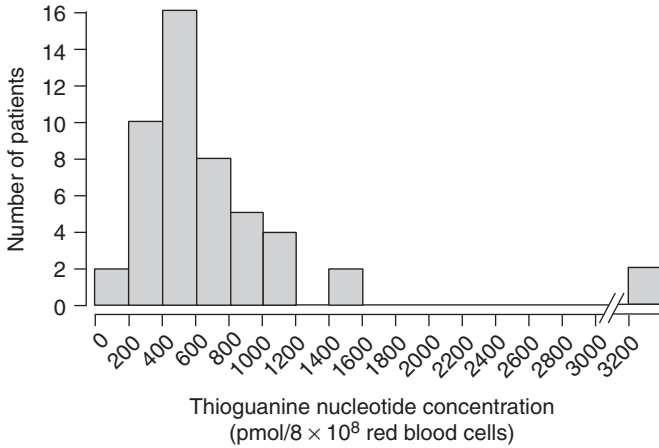


Figure 5.1 Histogram of red blood cell thioguanine nucleotide concentration (RBCTNC), in $pmol/8 \times 10^8$ red blood cells, in 49 children. Reprinted courtesy of Elsevier (*The Lancet* 2002, **354**, 34–9)

Percentiles

Percentiles are the values which divide an ordered set of data into 100 equal-sized groups. As an illustration, suppose you have birthweights for 1200 infants, which you’ve put in ascending order. If you identify the birthweight that has 1 per cent (i.e. 12) of the birthweight values below it, and 99 per cent (1188) above it, then this value is the *1st percentile*. Similarly, the birthweight which has 2 per cent of the birthweight values below it, and 98 per cent above it is the 2nd percentile. You could repeat this process until you reached the 99th percentile, which would have 99 per cent (1188) of birthweight values below it and only 1 per cent above. Notice that this makes the median the *50th percentile*, since it divides the data values into two equal halves, 50 per cent above the median and 50 per cent below.

Calculating a percentile value

How do you determine any particular percentile value? Take the example of the 30 birthweights in Table 2.5, which we reproduce below, but now in ascending order, along with their position in the order:

2860	2994	3193	3266	3287	3303	3388	3399	3400	3421	3447	3508	3541	3594	3613
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3615	3650	3666	3710	3798	3800	3886	3896	4006	4010	4090	4094	4200	4206	4490
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

The p th percentile is the value in the $p/100(n + 1)$ th position. For example, the 20th percentile is the $20/100(n + 1)$ th value. With the 30 birthweight values, the 20th percentile is therefore the $20/100(30 + 1)$ th value = 0.2×31 st value = 6.2th value. The 6th value is 3303 g and the

7th value is 3388g, a difference of 85g, so the 20th percentile is 3303g plus 0.2 of 85g, which is $3303\text{g} + 0.2 \times 85\text{g} = 3303\text{g} + 17\text{g} = 3320\text{g}$.

You might be thinking, this all seems a bit messy, but a computer will perform these calculations effortlessly. As well as percentiles, you might also encounter *deciles*, which sub-divide the data values into 10, not 100, equal divisions, and *quintiles*, which sub-divide the values into five equal-sized groups. Collectively, we call percentiles, deciles and quintiles, *n-tiles*.

Exercise 5.9 Calculate the 25th and 75th percentiles for the ICU per cent mortality values in Table 2.7, and explain your results.

Choosing the most appropriate measure

How do you choose the most appropriate measure of location for some given set of data? The main thing to remember is that the mean *cannot* be used with ordinal data (because they are not real numbers), and that the median can be used for both ordinal and metric data (particularly when the latter is skewed).

As an illustration of the last point, look again at Figure 3.7 which shows the distribution of the number of measles cases in 37 schools. Not only is this distribution positively skewed, it has a single high-valued outlier. The median number of measles cases is 1.00, but the mean number is 2.91, almost three times as many! The problem is that the long positive tail and the outlier are dragging the mean to the right. In this case, the median value of 1 seems to be more representative of the data than the mean. I have summarised the choices of a measure of location in Table 5.2.

Table 5.2 A guide to choosing an appropriate measure of location

Type of variable	Summary measure of location		
	mode	median	mean
Nominal	yes	no	no
Ordinal	yes	yes	no
Metric discrete	yes	yes, if distribution	yes
Metric continuous	no	is markedly skewed	yes

Summary measures of spread

As well as a summary measure of location, a summary measure of spread or dispersion can also be very useful. There are three main measures in common use, and once again, as you will see, the type of data influences the choice of an appropriate measure.

The range

The *range* is the distance from the smallest value to the largest. The range is not affected by skewness, but is sensitive to the addition or removal of an outlier value. As an example, the range of the 30 birthweights in Table 2.5 is (2860.0 to 4490.0) g. The range is best written like this, rather than as the single-valued difference, i.e. as 1630 g, in this example, which is much less informative.

Exercise 5.10 What are the ranges for age among those infants breast-fed, and those bottle-fed in Table 3.2?

The interquartile range (iqr)

One solution to the problem of the sensitivity of the range to extreme value (outliers) is to chop a quarter (25 per cent) of the values off both ends of the distribution (which removes any troublesome outliers), and then measure the range of the remaining values. This distance is called the *interquartile range*, or *iqr*. The interquartile range is not affected either by outliers or skewness, but it does not use all of the information in the data set since it ignores the bottom and top quarter of values.

Calculating interquartile range by hand (avoid if possible!)

To calculate the interquartile range by hand, you need first to determine two values:

- The value which cuts off the bottom 25 per cent of values; this is known as the *first quartile* and denoted $Q1$.
- The value which cuts off the top 25 per cent of values, known as the *third quartile* and denoted $Q3$.³

The interquartile range is then written as ($Q1$ to $Q3$). With the birthweight data: $Q1 = 3396.25$ g, and $Q3 = 3923.50$ g. Therefore: interquartile range = (3396.25 to 3923.50) g. This result tells you that the middle 50 per cent of infants (by weight) weighed between 3396.25 g and 3923.50 g.

An example from practice

Table 5.3 describes the baseline characteristics of 56 patients in an investigation into the use of analgesics in the prevention of stump and phantom pain in lower-limb amputation (Nikolajsen

³The median is sometimes denoted as $Q2$.

et al. 1997). The ‘blockade’ group of patients were given bupivacaine and morphine, the control (comparison) group, were given an identically administered saline placebo.

As you can see, two variables, ‘pain in week before amputation’, and ‘daily opioid consumption at admission (mg)’, were summarised with median and interquartile range values. Pain was measured using a visual analogue scale (VAS⁴), which of course produces ordinal data, so the mean is not appropriate, and the authors have used the median and interquartile range as their summary measures of location and spread.

The median level of pain in the blockade group is 51, with an iqr of (23.8 to 87.8).⁵ This means that 25 per cent of this group had a pain level of less than 23.8, and 25 per cent a pain level greater than 87.8. The middle 50 per cent had a pain level between 23.8 and 87.8. I’ll return to the opioid consumption variable shortly.

Table 5.3 The baseline characteristics of 56 patients in an investigation into the use of analgesics in the prevention of stump and phantom pain in lower-limb amputation. Reproduced from *The Lancet*, 1994, **344**, 1724–26, courtesy of Elsevier

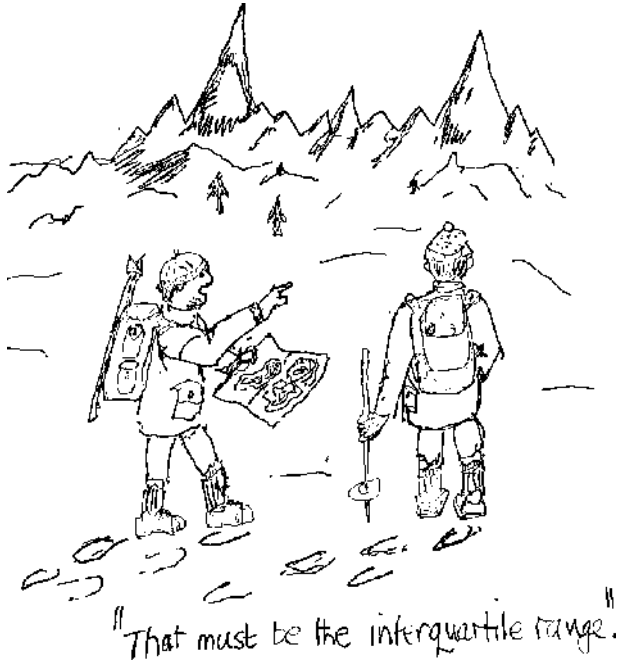
Characteristics of patients	Blockade group (n = 27)	Control group (n = 29)
Men/women	15/12	18/11
Mean (SD) age in years	72.8 (13.2)	70.8 (11.4)
Diabetes	10	14
Concurrent treatment because of cardiovascular disease	18	19
Previous stroke	3	2
Previous contralateral amputation	7	3
Median (IQR) pain in week before amputation (VAS, 0–100 mm)	51 (23.8–87.8)	44 (25.3–68)
Median (IQR) daily opioid consumption at admission (mg)	50 (20–68.8)	30 (5–62.5)
Level of amputation		
Below knee	15	16
Through knee-joint	5	2
Above knee	7	11
Reamputations during follow-up	3	2
Died during follow-up	10	10

Exercise 5.11 Calculate the iqr for the ICU percentage mortality values in Table 2.7. (You have already calculated the 25th and 75th percentiles in Exercise 5.9).

Exercise 5.12 Interpret the median and interquartile range values for pain in the week before amputation, for the control group in Table 5.3.

⁴ See Chapter 1.

⁵ The table contains a typographical error, recording 87.8 as ‘8–78’.



Estimating the median and interquartile range from the ogive

As I indicated earlier, you can estimate the median and the interquartile range from the cumulative frequency curve (the ogive). Figure 5.2 shows the ogive for the cumulative birthweight data in Table 3.3.

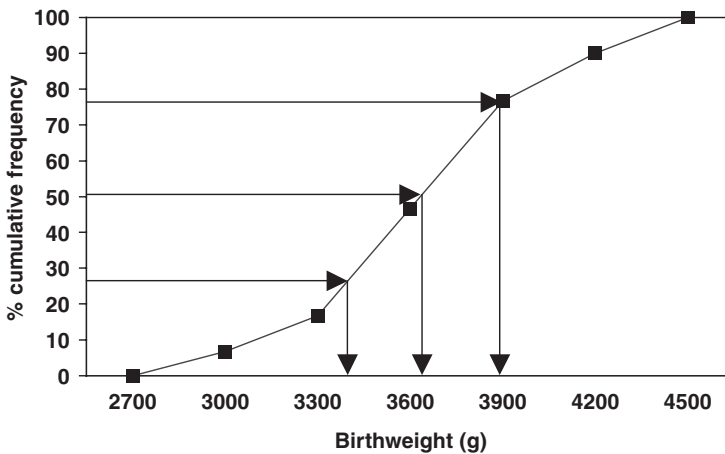


Figure 5.2 Using the relative cumulative frequency curve (or ogive) of birthweight to estimate the median and interquartile range values (Note that this *should* be a smooth curve)

If you draw horizontal lines from the values 25 per cent, 50 per cent and 75 per cent on the y axis, to the ogive, and then down to the x axis, the points of intersection on the x axis approximate values for Q1, Q2 (the median), and Q3, of 3400 g, 3650 g and 3900 g. Thus, if you happen to have an ogive handy, these approximations can be helpful. I plotted *per cent* cumulative frequency because it makes it slightly easier to do find the percentage values. Notice that you can also use the ogive to answer questions like, ‘What percentage of infants weighed less than, say, 4000 g?’ The answer is that a value of 4000 g on the x axis produces a value of 80 per cent for cumulative frequency on the y axis.

Exercise 5.13 Estimate the median and iqr for total blood cholesterol for the control group from the ogive in Figure 3.12.

The boxplot

Now that we have discussed the median and interquartile range, I can introduce the *boxplot* as I promised in Chapter 3. The general discussion on measures of spread continues overleaf if you want to continue with this and come back to consider the boxplot later. Boxplots provide a graphical summary of the three quartile values, the minimum and maximum values, and any outliers. They are usually plotted with value on the vertical axis. Like the pie chart, the boxplot can only represent one variable at a time, but a number of boxplots can be set alongside each other.

An example from practice

Figure 5.3 is from the same study as Figure 4.3, into the use of the mammography service in the 33 health districts of Ontario, in which investigators were interested in the variation in the mammography utilisation rate across age groups (Goel *et al.* 1997). They supplemented their

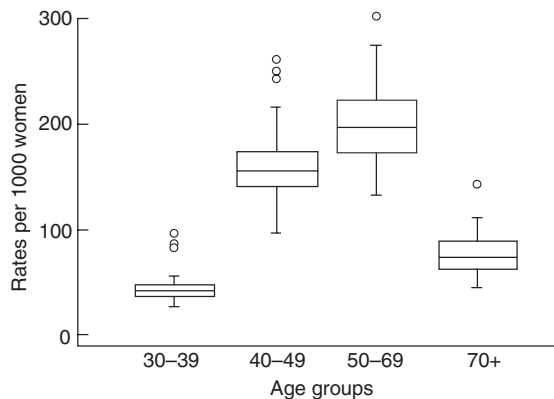


Figure 5.3 Boxplots of the rate of use of mammography services in 33 health districts in Ontario. Reproduced from *J. Epid. Comm. Health*, **51**, 378–82, courtesy of BMJ Publishing Group

results with the boxplots shown in the figure, for the age groups: (30–39); (40–49); (50–59); and 70+ years. The vertical axis is the mammography utilisation rate (visits per 1000 women), in the 33 health districts. Outliers are denoted by the small open circles.

Let's look at the third boxplot, that for the women aged 50–69:

- The bottom end of the lower 'whisker' (the line sticking out of the bottom of the box), corresponds to the minimum value – about 125 visits per 1000 women.
- The bottom of the box is the 1st quartile value, Q1. So about 25 per cent of women had a utilisation rate of 175 or less visits per 1000 women.
- The line across the inside of the box (it won't always be half-way up), is the median, Q2. So half of the women had a utilisation rate of less than about 200 consultations per 1000 women, and half a rate of more than 200. The more asymmetric (skewed) the distributional shape, the further away from the middle of the box will be the median line, closer to the top of the box is indicative of negative skew, closer to the bottom of the box – positive skew.
- The top of the box is the third quartile Q3. That is, about a quarter of women had a consultation rate of 225 or more per 1000.
- The top end of the upper whisker is the 'maximum' mammography utilisation rate – about 275 consultations per 1000 women. This is the maximum value that can be considered still to be part of the general mass of the data. Because. . .
- . . .there is one outlier. One of the health districts reported a utilisation rate of about 300 per 1000 women.⁶ This is, of course, the actual maximum value in the data.

Exercise 5.14 Sketch the box plot for the percentage mortality in ICUs shown in Table 2.7. (Note that you have already calculated the median and iqr values in Exercises 5.6 and 5.10). What can you glean from the boxplot about the shape of the distribution of the ICU percentage mortality rate?

Exercise 5.15 The boxplots in Figure 5.4 are from a study of sperm integrity in adult survivors of childhood cancer compared to a control group of non-cancer individuals (Thomson *et al.* 2002). What do the two boxplots tell you?

Standard deviation

The limitation of the interquartile range as a summary measure of spread is that (like the median) it doesn't use all of the information in the data, since it omits the top and bottom

⁶Outliers are defined in various ways by different computer programs. Outliers are here defined as any value more than three halves of the interquartile range greater than the third quartile, or less than the first quartile.

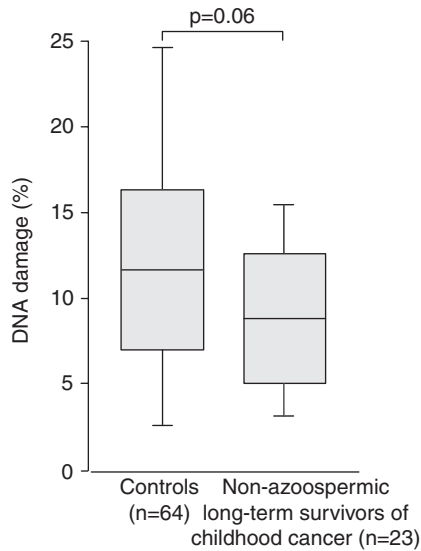


Figure 5.4 Boxplots from a study of sperm integrity in adult survivors of childhood cancer, compared to a control group of non-cancer individuals. Reprinted from *The Lancet* 2002, **360**, 361–6, Fig. 2, p. 364, courtesy of Elsevier

quarter of values. An alternative approach uses the idea of summarising spread by measuring the mean (average) distance of all the data values from the *overall* mean of all of the values. The smaller this mean distance is, the narrower the spread of values must be, and vice versa. This idea is the basis for what is known as the *standard deviation*, or s.d. The following way of calculating the sample standard deviation by hand illustrates this idea:⁷

- Subtract the mean of the sample from each of the n sample values in the sample, to give the *difference* values.
- Square each of these differences.
- Add these squared values together (called the *sum of squares*).
- Divide the sum of squares by $(n - 1)$; i.e. divide by 1 less than the sample size.⁸
- Take the square root. This is the standard deviation.

One advantage of the standard deviation is that, unlike the interquartile range, it uses all of the information in the data.

⁷ This is a very tedious procedure. If you have an s.d. key on your calculator use that. Better still, use a computer!

⁸ If we divide by n , as we normally would do to find a mean, we get a result which is slightly too small. Dividing by $(n - 1)$ adjusts for this. Technically, the sample s.d. is said to be a biased estimator of population s.d. See Chapter 7 for the meaning of sample and population.

Exercise 5.16 In Figure 4.6 the authors tell us that the mean cord platelet count is $308 \times 10^9/l$, and the standard deviation is $69 \times 10^9/l$ (notice the two measures have the same units).¹ Explain what this value means.

An example from practice

In Table 5.3, the analgesic/amputation pain study, the authors summarise the age of the patients in the study with the mean and standard deviation. As you can see, the spread of ages in the blockade group is wider than in the control group, 13.2 years around a blockade group's mean of 72.8 years, compared to 11.4 years around a control group's mean of 70.8 years.

The authors could also have used the mean and standard deviation for daily opioid consumption (mg), since this is a metric variable, but instead used the median and interquartile range; there are a number of possible reasons for this. First, the data may be noticeably skewed and/or contained outliers, perhaps making the mean a little too unrepresentative of the general mass of data. Or the investigators may have specifically wanted a summary measure of central-ness, which the median provides. Third, they may have felt that asking people to recall their opioid consumption last week was likely to lead to fuzzy, imprecise, values, and so have preferred to treat them as if they were ordinal.

Exercise 5.17 Calculate and interpret the standard deviation for the ICU percentage mortality values in Table 2.7. (You have already calculated the mean percentage mortality in Exercise 5.7). I would hesitate to do this without a calculator with a standard deviation function.

To sum up summary measures of spread: with ordinal data use either the range or the interquartile range. The standard deviation is not appropriate because of the non-numeric nature of ordinal data. With metric data use either the standard deviation, which uses all of the information in the data, or the interquartile range. The latter if the distribution is skewed, and/or you have already selected the median as your preferred measure of location. Don't mix-and-match measures – standard deviation goes with the mean, and iqr with the median. These points are summarised in Table 5.4.

Table 5.4 Choosing an appropriate measure of spread

Type of variable	Summary measure of spread		
	Range	Interquartile range	Standard deviation
Nominal	No	No	No
Ordinal	Yes	Yes	No
Metric	Yes	Yes, if skewed	Yes

¹ 10^9 means 1000 000 000.

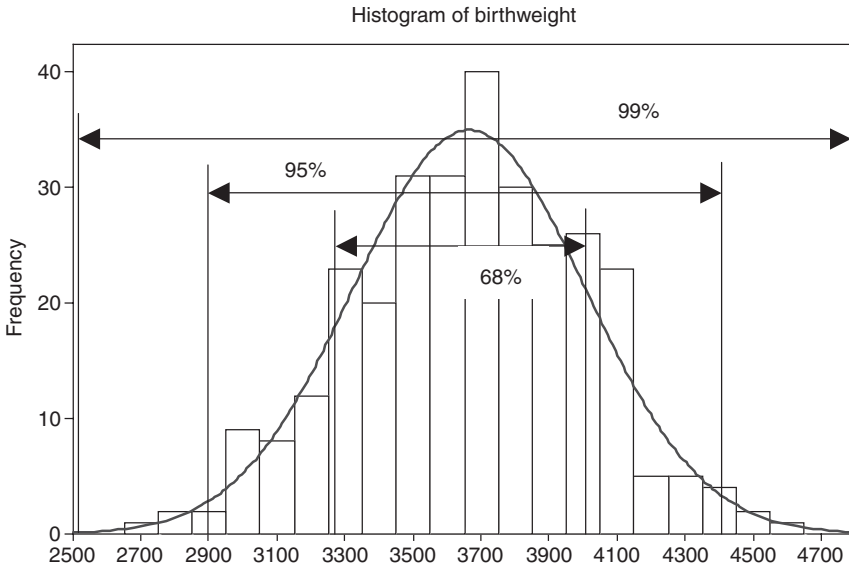


Figure 5.5 The area properties of the Normal distribution illustrated with the birthweight data

Standard deviation and the Normal distribution

If you are working with metric data which is distributed Normally, the standard deviation has one very useful property that relates to the percentage of data between certain values. These *area properties of the Normal distribution* are illustrated in Figure 5.5 for the histogram of birthweight data from Table 2.5,⁹ through which a Normal curve is drawn. Minitab calculates these birthweights to have a mean of 3644 g, and a standard deviation of 377 g. In words, the area properties are as follows:

- About 68 per cent of the birthweights will lie within one standard deviation either side of the mean. That is, from $3644\text{ g} - 377\text{ g}$ to $3644\text{ g} + 377\text{ g}$, or from 3267 g to 4021 g.
- About 95 per cent of the birthweights will lie within two standard deviations either side of the mean. That is, from $3644\text{ g} - 754\text{ g}$ to $3644\text{ g} + 754\text{ g}$, or from 2890 g to 4398 g.
- About 99 per cent of the birthweights will lie within three standard deviations either side of the mean. That is, from $3644\text{ g} - 1131\text{ g}$ to $3644\text{ g} + 1131\text{ g}$, or from 2513 g to 4775 g.

So, if you have some data that you know is Normally distributed, and you also know the values of the mean and standard deviation, then you can make statements such as, 'I know that 95 per cent of the values must lie between so-and-so and so-and-so.'

⁹ Which is reasonably Normally distributed.

An example from practice

To illustrate the usefulness of the Normal area properties, look again at the histogram of the cord platelet count for 4382 infants in Figure 4.6, which appears to be reasonably Normal, and has a mean of $308 \times 10^9/l$, and a standard deviation of $69 \times 10^9/l$. You can therefore say that about two-thirds (67 per cent) of the 4382 infants, i.e. 2936 infants, had a cord platelet count between $308 - 69$ and $308 + 69$, which is between 239 and $377 \times 10^9/l$.

Table 5.5 Output measures from a study of the effectiveness of lisinopril as a prophylactic for acute migraine. Figures are means (SD). Reproduced from *BMJ*, **322**, 19–22, courtesy of BMJ Publishing Group

	Lisinopril	Placebo	Mass % reduction (95% CI)
Primary efficacy parameter			
Hours with headache	129 (125)	162 (142)	20 (5 to 36)
Days with headache	19.7 (14)	23.7 (11)	17 (5 to 30)
Days with migraine	14.5 (11)	18.5 (10)	21 (9 to 34)
Secondary efficacy parameter			
Headache severity index	297 (325)	370 (310)	20 (3 to 37)
Triptan doses	15.7 (15)	20.2 (17)	22 (7 to 38)
Doses of analgesics	14.5 (23)	16.2 (20)	11 (–16 to 37)
Days with sick leave	2.30 (4.32)	2.09 (2.50)	–10 (–64 to 37)
Bodily pain*	63.7 (29)	53.8 (23)	–18 (–35 to –1)
General health*	73.6 (20)	74.1 (21)	1 (–6 to 7)
Vitality*	61.1 (24)	58.2 (21)	–5 (–18 to 8)
Social functioning*	81.4 (25)	79.5 (23)	–2 (–11 to 6)

* From SF-36.

Exercise 5.18 Table 5.5 is from a study of the effectiveness of lisinopril as a prophylactic for acute migraine, in which one group of patients was given lisinopril, and a second group a placebo (Schrader *et al.* 2001). Outcome measures included, ‘hours with headache’, ‘days with headache’ and ‘days with migraine’, all metric continuous variables. The mean and standard deviation for each of these variables for both groups is shown in the figure. Do you think they can be Normally or symmetrically distributed? Explain your answer.

Transforming data

Later in the book you will meet some procedures which require the data to be Normally distributed. But what if it isn’t? Happily some non-Normal data can be *transformed* to make the distribution more Normal (or at least more Normal than it was to start with). The most popular approach is to take the *log* of the data (to base 10); first because it works more often than other procedures, and second because the back-transformation (i.e. anti-logging the results at the end of the analysis) can be meaningfully interpreted.

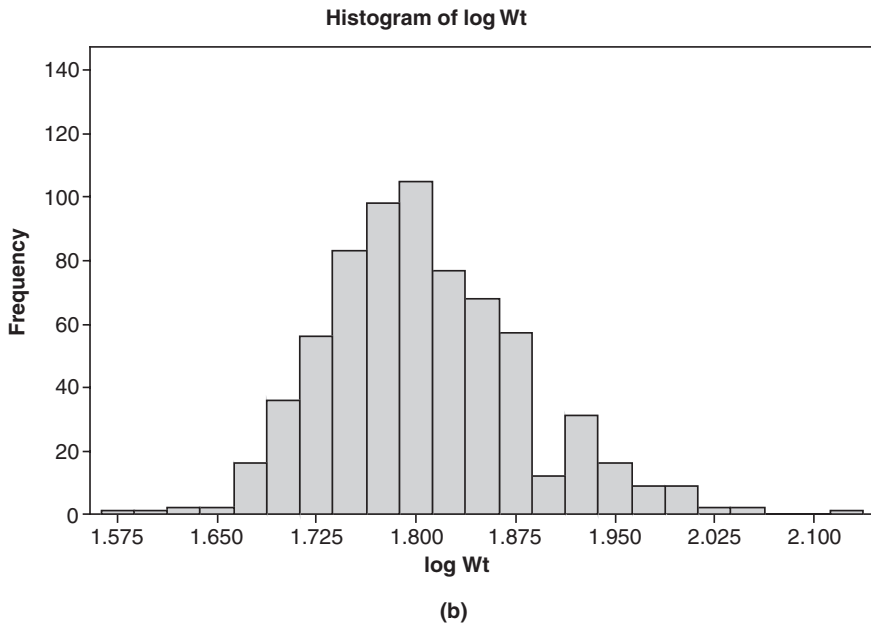
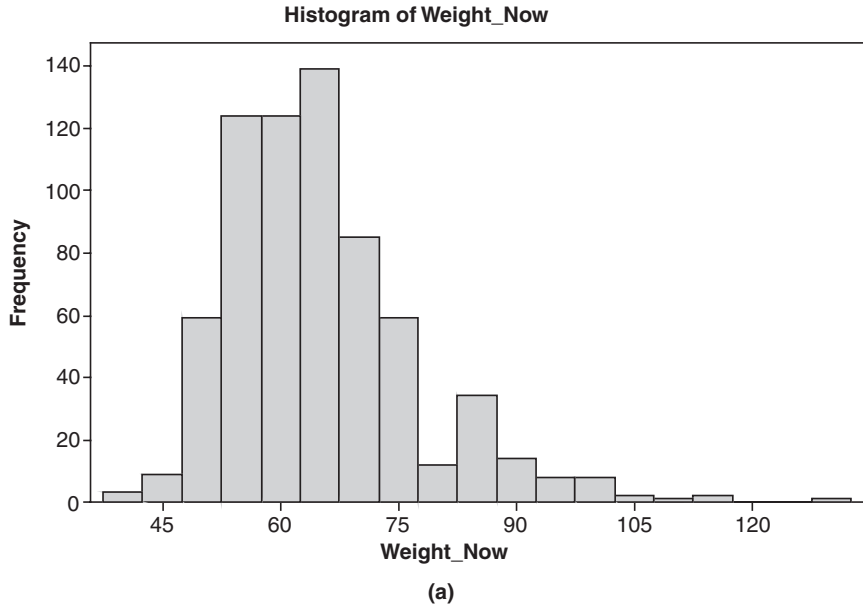


Figure 5.6 The effect of applying a \log_{10} transformation on the shape of the distribution of the weight of 658 women

An example from practice

Figure 5.6 shows histograms for the original and transformed data on the weight (kg) of 685 women in a diet and health cohort study.¹⁰ The original data is positively skewed, Figure 5.6a. If we transform the data by taking \log_{10} , you can see that the transformed data has a more Normal-ish shape, Figure 5.6b.

In Part II, I have discussed ways of looking at sample data – with tables, with charts, from its shape, and with numeric summary measures. Collectively these various procedures are labelled *descriptive statistics*. However, in all of the above, I assumed that you *already* had the data that you were describing, and I've said nothing so far about how you might collect the data in the first place. This is the question I will address in the following chapter.

¹⁰This data was kindly supplied by Professor Janet Cade of Leeds University Medical School.

III

Getting the Data

6

Doing it right first time – designing a study

Learning objectives

When you have finished this chapter you should be able to:

- Explain what a sample is, and what the difference between study and target populations is.
- Explain why it is important for a sample to be as representative of the population from which it is taken as possible.
- Define a random sample, and explain what a sampling frame is.
- Briefly outline what is meant by a contact sample, and by stratified and systematic samples.
- Explain the difference between observational and experimental studies.
- Explain the difference between matched and independent groups.
- Briefly describe case-series, cross-section, cohort and case-control studies, and their limitations and advantages.
- Explain the problem of confounding.

- Outline the general idea of the clinical trial.
- Explain the concept of randomisation, and why it is important, and demonstrate that you can use a random number table to perform a simple block randomisation.
- Describe the concept of blinding, and what it is intended to achieve.
- Outline and compare the design of the parallel and cross-over randomised controlled trials, and summarise their respective advantages and shortcomings.
- Explain what intention-to-treat means.
- Be able to choose an appropriate study design to answer some given research question.

Hey ho! Hey ho! It's off to work we go

There are two main threads here. First, the *study design* question, and second, the *data collection* question. Study design embraces issues like:

- What is the research question? What are we hypothesising?
- Which variables do we need to measure?
- Which is our main *outcome variable* (the variable we are most interested in)?
- How many subjects need to be included in the study?
- Who exactly are the subjects? How should we select them?
- How many groups do we need?
- Are we going to make some form of clinical intervention or simply observe?
- Do we need a comparison group?
- At what stage are we going to take measurements? Before, during, after, etc.?
- How long will the study take? And so on.

Study design is a systematic way of dealing with these issues, and offers a good-practice blueprint that is applicable in almost all research situations.

Second, the *data collection* question. Having decided an appropriate study design, we then have to consider the following:

- How are we going collect the data from the subjects?
- How do we ensure that the sample is as representative as possible?

I want to start with the data collection question. First, though, a brief mention of what we mean by a *population*.

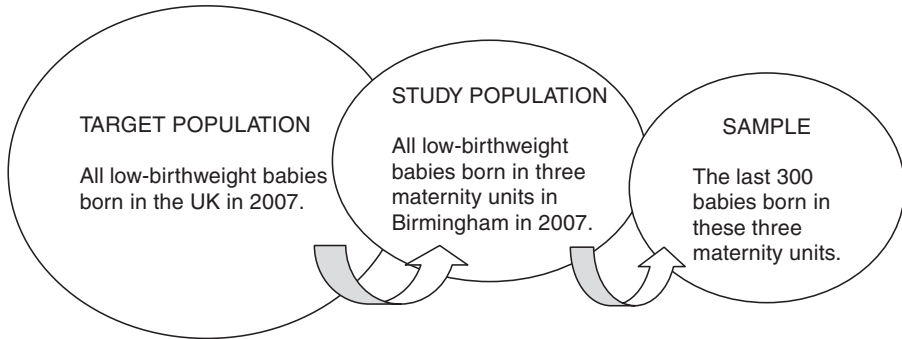


Figure 6.1 The target population, the study population and the sample

Samples and populations

In clinical research, we usually study a sample of individuals who are assumed to be representative of a wider group, to whom (with a good research design and appropriate sampling) the research might apply. This wider group is known as the *target population*, for example ‘all low-birthweight babies born in the UK in 2007’.

It would be impossible to study every single baby in such a large target population (or every member of *any* population). So instead, we might choose to take a sample from a (hopefully) more accessible group. For example, ‘all low-birthweight babies born in three maternity units in Birmingham in 2007’. This more restricted group is the *study population*. Suppose we take as our sample the last 300 babies born in these three maternity units. What we find out from this sample we hope will also be true of the study population, and ultimately of the target population. The degree to which this will be the case depends largely on the representativeness of our sample. These ideas are shown schematically in Figure 6.1. I’ll have more to say about this process in Chapter 7.

Exercise 6.1 Explain the differences between a target population, a study population and a sample. Explain, with an example, why it is almost never possible to study every member of a population.

Sampling error

Needless to say, samples are never perfect replicas of their populations, so when we draw a conclusion about a population based on a sample, there will always be what is known as *sampling error*. For example, if the percentage of women in the UK population with genital chlamydia is 3.50 per cent (we wouldn’t know this of course), and a sample produces a sample percentage of 2.90 per cent, then the difference between these two values, 0.60 per cent, is the sampling error. We can never completely eliminate sampling error, since this is an inherent feature of any sample.

Collecting the data – types of sample

Now the data collection question. There are many books wholly dedicated to the various methods of collecting sample data. I am going to do little more than mention a couple of these methods by name. Those interested in more details of the methods referred to should consult other readily available sources.

The simple random sample and its offspring

The most important consideration is that any sample should be *representative* of the population from which it is taken. For example, if your population has equal numbers of male and female babies, but your sample consists of twice as many male babies as female, then any conclusions you draw are likely to be, at least, misleading. Generally, the most representative sample is a *simple random sample*. The only way that a simple random sample will differ from the population will be *due to chance* alone.

For a sample to be truly random, every member of the population must have an equal chance of being included in the sample. Unfortunately, this is rarely possible in practice, since this would require a complete and up-to-date list (name and contact details) of, for example, *every* low-birthweight baby born in the UK in 2007. Such a list is called a *sampling frame*. In practice, compiling an accurate sampling frame for any population is hardly ever going to be feasible!

This same problem applies also to two close relatives of simple random sampling – *systematic* random sampling, and *stratified* random sampling. In the former, some fixed fraction of the sampling frame is selected, say every 10th or every 50th member, until a sample of the required size is obtained. Provided there are no hidden patterns in the sampling frame, this method will produce samples as representative as a random sample. In stratified sampling, the sampling frame is first broken down into strata relevant to the study, for example men and women; or non-smokers, ex-smokers and smokers. Then each separate stratum is sampled using a systematic sampling approach, and finally these strata samples are combined. But both methods require a sampling frame.

Contact or consecutive samples

The need for an accurate sampling frame makes random sampling impractical in any realistic clinical setting. One common alternative is to take as a sample, individuals in current or recent *contact* with the clinical services, such as consecutive attendees at a clinic. For example, in the study of stress as a risk factor for breast cancer (Table 1.6), the researchers took as their sample 332 women attending a clinic at Leeds General Infirmary for a breast lump biopsy.

Alternatively, researchers may study a group of subjects *in situ*, for example on a ward, or in some other setting. In the nit lotion study (Table 2.1), researchers took as their sample all infested children from a number of Parisian primary schools, based on the high rates of infestation in those same schools the previous year.

If your sample is not a random sample, then the obvious question is, ‘How representative is it of the population?’ And, moreover, which population are we talking about here? In the breast cancer study, if the researchers were confident that their sample of 332 women was

reasonably representative of *all such women* in the Leeds area (their study population), then they would perhaps have felt justified in generalising their findings to this population, and maybe to all women in the UK (a possible target population). But if they knew that the women in their sample were all from a particularly deprived (or particularly affluent) part of the city, or if some ethnic minority formed a noticeably large proportion of the women, then such a generalisation would be more risky.

Exercise 6.2 What is the principal advantage of random sampling? What is the principal drawback of this approach? Describe another method of getting samples that is used in clinical research.

Types of study

With this brief look at the data collection problem, I want to return now to the study design question. Study design divides into two main types. Some alternative ways of classifying these are:

- Observational versus experimental studies.
- Prospective versus retrospective studies.
- Longitudinal versus cross-sectional studies.

I am going to use the first classification, although I will explain the other terms along the way. Broadly speaking, an *observational* study is one in which researchers *actively* observe the subjects involved, perhaps asking questions, or taking some measurements, or looking at clinical records, but they *don't* control, change or effect in any way, their selection, treatment or care. An *experimental* study, on the other hand, does involve some sort of *active* intervention with the subjects. I will first discuss a number of types of observational study designs.

Exercise 6.3 What is the fundamental difference between an observational study and an experimental study?

Observational studies

There are four principal types of observational study:

- Case-series.
- Cross-section studies.
- Cohort studies.
- Case-control studies.

Case-series studies

A health carer may see a series of patients (cases) with similar but unusual symptoms or outcomes, find something interesting and write it up as a study. This is a *case-series*.

An example from practice

In 1981 a drug technician at the Centre for Disease Control in the USA, noticed an unusually high number of requests for the drug pentamidine, used to treat *Pneumocystis carinii* pneumonia (PCP). This led to a scientific report, in effect a case-series study, of PCP occurring unusually in five gay men in Los Angeles. At the same time a similar outbreak of Kaposi's Sarcoma (previously rare except in elderly men) in a small number of young gay men in New York, also began to raise questions. These events signalled the arrival of HIV in the USA.

In the same way, new variant CJD was also first suspected from an unusual series of deaths of young people in the UK, from an apparent dementia-like illness, a disease normally associated with the elderly. Case-series studies often point to a need for further investigations, as was the case in each one of these quoted examples.

Cross-section studies

A cross-section study aims to take a 'snapshot' of some situation at some particular point in time,¹ but notably data on one or more variables from each subject in the study is collected only once.

An example from practice

The following extract is from a cross-section study carried out in 1993 on 2542 rural Chinese subjects, into the relationship between body mass index² and cardiovascular disease, in a rural Chinese population (1st paragraph in text below) (Hu *et al.* 2000). The population of this region of China was about 6 million, and the 2542 individuals included in the sample were selected using a two-stage sampling process, as the 2nd paragraph explains. Each subject was then interviewed and the necessary measurements were taken (3rd paragraph).

A total of 2 542 subjects aged 20–70 years from a rural area of Anqing, China, participated in a **cross-sectional survey**, and 1 610 provided blood samples in 1993. Mean BMI (kg/m²) was 20.7 for men and 20.9 for women. . .

¹ In practice this 'point' in time may in fact be a short-ish period of time.

² Body mass index, used to measure obesity, is equal to a person's weight (kg) divided by their height squared (m)². A bmi of between 20 to 25 is considered 'normal', 25 to 30 indicates a degree of obesity. Higher scores indicate greater levels of obesity.

... These participants were selected from 20 townships in four counties based on a two-stage sampling approach. The sampling unit is a village in the first stage and a nuclear family in the second stage, based on the following criteria: 1) both parents are alive; and 2) there are at least two children in the family. We limited the analysis to 2 542 participants aged 20 years or older from 776 families. . .

... Trained interviewers administered questionnaires to gather information on each participant's date of birth, occupation, education level, current cigarette smoking, and alcohol use. . . measurements, including height and weight, were taken using standard protocols, with subjects not wearing shoes or outer-wear. BMI was calculated as weight (kg)/height (m²). Blood pressure measurements were obtained by trained nurses after subjects had been seated for 10 minutes by using a mercury manometer and appropriately sized cuffs, according to standard protocols.

Note that there is no intervention by the researchers into any aspect of the subjects' care or treatment – the observers only take measurements, ask some questions or study records. The results from the above study showed that subjects in the sample with higher body mass index values were also likely to have higher blood pressures. The researchers might reasonably claim that this link would also exist in the province's population of 6 million – that's their *inference* – but the truth of this would depend on how representative the sample was of the whole Anqing population. Whether or not the finding could be extended to the rest of the diverse Chinese population is more questionable. To sum up, cross-section studies:

- Take only one measurement from each subject at one moment in, or during one period of, time. Data from one or more than one variable may be collected.
- Can be used to investigate a link between two or more variables, but not the *direction* of any causal relationship. The Anqing study does not reveal whether a higher body mass index leads to higher blood pressures (more strain on the heart, for example), or whether higher blood pressures lead to higher body mass index (maybe higher blood pressures increase appetite), it simply establishes some sort of association.
- Are not particularly helpful if the condition being investigated is rare. If, for example, only 0.1 per cent of a population has some particular disease, then a very large sample would be needed to provide any reliable results. Too small a sample might lead you to conclude that nobody in the population had the disease!
- Can be more limited in scope and aim only to *describe* some existing state of affairs, such as the *prevalence* of some condition – for example, the percentage of 16+ UK individuals who have taken ecstasy. Only one variable is measured – use of ecstasy, yes or no. Since this is the only variable measured, no link with any other variable can be explored.
- That aim to uncover attitudes, opinions or behaviours, are often referred to as *surveys*. For example, the views of clinical staff towards having patients' relatives in Emergency Department trauma rooms.

Exercise 6.4 Give two examples of the application of the cross-section design in a clinical setting.

From here to eternity – cohort studies

The main objective of a cohort study is to identify risk factors causing a particular outcome, for example death, or lung cancer, or stroke, or low-birthweight babies and so on. The principle structure of a cohort study (also known as a *follow-up*, *prospective*, or *longitudinal* study) is as follows:

- A group of individuals is selected at random from the general population, for example all women living in Manchester. . .
- . . .or from a particular population, for example all call-centre workers. . .
- . . .or via a clinical setting, for example women diagnosed with breast cancer.
- The group is followed forward over a period of time,³ and the subjects monitored on their exposure to suspected risk factors, or to different clinical interventions.
- At the end of the study, a comparison is made between groups with and without the outcome of interest (say cardio-vascular disease), in terms of their exposure over the course of the study to a suspected risk factor (e.g. smoking, lack of exercise, diet, etc.).
- A reasoned conclusion is drawn about the relationship between the outcome of interest and the suspected risk factor or intervention.

A well-known prospective cohort study was that conducted by Doll and Hill into a possible connection between mortality and cigarette smoking. They recruited about 60 per cent of the doctors in the UK, determined their age and smoking status (among other things), and then followed them up over the ensuing years, recording deaths as they arose. Very quickly the data began to show significantly higher mortality among doctors who smoked.

In some cohort studies, the data may be collected from existing historical records, and subjects followed from some time starting in the past, as the following example demonstrates.

An example from practice

An investigation of the relationship between weight in infancy and the prevalence of coronary heart disease (CHD) in adult life used a sample of 290 men born between 1911 and 1930, and living in Hertfordshire, whose birthweights and weights at one year were on record. In 1994

³ Note that ‘forward’ doesn’t necessarily mean from *today*, although *prospective* cohort studies *do* follow subjects forward from the time the study is initiated.

various measurements were made on the 290 men, including the presence or not of CHD (Fall *et al.* 1995). So ‘forward’ here means from each birth year between 1911 and 1930, up to 1944.

The researchers found that 42 men had CHD, a prevalence of 14 per cent, $(42/290) \times 100$. But weight at *birth* was not influential on adult CHD. However, men who weighed 18 lbs (8.2kg) or less, at *one year*, had almost twice the risk of CHD as men who weighed more than 18 lbs. This of course is only the sample evidence. Whether this finding applies to the population of *all* men born in Hertfordshire during this period, or today, or indeed in the UK, depends on how representative this sample is of either of these populations.

Table 6.1 shows this cohort study expressed as a contingency table (see Chapter 2). The subjects are grouped according to their exposure or non-exposure to the risk factor (in this case weighing 18 lbs or less at one year is taken to be the risk factor), and these groups form the columns of the table. The rows identify the presence or otherwise of the *outcome*, CHD. Clearly this design does suggest (but certainly does not prove) a cause and effect – low weight at one year seems to lead to coronary heart disease in adult life. Cohort studies suffer a number of drawbacks, among which are the following:

- Selection of appropriate subjects may cause difficulties. If subjects are chosen using a contact sample, for example attendees at a clinic, then the outcomes for these individuals may be different from those in the general population.
- If the condition is rare in the population, i.e. has low prevalence, it may require a very large cohort to capture enough cases to make the exercise worthwhile.
- The subjects will have to be followed-up for a long time, possibly many years, before any worthwhile results are obtained. This can be expensive as well as frustrating, and not good if a quick answer is needed. Moreover, this long time-period allows for considerable losses, as subjects drop out for a variety of reasons - they move away, they die from other non-related causes, and so on.
- Over a long period a significant proportion of the subjects may change their habits, quit smoking, for example, or take up regular exercise. However, this problem can be monitored with frequent checks of the state of the cohort.

Table 6.1 The cohort study of weight at one year and its effect on the presence of coronary heart disease (CHD) in adult life, expressed in the form of a contingency table

		Group by exposure to risk factor – weighed \leq 18 lbs at 1 year		Totals
		Yes	No	
Has CHD	Yes	4	38	42
	No	11	237	248
Totals		15	275	290

Finally, note again that the selection of the groups in the cohort contingency table is based on *whether individuals have or have not been exposed to the risk factor*, for example weighing 18 lbs or less at one year (or smoking, or exposure to asbestos, or whatever).

Back to the future – case-control studies

A number of the limitations of the cohort design are addressed by the *case-control* design, although it is itself far from perfect, as you will see. In a cohort study, a group of subjects is followed up to see if they develop an outcome (a condition) of interest. In contrast, in a case-control study the groups are selected on the basis of having or not having the outcome or condition. The objective is the same in both types of study – can the outcome of interest be related to the candidate risk factor? The structure of a case-control study (also known as a *longitudinal* or *retrospective* study) is as follows:

- Two groups of subjects are selected on the basis of whether they have or do not have some condition of interest (for example, sudden infant death, or stroke, or depression, etc.).
- One group, the *cases*, will *have* the condition of interest.
- The other group, the *controls*, will *not* have the condition, but will be as similar to the cases as possible in all other ways.
- Individuals in both groups are then questioned about past exposure to possible risk factors.
- A reasoned conclusion is then drawn about the relationship between the condition in question and exposure to the suspected risk factor.



It was the outcome from such a case-control study by Doll and Hill that led them to conduct the later cohort study referred to above. Before I discuss the case-control design in more detail, there are a couple of important ideas to be dealt with first.

Confounding

Why do we want to ensure that the cases and controls are broadly similar (on age and sex, if nothing else). The reason is that it would be very difficult to identify smoking, say, as a risk factor for lung cancer in the cases, if these were on average twice as old as the controls. Who is to say that it is not increased age that causes a corresponding increased risk of lung cancer and not smoking. Consider the following situation.

Researchers noticed that mothers who smoke more have fewer Down syndrome babies than mothers who smoke less (or don't smoke at all) (Chi-Ling *et al.* 1999). So at first glance smoking less seems to be a risk factor for Down syndrome. It would appear that if a mother wants to reduce the risk of having a baby with Down syndrome she should smoke a lot! However, the fact is that younger mothers have fewer Down syndrome babies but smoke more, while older mothers have more Down syndrome babies but smoke less. Thus the apparent connection between smoking and Down syndrome babies is a mirage. It disappears when we take age into account. We say that age is confounding the relationship between smoking and Down syndrome, i.e. age is a *confounder*.

To be a confounder, a variable must be associated with *both* the risk factor (smoking) *and* the outcome of interest (Down syndrome). Age satisfies this condition since smoking is connected with age, and having a Down syndrome baby is also connected with age. Age is commonly found to be a confounder, as is sex. When we allow for the effects of possible confounders, we are said to be *controlling* or *adjusting* for confounders. Results which are based on unadjusted data are said to be 'crude' results. I'll have more to say about confounding later in the book.

Matching

One way to make cases and controls more similar is to *match* them. How we match cases and controls divides case-control studies into two types – the matched and the unmatched designs. To qualify as a *matched case-control* each control must be *individually* matched (or paired), *person-to-person*, with a case. If cases and controls are independently selected, or are only *broadly* matched (for example, the same *broad mix* of ages, same *proportions* of males and females – known as *frequency matching*), then this is an *unmatched case-control* design. Finally, bear in mind that variables on which the subjects are matched cannot be used to shed any light on the relationship between outcome and risk. For example, if we are interested in coffee as one possible risk factor for people with pancreatic cancer (the cases), we should certainly not match cases and controls so that *both* groups drink lots of coffee.

Unmatched case-control design – an example from practice

In the following extract, from a frequency-matched case-control study into the possible connection between lifelong exercise and stroke (Shinton and Sagar 1993), the authors describe the selection of the cases and the controls.

SUBJECTS

Between 1 October 1989 and 30 September 1990 we recruited men and women who had just had their first stroke and were aged 35–74. The patients were assessed by one of us using the standard criteria (for stroke) of the World Health Organisation.

Control subjects were randomly selected from the general practice population to broadly match the distribution of age and sex among the patients with stroke (frequency matching). All those on the register of the 11 participating practices aged 35–74 were eligible for inclusion. The controls were each sent a letter signed by their general practitioner, which was followed up by a telephone call or visit to arrange an appointment for assessment, usually at their practice surgery.

Table 6.2 Outcome from the exercise and stroke unmatched case-control study for those subjects who had and who had not exercised between the ages of 15 and 25

		Group by disease or condition	
		Cases (stroke)	Controls
Risk factor: exercise	Yes	55	130
undertaken when aged 15–25	No	70	68

The researchers came up with 125 cases with stroke and 198 controls, broadly matched by age and sex. Notice that the numbers of cases and controls need not be the same (and usually aren't). All subjects (or their relatives if necessary), were interviewed and asked about their history of regular vigorous exercise at various times in the past. Table 6.2 shows the results for those subjects who had, and had not, taken exercise between the ages of 15 and 25.

In contrast to cohort studies, in case-control study tables you group by 'has outcome (e.g. disease) or not', for the columns. The rows correspond to whether or not subjects were exposed to the risk factor. From these results you can calculate (you'll see how later) that among those who had had a stroke, the chance that they had exercised in their youth was only about half the chance that somebody without a stroke had exercised. Notice that Table 6.2 is not a contingency table since you now have more than one group, the cases and the controls.

Matched case-control studies

With individuals matched person-to person, you have matched or paired data, which means that the groups of cases and controls are necessarily the same size. Otherwise, the matched design has the same underlying principle as the unmatched design. With individual matching the problem of confounding variables is much reduced. However, one practical difficulty is that it is sometimes quite hard to find a suitable control to match each of the cases on anything more than age and sex.

Comparing cohort and case-control designs

The case-control design has a number of advantages over the cohort study:

- With a cohort study, as you saw above, rare conditions require large samples, but with a case-control study, the availability of potential cases is much greater and sample size can be smaller. Cases will often be contact samples, i.e. selected from patients attending particular clinics.
- Case-control studies are cheaper and easier to conduct.
- Case-control studies give results much more quickly.

But they do have a number of limitations:

- Problems with the selection of suitable control subjects. You want subjects who, apart from not having the condition in question, are otherwise similar to the cases. But such individuals are often not easily found.
- Problems with the selection of cases. One problem is that many conditions vary in their type and nature and it is thus difficult to decide which cases should be included.
- The problem of recall bias. In case-control studies you are asking people to recall events in their past. Memories are not always reliable. Moreover cases may have a better recall of relevant past events than controls – over the years their illness may provide more easily remembered signposts, and they have a better motive for remembering – to get better!

Because of these various difficulties, case-control studies often provide results which seem to conflict with findings of other apparently similar case-control studies. For reliable conclusions, cohort studies are generally preferred – but are not always a practical alternative.

Exercise 6.5 (a) What advantages does a case-control study have over a cohort study?
(b) What are the principal shortcomings of a case-control study?

Getting stuck in – experimental studies

We can now turn to designs, where, in contrast to observational studies, the investigators actively participate in some aspect of the recruitment, treatment or care of the subjects in the study.

Clinical trials

Clinical trials are *experiments* to compare two or more clinical treatments. I use the word ‘treatment’ here, to mean any sort of clinical intervention, from kind words to new drugs. Many books have been written wholly on clinical trials, and I can only touch briefly upon some of the more important aspects of this design. Consider the following imaginary scenario. A new drug, *Arabarb*, has been developed for treating hypertension. You want to investigate its efficacy compared to the existing drug of choice. Here’s what you need to do:

- Decide on an outcome measure – diastolic blood pressure seems a good candidate.
- Select a sample of individuals with hypertension. Divide into two groups (we’ll see how below)
- Ensure that the two groups are as similar as possible. Similar, not only for the obvious variables, such as sex and age, but similar also for other variables whose existence you’re aware of but can’t easily measure. For example, emotional state of mind, lifestyles, genetic differences and so on. But also similar in terms of other variables whose existence you are *not* even aware of.
- Give one group the new drug, Arabarb. This is the *treatment* group.
- Give the other group the existing drug. This is the comparison or *control group*. A control group is imperative. If you have only one group of people, and you measure their diastolic blood pressure before and after they get the Arabarb, you cannot conclude that any decrease in diastolic blood pressure is caused necessarily by the drug. Being in a calm, quiet clinical setting, or having someone fussing over them, might reduce diastolic blood pressure.
- Group similarity is a possible answer to the *confounding* problem. If the groups were *identical* in every respect, the only difference being that one group got Arabarb, while the other got the existing drug, then any *greater* reduction in diastolic blood pressure in the treatment group is likely to be due to the new drug. We know it can’t be due to the fact that the subjects in one group were slightly older, or contained more people who lived alone, or had a greater proportion of males, etc. because we have set out to make the groups identical with respect to these variables. So how do we do this?

Randomisation

The solution is to allocate subjects to one group or the other, using some random procedure. We could toss a coin – heads they go to the treatment group, tails to the control group. This method has the added virtue, not only of making the groups similar, but also of taking the allocation process out of the hands of the researcher. He or she might unconsciously introduce *selection bias* in the allocation, for example by choosing the least well patients for the treatment group. If the *randomisation* is successful, and the original sample is large enough, then the two

groups should be more or less identical, differing *only by chance*. This design is thus called the *randomised controlled trial* (RCT).

Coin tossing is a little impractical of course, and instead a table of *random numbers* (there's one in the Appendix) can be used for the allocation process. Let's see how we might use this method to randomly allocate 12 patients.

You decide to allocate a patient to the treatment group (T), if the random number is *even*, say, and to the control group (C), if *odd*. You then need to determine a starting point in the random number table, maybe by sticking a pin in the table and identifying a start number. Suppose, to keep things simple, you start at the top of column 1 and go down the column; the first six rows contain the values: 23157, 05545, 14871, 38976, 97312, 11742. Combining these three rows gives:

The numbers: 2 3 1 5 7 0 5 5 4 5 1 4
 The allocations: T C C C C T C C T C C T

This gives you four treatment group subjects and eight control group subjects. This is a problem because if possible you want your groups to be the same size. You can fix this with *block randomisation*.

Block randomisation

Here's how it works. You decide on a block size, let's say blocks of four, and write down all combinations that contain *equal* numbers of Cs and Ts. Since there are six such possible combinations, you will have six blocks:

Block1 : CCTT
 Block2 : CTCT
 Block3 : CTTC
 Block4 : TCTC
 Block5 : TCCT
 Block6 : TTCC

With the same random numbers as before, the first number was 2, so the first four subjects are allocated according to Block 2, i.e. CTCT. The next number was 3, so the next four subjects are allocated as Block 3, i.e. CTTC. The next number was 1, giving the allocation CCTT, and so on. Obviously random numbers greater than 6 are ignored. You will end up with the allocation:

CTCT CTTC CCTT

which gives equal numbers, six, in both groups.

Blinding

If at all possible, you don't want the patients to know whether they are in the treatment group or the control groups. This is to avoid the possibility of *response* or *placebo bias*. If a patient knows, or thinks they know, that they are getting the active drug, their psychological response to this knowledge may cause a physical, i.e. a biochemical, response, which conceivably might in turn affect their diastolic blood pressure. In the Arabarb trial, you could achieve this 'blinding' of the patients to their treatment, for example, by giving them all identical tablets, one containing the Arabarb, the other a placebo. This blinding is not always possible. For example, you might be testing out a new walking frame for elderly infirm patients. It will be difficult to disguise this from the older existing frame with which they are all familiar.

A further desirable precaution is also to blind the investigator to the allocation process. If the investigator doesn't know which subject is receiving the drug and which the placebo, their treatment of the subjects will remain impartial and even-handed. Human nature being what it is, there may be an unconscious inclination to treat a patient who is known to be in the treatment group differently to one in the control group. This effect is known as *treatment bias*, and can be avoided by blinding the investigator. We can do this by entrusting a disinterested third party to obtain the random numbers and decide on the allocation rules. Only this person will know which group any given subject is in, and will not reveal this until after the treatment is complete and the results collected and analysed.

Assessment bias can also be overcome by blinding the investigator. This applies to where an *assessment* of some condition after treatment, is required. For example, in trials of a drug to control agitation or anxiety, where proper *measurement* is not possible, then an investigator, knowing that a patient got the active drug, might then judge a patient's condition to be more 'improved', than would an uninvolved outsider, who should thus be involved in the process.

When both subject and investigator are blinded, we refer to the design as a *double-blind randomised controlled trial* – the gold standard among experimental designs. Without blinding the trial is referred to as being *open*. Compared to other designs, the RCT gives the most robust and dependable results.

The design described above, in which two groups receive identical treatment (except for the difference in drugs) throughout the period of the trial, is known as a *parallel* design.

The cross-over randomised controlled trial

A variation on the parallel design is the *cross-over* design, shown schematically in Figure 6.2. In this design one group gets drug A, say, for some fixed period of time and the second group get drug B (or placebo). Then, after a *wash-out* period to prevent drug effect carry-over, the groups are reversed. The group which got drug A now gets drug B, and vice versa, and for the same period of time. Which group gets which treatment first is decided randomly.

The advantage of this method is that each subject gets both treatments, and thus acts as his or her own control. 'Same-subject' matching, if you like. As a consequence of the matched-pair feature, this design requires smaller samples to achieve the same degree of efficiency. Unfortunately, there are a number of problems with this approach.

- A subject may undergo changes between the first treatment period and the second.

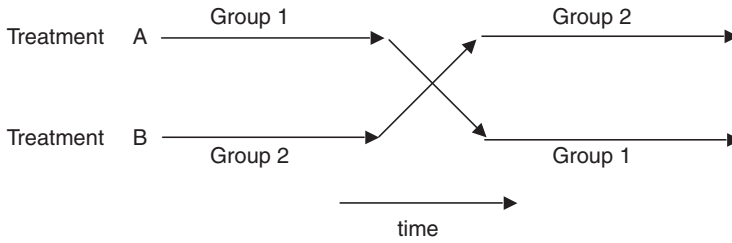


Figure 6.2 Schematic of a cross-over randomised controlled trial

- The method doesn't work well if the drug or treatment to be investigated requires a long time to become effective - for practical reasons cross-over trials are generally of relatively short duration (one reason is to avoid excessive drop out).
- Despite a wash-out interval, there may still be a drug carry-over effect. If carry-over is detected the second half of the trial has to be abandoned.
- The cross-over design is also inappropriate for conditions which can be cured – most of the subjects in the active drug half of the study might be cured by the end of the first period!

An example from practice

The following extract describes the design of a randomised cross-over trial of regular versus as-needed salbutamol in asthma control (Chapman *et al.* 1994).

If inclusion criteria were met at the first clinic visit, patients were enrolled in a four-week randomised crossover assessment of regular vs. as-needed salbutamol. Patients took either 2 puffs (200 mg) metered dose salbutamol from a coded inhaler or matching placebo four times daily for two weeks. On return to the clinic, diary cards were reviewed and patients assigned to receive the crossover treatment for two weeks. During both treatment arms patients carried a salbutamol inhaler for relief of episodic asthma symptoms. Thus, the placebo treatment arm constituted as-needed salbutamol.

Patients were instructed to record their peak expiratory flow rate (PEFR) twice daily: in the early morning and late at night, before inhaler use. Patients also recorded in a diary the number of daytime and night-time asthma episodes suffered and the number of as-needed salbutamol puffs used for symptom relief.

Data from the last eight days of each treatment period were analysed; the first six acted as an active run-in or washout period. Two investigators, blinded to the treatment assignment, examined these comparisons for each patient, and categorised each patient as: showing no difference in asthma control between treatment periods; greater control during the first treatment

period; greater control during the second treatment period; or differences between treatment periods that did not indicate control to be clearly better during either.

Selection of subjects

Just a brief word about selecting subjects for the RCT. Essentially you want a sample of subjects (and they will usually be patients of some sort), who represent a cohesive and clearly defined population. Thus you might want to exclude subjects who, although they have the condition of interest, have a complicated or more advanced form of it, or simultaneously have other significant illnesses or conditions, or are taking drugs for another condition – indeed anything which you feel makes them untypical of the population you have in mind. If your sample is not truly representative of the population you are investigating (a problem known as *selection bias*), then any conclusions you arrive at about your target population are unlikely to be at all reliable.

An example from practice

The following extract is from a RCT to compare the efficacy of having midwives solely manage the care of pregnant Glasgow women, with the more usual arrangements of care being shared between midwife, hospital doctors, and GPs (Turnbull *et al.* 1996). Outcomes were the number of interventions and complications, maternal and fetal outcomes, and maternal satisfaction with the care received. The first paragraph details the selection criteria, the second and third paragraphs describe the random allocation and the blinding processes.

Methods

Design and participants

The study was carried out at Glasgow Royal Maternity Hospital, a major urban teaching hospital with around 5000 deliveries per year, serving a largely disadvantaged community. Between Jan 11, 1993, and Feb 25, 1994, all women booking for routine care at hospital-based consultant clinics were screened for eligibility; the criteria were residence within the hospital's catchment area, booking for antenatal care within 16 completed weeks of pregnancy, and absence of medical or obstetric complications (based on criteria developed by members of the clinical midwifery management team in consultation with obstetricians; available from the MDU).

The women were randomly assigned equally between the two types of care without stratification. A restricted randomisation scheme (random permuted blocks of ten) by random number tables was prepared for each clinic by a clerical officer who was not involved in determining eligibility, administering care, or assessing outcome. The research team telephoned a clerical officer in a separate office for care allocation for each woman.

Women in the control group had no identifying mark on their records, and clinical staff were unaware whether a particular woman was in the control group or was not

in the study. We decided not to identify control women. . .because of concern that the identification of the control group would prompt clinical staff to treat these women differently (i.e., the Hawthorne effect).

Intention-to-treat

One problem that often arises in an RCT, after the randomisation process has taken place, is the loss of subjects, principally through drop-out (moving away, refusing further treatment, dying from non-related causes, etc.), and withdrawal for clinical reasons (perhaps they cannot tolerate the treatment). Unfortunately, such losses may adversely affect the balance of the two groups achieved through randomisation. In these circumstances it is good practice to analyse the data as if the lost subjects were still in the study, as you originally intended – even if all of their measurements are not complete. This is known as *intention-to-treat* analysis. It does, however, require that you have information on the outcome variable for all participants who were originally randomised, even if they didn't complete the course of treatment in the trial. Unfortunately this information is not always available, and in many studies therefore intention-to-treat may be more an aspiration than a reality.

Exercise 6.6 Explain how the possibility of treatment and assessment bias, and response bias, is overcome in the design of a RCT.

Exercise 6.7 (a) What is the principle purpose of randomisation in clinical trials? (b) Using block randomisation, with blocks of four, and a random number table, allocate 40 subjects into two groups, each with 20 individuals.

Exercise 6.8 The following paragraphs contain the stated objective or hypothesis (the wording might have been changed slightly in some cases), in each of a number of recently published clinical research papers. In each case: (a) suggest a suitable outcome variable; (b) suggest an appropriate study design or designs (there's usually more than one way to skin a cat), which would enable the investigators to achieve their stated objective(s); (c) identify possible confounders (if appropriate); (d) comment on the appropriateness of the designs and methods actually chosen by the researchers.

- (a) To determine whether a child's tendency to atopic diseases (asthma, hay fever, eczema, etc.), is affected by the number of siblings that child has.
- (b) To compare two drugs, ciprofloxacin (CF) and pivmecillinam (PM), for the treatment of childhood shigellosis (dysentery).
- (c) To study the effect of maternal chronic hypertension on the risk of small-for-gestational age birthweight.
- (d) To evaluate a possible association between maternal smoking and the birth of a Down syndrome child.

- (e) To compare a community-based service (patients living and treated at home), with a hospital-based service (patients admitted to and treated in hospital), for patients with acute, severe psychiatric illness, with reference to psychiatric outcomes, the burden on relatives and relatives' satisfaction with the service.
- (f) To compare regular with as-needed inhaled salbutamol in asthma control.
- (g) To evaluate the impact of counselling on: client symptomatology, self-esteem and quality of life; drug prescribing; referrals to other mental health professionals; and client and GP satisfaction.

IV

From Little to Large – Statistical Inference

7

From samples to populations – making inferences

Learning objectives

When you have finished this chapter you should be able to:

- Show that you understand the difference, and the connection, between a population parameter and a sample statistic.
- Explain what statistical inference is.
- Explain what an estimate is and why this is unlikely to be exactly the same as the population parameter being estimated.

Statistical inference

You saw in the previous chapter, that when we want to discover things that interest us about a population, we take a sample. We then hope to generalise our sample findings, first to the study population and ultimately to the target population. Statisticians call this process, of generalising from a sample to a population, *statistical inference* or *inferential statistics*.

To take an example (Grun *et al.* 1997): researchers were interested in comparing two methods of screening for genital chlamydia in women attending general practice. Their target population was, ‘all asymptomatic women attending general practice’.¹ Their study population was four

¹ They don’t say whether this is all such women in London, or England, or Wales, or the UK!

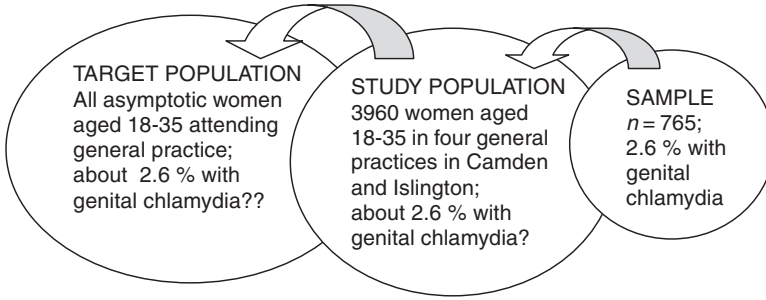


Figure 7.1 The process of statistical inference – from sample to population

general practices in the London Boroughs of Camden and Islington, with a total of 37 000 patients. All women aged between 18 and 35 were invited to take part in the study. A total study population of 3960 women were eligible for inclusion. After exclusions for various reasons, a total sample of 765 women were finally included. As well as the results of their cervical smear for genital chlamydia, data from a brief questionnaire on demographic details, history of urogenital problems and information on sexual history, was also included in the sample data.

The prevalence of genital chlamydia in the sample was found to be 2.6 per cent. The authors might then have inferred from this sample result that the prevalence of genital chlamydia in the study population of 3960 women in the four practices, was also *about* 2.6 per cent. And by extension, was also true of the target population of all asymptomatic women attending general practice.

The accuracy of this *estimate* would depend on how typical the 765 women in the sample were of all the 3960 women in the study population, and in turn how typical these women were of all the women in the target population – all women 18–35 in the UK attending GP practice. This particular statistical inference process is illustrated in Figure 7.1.

I have used the word ‘estimate’² here deliberately, because the value you get from your sample (from *any* sample) is never going to be exactly the same as the population value. You have to accept that the percentage with genital chlamydia in the population is probably *around* 2.6 per cent, *give or take a bit*. The size of the ‘bit’ depends on how similar your sample is to its population – and on sampling error. I’ll have a lot more to say on this later in the book.

For the moment, the meaning of a few terms. The feature or characteristic of a population whose value you want to determine is known as a *population parameter*. For example, the mean or the median of some variable in a population are both population parameters. In the genital chlamydia example, the population parameter you want to estimate is the *percentage* with genital chlamydia.

The value that you get from your sample, in this case the *sample percentage* with genital chlamydia (on which you are going to base your estimate of the population value) is called the *sample statistic*. This is why we are so interested in the summary descriptive measures, such as the sample mean and the sample median, described in Chapter 6. In other words, you can use the sample mean, for example, to estimate the population mean, the sample median to estimate the population median and so on.

² An *estimate* is just a fancy word for an informed guess.

Actually, estimation is not the only way of making inferences about population parameter values. An alternative approach is to hypothesise that a population parameter has a particular value, and then see if the value of the corresponding sample statistic is compatible with your hypothesis. This approach is called hypothesis testing. In Chapters 9 to 11, I am going to discuss some common estimation procedures and in Chapters 12 to 14, I will discuss the alternative hypothesis test approach. First, however, I need to say a few words on probability, and some other related stuff; this I will do in the next chapter.

Exercise 7.1 (a) Explain the meaning of and the difference between a population parameter and a sample statistic. (b) Why is a sample, however well chosen, never going to be *exactly* representative of the sampled population? (c) Give a couple of examples that illustrate the difference between a target and a study population?

Exercise 7.2 Give a few reasons why women aged 18–35 in the London boroughs of Camden and Islington may not be typical of all women in London, or of all women in the UK.

8

Probability, risk and odds

Learning objectives

When you have finished this chapter you should be able to:

- Define probability, explain what an event is and calculate simple probabilities.
- Explain the proportional frequency approach to calculating probability.
- Explain how probability can be used with the area properties of the Normal distribution.
- Define and explain the idea of risk and its relationship with probability.
- Calculate the risk of some outcome from a contingency table and interpret the result.
- Define and explain the idea of odds.
- Calculate odds from a case-control 2×2 table and interpret the result.
- State the equation linking probability and odds and be able to calculate one given the other.
- Explain what the risk ratio of some outcome is, calculate a risk ratio and interpret the result.
- Explain what the odds ratio for some outcome is, calculate an odds ratio and interpret the result.

- Explain why it's not possible to calculate a risk ratio in a case-control study.
- Define number needed to treat, explain its use and calculate NNT in a simple example.

Chance would be a fine thing – the idea of probability

Probability is a measure of the chance of getting some outcome of interest from some event. The event might be rolling a dice and the outcome of interest might be getting a six; or the event might be performing a biopsy with the outcome of interest being evidence of malignancy and so on. Some basic ideas about probability:

- The probability of a particular outcome from an event will lie between zero and one.
- The probability of an event that is certain to happen is equal to one. For example, the probability that everybody dies eventually.
- The probability of an event that is impossible is zero. For example, throwing a seven with a normal dice.
- If an event has as much chance of happening as of not happening (like tossing a coin and getting a head), then it has a probability of $\frac{1}{2}$ or 0.5.
- If the probability of an event happening is p , then the probability of the event *not* happening is $1 - p$.



Table 8.1 Frequency table showing causes of blunt injury to limbs in 75 patients

Cause of injury	Frequency (number of patients) $n = 75$	Proportional frequency
Falls	46	0.613
Crush	20	0.267
Motor vehicle crash	6	0.080
Other	3	0.040

$$46/75 = 0.613$$

Calculating probability

You can calculate the probability of a particular outcome from an event with the following expression:

The probability of a particular outcome from an event is equal to the number of outcomes that favour that event, divided by the *total* number of possible outcomes.

To take a simple example: What is the probability of getting an even number when you roll a dice?

Total number of possible outcomes = 6 (1 or 2 or 3 or 4 or 5 or 6)

Total number of outcomes favouring the event 'an even number' = 3 (i.e. 2 or 4 or 6)

So probability of getting an even number = $3/6 = 1/2 = 0.5$

The above method for determining probability works well with experiments where all of the outcomes have the same probability, e.g. rolling dice, tossing a coin, etc. In the real world you will often have to use what is called the *proportional frequency* approach, which uses existing frequency data as the basis for probability calculations.

As an example, look at Table 8.1 (which is Table 2.3 reproduced for convenience) which shows the causes of blunt injury to limbs. I have added an extra column showing the *proportional* frequency (category frequency divided by total frequency). Notice that the proportional frequencies sum to one.

Exercise 8.1 Table 1.6 shows the basic characteristics of the two groups of women receiving a breast lump diagnosis in the stress and breast cancer study. What is the probability that a woman chosen at random: (a) will have had her breast lump diagnosed as (i) benign? (ii) malignant?; (b) will be post-menopausal?; (c) will have had three or more children?

Exercise 8.2 Table 1.7 is from a study of thrombotic risk during pregnancy. What is the probability (under classification 1) that a subject chosen at random will be aged: (a) less than 30?; (b) more than 29?

Now ask the question, ‘What is the probability that if you chose one of these 75 patients at random their injury will have been caused by a fall?’. The answer is the proportional frequency for the ‘fall’ category, i.e. 0.613. In other words, we can interpret proportions as equivalent to probabilities. Probability is a huge subject with many textbooks devoted to it, but for our purposes in this book we don’t really need to know any more.

Probability and the Normal distribution

We know that if data is Normally distributed then about 95 per cent of the values will lie no further than two standard deviations from the mean (see Figure 5.5). In probability terms, we can say that there is a probability of 0.95 that a single value chosen at random will lie no further than two standard deviations from the mean. In the case of the Normally distributed birthweight data, this means that there is a probability of 0.95 that the birthweight of one of these infants chosen at random will be between 2890 g and 4398 g.

Exercise 8.3 Using the information on cord platelet count in Figure 4.6, determine the probability that one infant chosen at random from this sample will have a cord platelet count: (a) between $101 \times 10^9/l$ and $515 \times 10^9/l$; (b) less than $239 \times 10^9/l$.

Risk

As I mentioned earlier a *risk* is the same as a probability, but the former word tends to be favoured in the clinical arena. So the definition of probability given earlier applies equally here to risk. In other words, the risk of any particular outcome from an event is equal to the number of favourable outcomes divided by the total number of outcomes. Risk accordingly can vary between zero and one.

As an example, and also to re-visit the contingency table, look again at the table in Table 6.1 from the cohort study of coronary heart disease (CHD) in adult life and the risk factor ‘weighing 18 lbs or less at one year’. The risk (or probability) that those adults who as infants weighed 18 lbs or less at one year will have CHD, is equal to the number who weighed 18 lbs or less at one year and had CHD, divided by the total number who weighed 18 lbs or less. This is equal to $4/15 = 0.2667$.

Similarly, the risk (or probability) for those who weighed more than 18 lbs at one year will have CHD equals the number who weighed more than 18 lbs at one year and had CHD, divided by the total number who weighed more than 18 lbs. This is equal to $38/275 = 0.1382$ and thus is only half the risk of those weighing 18 lbs or less.

The risk for a single group, as it is described it above, is also known as the *absolute risk*, mainly to distinguish it from *relative risk*, which is the risk for one group *compared* to the risk for some other group (which we’ll come to shortly).

Table 8.2 The distribution of alcohol intake and deaths by sex and level of alcohol intake. Reproduced from *BMJ*, 308, 302–6, courtesy of BMJ Publishing Group

Alcohol intake (beverages a week)*	Men		Women	
	No of subjects	No (%) of deaths	No of subjects	No (%) of deaths
<1	625	195 (31.2)	2472	394 (15.9)
1–6	1183	252 (21.3)	3079	283 (9.2)
7–13	1825	383 (21.0)	1019	96 (9.4)
14–27	1234	285 (23.1)	543	46 (8.5)
28–41	585	118 (20.2)	72	6 (8.3)
42–69	388	99 (25.5)	29	5 (17.2)
> 69	211	66 (31.3)	20	1 (5.0)
Total	6051	1398 (23.1)	7234	831 (11.5)

*One beverage contains 9–13 g alcohol.

Exercise 8.4 Table 8.2 is from a cohort study into the influence of sex, age, body mass index and smoking on alcohol intake and mortality in Danish men and women aged between 30 and 79 years (Gronbaek *et al.* 1994). The table shows the distribution of alcohol intake and deaths by sex and level of alcohol intake. Use the information in the table to construct an appropriate contingency table for: (a) men; (b) women. Calculate the absolute risk of death among those subjects who consume: (i) less than one beverage a week; (ii) more than 69 beverages a week. Interpret your results.

Odds

The *odds* for a particular outcome from an event is closely related to probability, is perhaps a more difficult concept, but important in medical statistics, and we will meet it again later in the book. As you saw above, the probability (or risk) of a particular outcome from an event is the number of outcomes favourable to the event divided by the *total* number of outcomes. But:

The *odds* for an event is equal to the number of outcomes favourable to the event divided by the number of outcomes not favourable to the event.

Notice that:

- The value of the odds for an outcome can vary from zero to infinity.
- When the odds for an outcome are less than one, the odds are *unfavourable* to the outcome; the outcome is *less* likely to happen than it is *to* happen.
- When the odds are equal to one, the outcome is as likely to happen as not.

- When the odds are greater than one, the odds are *favourable* to the outcome; the outcome is *more* likely to happen than not.

Let's go back to the dice rolling game. The *odds* in favour of the outcome 'an even number', is the number of outcomes favourable to the event (the number of *even* numbers, i.e. 2, 4, 6), divided by the number of outcomes not favourable to the event (the number of *not even* numbers, i.e. 1, 3, 5), which is $3/3 = 1/1$ or one to one.

So the odds of getting an even number are the same as the odds of getting an odd number. Nearly all the odds in health statistics are expressed as 'something' to one. We call this value of one the *reference value*.

As a further more relevant example, we can also calculate odds from a table such as that for the exercise and stroke case-control study in Table 6.2. For instance:

- Among those patients who'd *had* a stroke, 55 had exercised (been exposed to the 'risk' of exercising) and 70 had not, so the odds that those with a stroke had exercised is $55/70 = 0.7857$.
- Among those patients who *hadn't* had a stroke, 130 had exercised and 68 had not, so the odds that they had exercised is $130/68 = 1.9118$.

In other words, among those who'd had a stroke, the odds that they had exercised was less than half the odds ($0.7857/1.9118$) of those who hadn't had a stroke. We can conclude on the basis of this sample that exercise when young seems to confer protection against a stroke.

Exercise 8.5 Table 8.3 is from a matched case-control study into maternal smoking during pregnancy and Down syndrome (Chi-Ling *et al.* 1999). It shows the basic characteristics of mothers giving birth to babies with Down syndrome (cases), and without Down syndrome (controls). Use the information in the table to construct appropriate separate 2×2 contingency tables for women: (a) aged under 35; (b) aged 35 and over. Hence calculate the odds that they had smoked during pregnancy among mothers giving birth to: (i) a Down syndrome baby; (ii) a healthy baby. What do you conclude?

Why you can't calculate risk in a case-control study

For most people the *risk* of an event, being akin to probability, makes more sense and is easier to interpret than the odds for that same event. That being so, maybe it would be more helpful to express the stroke/exercise result as a risk rather than as odds. Unfortunately we can't, and here's why.

To calculate the risk that those with a stroke had exercised, you need to know two things: the total number who'd had a stroke, and the number of these who had been exposed to the risk (of exercise). You then divide the latter by the former. In a cohort study you would select the groups on this basis – whether they had been exposed to the risk (of exercising) or not. So one group would contain individuals exposed to the risk and the other those not exposed.

Table 8.3 Basic characteristics of mothers in a case-control study of maternal smoking and Down syndrome. Reproduced from *Amer. J. Epid.*, **149**, 442–6, courtesy of Oxford University Press

	Cases (n = 775)		Controls (n = 7750)	
	No.	%	No.	%
Selected characteristics of Down syndrome cases and birth-matched controls. Washington State, 1984–1994				
Smoking during pregnancy				
Age < 35 years				
Yes	112	20.0	1411	20.2
No	421	75.0	5214	74.6
Unknown	28	5.0	363	5.2
Aged ≥ 35 years				
Yes	15	7.0	108	14.2
No	186	86.9	611	80.2
Unknown	13	6.1	43	5.6

But in a case-control study you don't select on the basis of whether people have been exposed to the risk or not, but on the basis of whether they have some condition (a stroke) or not. So you have one group composed of individuals who have had a stroke, and one group who haven't, but *both* groups will contain individuals who were and were not exposed to the risk (of exercising). Moreover, you can select whatever number of cases and controls you want. You could for example halve the number of cases and double the number of controls. This means the column totals, which you would otherwise need for your risk calculation, are meaningless.

The link between probability and odds

The connection between probability (risk) and odds means that it is possible to derive one from another:

$$\text{risk or probability} = \text{odds} / (1 + \text{odds})$$

$$\text{odds} = \text{probability} / (1 - \text{probability})$$

Exercise 8.6 Following on from Exercise 8.5, what is the probability that a mother chosen at random from those aged ≥ 35 , will have smoked during pregnancy if they are: (a) mothers of Down syndrome babies; (b) mothers of healthy babies?

Table 8.4 Generalised contingency table for risk ratio calculations in a cohort study

		Group by exposed to risk factor		
		Yes	No	Totals
Outcome: has disease	Yes	a	b	$(a + b)$
	No	c	d	$(c + d)$
	Totals	$(a + c)$	$(b + d)$	

The risk ratio

In practice, risks and odds for a single group are not nearly as interesting as a *comparison* of risks and odds between *two* groups. For risk you can make these comparisons by dividing the risk for one group (usually the group exposed to the risk factor) by the risk for the second, non-exposed, group. This gives us the *risk ratio*.¹ Let's calculate the risk ratio for the data in Table 6.1, from the cohort study of coronary heart disease (CHD) in adult life and weighing 18 lbs or less at one year, using the results obtained on page 100:

Among those weighing 18 lbs or less at one year, the risk of CHD = 0.2667

Among those weighing more than 18 lbs at one year, the risk of CHD = 0.1382

So the risk *ratio* for CHD among those weighing 18 lbs or less at one year compared to those weighing more than 18 lbs = $0.2667/0.1382 = 1.9298$. We interpret this result as follows: adults who weighed 18 lbs or less at one year old have nearly twice the risk of CHD as those who weighed more than 18 lbs.

We can generalise the risk ratio calculation with the help of the contingency table as in Table 8.4, where the cell values are represented as a , b , c and d .

- Among those exposed to the risk factor, the risk of disease = $a/(a + c)$.
- Among those not exposed, the risk of disease = $b/(b + d)$.
- Therefore : risk ratio = $\frac{a}{(a+c)} / \frac{b}{(b+d)} = \frac{a(b+d)}{b(a+c)}$

Exercise 8.7 Use the results you obtained in Exercise 8.4 to calculate the risk ratio of death for those who consumed more than 69 beverages a week, compared to those who consumed less than one beverage per week (which we'll define as the reference group), for: (a) men; (b) women. Interpret your results.

¹ Risk ratio is also commonly known as *relative risk*.

Table 8.5 Generalised 2×2 table for odds ratio calculations in a case-control study

		Group by outcome (e.g. disease)	
		Cases	Controls
Exposed to risk factor?	Yes	a	b
	No	c	d

The odds ratio

With a case-control study you can compare the odds that those with a disease will have been exposed to the risk factor, with the odds that those who don't have the disease will have been exposed. If you divide the former by the latter you get the *odds ratio*.

On p. 102 you calculated the following odds for the stroke and exercise study (where we are treating exercise as the risk factor): the odds that those with a stroke had exercised = $55/70 = 0.7857$; and the odds that those without a stroke had exercised = $130/68 = 1.9118$. Dividing the former by the latter, you get the odds ratio = $0.7857/1.9118 = 0.4110$. This result suggests that those with a stroke are less than half as likely to have exercised when young as the healthy controls. It would seem that exercise is a *beneficial* 'risk' factor. We can generalise the odds ratio calculation with the help of the 2×2 table in Table 8.5.

- The odds of exposure to the risk factor among those with the disease = a/c ,
- The odds of exposure to the risk factor among the healthy controls = b/d .
- Therefore: odds ratio = $\frac{a/c}{b/d} = ad/bc$.

Exercise 8.8 Use the results from Exercise 8.5 to calculate the odds ratio for smoking among the mothers of Down syndrome babies compared to mothers of healthy babies, for: (a) mothers aged under 35; (b) mothers aged 35 and over. Interpret your results.

Remember that the risk ratios and odds ratios in the coronary heart disease and in the stroke examples above are *sample* risk and odds ratios. For instance, from the *sample* risk ratio of 1.928 in the CHD/weight at one year study, you can infer that the *population* risk ratio is also *about* $1.93 \pm$ a 'bit'. But how big is this 'bit', how precise is your estimate? This is a question I'll address in Chapter 11.

Finally, I mentioned earlier that most people are happier with the concept of 'risk' than with 'odds', but that you can't calculate risk in a case-control study. However, there is a happy ending. The odds ratio in a case-control study is a reasonably good estimator of the

equivalent risk ratio, so you can at least approximate its value with the corresponding odds ratio.

Number needed to treat (NNT)

This seems as good a time as any to discuss a measure of the effectiveness of a clinical procedure which is related to risk; more precisely, to absolute risk. This is the *number needed to treat*, or NNT. NNT is the number of patients who would need to be treated with the active procedure, rather than a placebo (or alternative procedure), in order to reduce by one the number of patients experiencing the condition.

To explain NNT let's go back to the example for weighing 18 lbs or less at one year as a risk factor for coronary heart disease (CHD). The absolute risk of CHD among those weighing 18 lbs or less was 0.2667. The absolute risk of CHD for those weighing more than 18 lbs was 0.1382.

We need now to define the *absolute risk reduction* or ARR as the difference between two absolute risks. So in this example, the absolute risk reduction is the difference in these two absolute risks – the reduction in risk gained by weighing more than 18 lbs at one year rather than weighing 18 lbs or less. In this case:

$$\text{ARR} = 0.2667 - 0.1382 = 0.1285$$

Now the number needed to treat is defined as follows: $\text{NNT} = 1/\text{ARR}$

Thus in this case: $\text{NNT} = 1/0.1285 = 7.78$

In other words, if you had some treatment (infant-care advice for vulnerable parents, for example), which would cause infants who would otherwise have weighed less than 18 lbs at one year to weigh 18 lbs or more, then you would need to 'treat' eight infants (or their parents) to ensure that one of these infants did not develop coronary heart disease when an adult.² NNT is often used to give a familiar and practical meaning to outcomes from clinical trials and systematic reviews,³ where measures of risk, and risk ratios, may be difficult to translate into the potential benefit to patients.

An example from practice

Table 8.6 is from the follow-up (cohort) study into the effectiveness of carotid endarterectomy in ipsilateral stroke prevention first referred to in Figure 3.2 (Inzitari *et al.* 2000). The table shows that for any stroke, the (absolute) risk if treated medically is 0.110 (11.0 per cent), and if treated surgically is 0.051 (5.1 per cent). The reduction in absolute risk, $\text{ARR} = 0.110 - 0.051 = 0.059$ (5.9 per cent). So $\text{NNT} = 1/0.059 = 16.95$ or 17, at five years. In other words, 17 patients would have to be treated with carotid endarterectomy to prevent one patient from having a stroke within five years who, without the treatment, would otherwise have done so.

² The number must always be rounded up.

³ Systematic review is the systematic collection of all the results from as many similarly-designed studies as possible dealing with the same clinical problem. I discuss this procedure in Chapter 20.

Table 8.6 Example of numbers needed to treat (NNT), at five years and two years from a follow-up (cohort) study into the effectiveness of carotid endarterectomy in stroke prevention. Reproduced from *NEJM*, **342**, 1693–9, by permission of Massachusetts Medical Society

Cause	Medically Treated Group	Surgically Treated Group	Reduction in Risk	Absolute Difference in Risk	No. Needed to Treat*	
					at 5 yr	at 2 yr
Any stroke [†]	11.0	5.1	54	5.9	17	67
Large-artery stroke [‡]	6.6	3.1	54	3.5	29	111

*The number needed to treat is calculated as the reciprocal of the difference in risk. At two years, the number needed to treat is based on estimated differences in risk of 1.5 percent for stroke of any cause and 0.9 percent for large-artery stroke.

[†]The risk of stroke from any cause in the medical and surgical groups in the Asymptomatic Carotid Atherosclerosis Study is shown.

[‡]The estimates of the risk of large-artery stroke were based on the observations that for subjects in the NASCET with 60 to 99 percent stenosis, the ratio of the risk of large-artery stroke to the risk of stroke from any cause in the territory of a symptomatic artery was similar in the medically and surgically treated subjects, and the risk of large-artery stroke was approximately 60 percent of the risk of stroke from any cause in the territory of an asymptomatic artery (i.e., 6.6 percent = 60 percent of 11.0 percent, and 3.1 percent = 60 percent of 5.1 percent).

Exercise 8.9 In a cohort study of a possible connection between dental disease and coronary heart disease (CHD), subjects were tracked for 14 years (deStefano *et al.*). Of 3542 subjects with no dental disease, 92 died from CHD, while of 1786 subjects with periodontitis, 151 died from CHD. How many people must be successfully treated for periodontitis to prevent one person dying from CHD?

V

**The Informed Guess –
Confidence Interval Estimation**

9

Estimating the value of a *single* population parameter – the idea of confidence intervals

Learning objectives

When you have finished this chapter you should be able to:

- Describe the sampling distribution of the sample mean and the characteristics of its distribution.
- Explain what the standard error of the sample mean is and calculate its value.
- Explain how you can use the probability properties of the Normal distribution to measure the preciseness of the sample mean as an estimator of the population mean.
- Derive an expression for the confidence interval of the population mean.
- Calculate and interpret a 95 per cent confidence interval for a population mean.
- Calculate and interpret a 95 per cent confidence interval for a population proportion.
- Explain and interpret a 95 per cent confidence interval for a population median.

Confidence interval estimation for a population mean

You saw at the beginning of Chapter 6 that we can use a sample statistic to make an informed guess, or *estimate*, of the value of the corresponding *population* parameter. For example, the sample mean birthweight for the 30 infants in Table 2.5 was 3644.4 g, so you can estimate the population mean birthweight of *all* infants of whom this sample is representative, also to be *about* 3644 g,¹ plus or minus some (hopefully) small random or *sampling error*. The obvious questions are:

- How small is this ‘plus or minus’ bit?
- Can it be *quantified*?
- Can we establish how *precise* our *sample* mean birthweight is as an estimate of population mean birthweight?
- How close to a population mean can you expect any given sample mean to be?

As you can see these are all essentially the same question, ‘How big an *error* might we be making when we use the sample mean as an estimate of the population mean?’. This question can be answered with what is known as a *confidence interval estimator*, which is a numeric expression that quantifies the likely size of the sampling error. But to get a confidence interval we need first to introduce an important concept in statistical inference – the *standard error*.

The standard error of the mean

Our sample of 30 infants produced a sample mean birthweight of 3644.4 g. You could take a second, different, sample of 30 infants from the same population, and this sample would produce a different value for the sample mean. And a third sample, and a fourth and so on. In fact from any realistic population you could (*in theory*), take a huge number of different same-size samples, each of which would produce a different sample mean. You would end up with a large number of sample means, and if you were to arrange all of these sample means into a frequency curve, you would find:

- That it was Normal. This Normal-ness of the distribution of sample means is a very useful quality (to say the least); we will depend on it a lot in what is to come.
- That it was centred around the true population mean. In other words, the mean of all possible sample means is the same as the population mean.

This is very re-assuring. It means that, *on average*, the sample mean estimates the population mean exactly. But note the ‘on average’. A particular *single* sample mean may still be some distance from the true mean.

¹ The value of the sample mean of 3644.4g is known as the *point estimate* of the population mean. It’s the *single best guess* you could make as to the value of the population mean.

We can measure the spread of all of these different sample means in the usual way - with the standard deviation. However, to distinguish it from the spread of values in a *single* sample, we call it the *standard error*.² It is usually abbreviated as $s.e.(\bar{x})$, where the symbol \bar{x} stands for the sample mean. Remember that the standard deviation is a measure of the spread of the data in a *single* sample. The standard error is a measure of the spread in *all* (same-size) sample means from a population.

We can very easily *estimate* the standard error with the equation: $s.e.(\bar{x}) = s/\sqrt{n}$. Here s is the sample standard deviation and n is the sample size. Notice that as the sample size n increases, the standard error decreases. In other words, the bigger the sample, the smaller the error in our estimate of population mean. Intuitively this feels right.

For example, if we took a sample of size $n = 100$ from a population, and measured systolic blood pressure, and obtained a sample mean of 135 mmHg and a sample standard deviation of 3 mmHg, then the estimated standard error would be:

$$s.e.(\bar{X}) = 3/\sqrt{100} = 3/10 = 0.33 \text{ mmHg}$$

Since the distribution of sample means is Normal, we can make use of the area properties of the Normal distribution (see Figure 5.5). If the sample standard deviation is 3 mmHg and sample size $n = 100$, then the standard error = 0.33 mmHg. Because the distribution of sample means is Normal, this means that about 95 per cent of sample means will lie within plus or minus two standard errors of the population mean. That is within plus or minus 0.66 mmHg of the population mean. In other words there's a pretty good chance (a probability of 0.95 in fact) that any single sample mean will be no further than 0.66 mmHg from the (unknown) population mean.

The above discussion about taking lots of different samples from a population is entirely theoretical. In practice, you will usually only get to take *one* sample from a population, the value of whose mean you will never know. To sum up, the standard error is a measure of the preciseness of the sample mean as an estimator of the population mean. Smaller is better. If you are comparing the precision of two different sample means as estimates of a population mean, the sample mean with the smallest standard error is likely to be the more precise.

Exercise 9.1 A team of researchers used a cohort study to investigate the intake of vitamins E and C and the risk of lung cancer, 19 years into the study (Yong *et al.* 1997). They calculated the mean (and the standard error) intake of vitamins E and C, of individuals with and without lung cancer (cases and non-cases respectively). These were:

Vitamin E.	Cases: 6.03 mg (0.35 mg);	non-cases: 6.30 mg (0.05 mg).
Vitamin C.	Cases: 64.18 mg (5.06 mg);	non-cases: 82.21 mg (0.80 mg).

How would you interpret these results in terms of the likely precision of each of the sample means as estimators of their respective population means?

² To give it its full name, the *standard error of the sampling distribution of the sample mean* (quite a mouthful), but thankfully, it is usually just called the *standard error*.

How we use the standard error of the mean to calculate a confidence interval for a population mean

With the standard error under our belt we can now get to grips with the confidence interval. You have seen that we can be 95 per cent confident that any sample mean is going to be within plus or minus two standard errors of the population mean.³ From this we can show that:

$$\text{Population mean} = \text{sample mean} \pm 2 \times \text{standard error}$$

That is:

- We can be 95 per cent confident that the interval, from the sample mean $- 2 \times$ standard error, to the sample mean $+ 2 \times$ standard error, will include the population mean.
- Or in probability terms, there is a probability of 0.95 that the interval from the sample mean $- 2 \times$ standard error, to the sample mean $+ 2 \times$ standard error, will contain the population mean.

In other words, if you pick one out of all the possible sample means at random, there is a probability of 0.95 that it will lie within two standard errors of the population mean. We call the distance from the sample mean $- 2 \times \text{s.e.}(\bar{x})$, to the sample mean $+ 2 \times \text{s.e.}(\bar{x})$, the *confidence interval*.

The above result means that you now quantify just how close a sample mean is likely to be to the population mean. For obvious reasons the value you get when you put some figures into this expression is known as the *95 per cent confidence interval estimate* of the population mean. A 95 per cent *confidence level* is most common, but 99 per cent confidence intervals are also used on occasion. Note that the confidence interval is sometimes said to represent a *plausible range of values* for the population parameter.

A worked example from practice

In the cord-platelet count histogram in Figure 4.6, the mean cord platelet count in a sample of 4382 infants is $306 \times 10^9/l$, and the standard deviation is $69 \times 10^9/l$, so the standard error of the mean is:

$$\text{s.e.}(\bar{X}) = 69 \times 10^9 / \sqrt{4382} = 1.042 \times 10^9 / l$$

³ I have used the value two in all of these expressions as a convenient *approximation* to the exact value (which in any case will be very close to two, when the probability is 0.95). The exact value comes from what is known as the *t distribution*. The *t* distribution is similar to the Normal distribution, but for small sample sizes is slightly wider and flatter. It is used instead of the Normal distribution for reasons connected to inferences about the population standard deviation, which we don't need to go into here. Anyway, in practice you will use a computer to obtain your confidence interval result. This will use the proper value.

Therefore the 95 per cent confidence interval for the population mean cord platelet count is:

$$(306 - 2 \times 1.042 \text{ to } 306 + 2 \times 1.042) \text{ g or } (303.916 \text{ to } 308.084) \times 10^9 / l$$

Which we can interpret as follows: we can be 95 per cent confident that the population mean cord platelet count is between $303.916 \times 10^9/l$ and $308.084 \times 10^9/l$, or alternatively that there's a probability of 0.95 that the interval from 303.916 to 308.084 will contain the population mean value. Of course there's also a 5 per cent chance (or a 0.05 probability), that it will not!

Alternatively we can say that the interval $(303.916 \text{ to } 308.084) \times 10^9/l$ represents a *plausible range of values* for the population mean cord platelet count. The narrower the confidence interval the more precise is the estimator. In the cord platelet example, the small width, and therefore high precision of the confidence interval, is due to the large sample. By the way, it's good practice to put the confidence interval in brackets and use the 'to' in the middle and not a '-' sign, since this may be confusing if the confidence interval has a negative value(s).

Exercise 9.2 Use the summary age measures given in Table 1.6 for the life events and breast cancer study, to calculate the standard error and the 95 per cent confidence intervals for population mean age of: (a) the cases; (b) the controls. Interpret your confidence intervals. What do you make of the fact that the two confidence intervals don't overlap?

An example from practice

The results in Table 9.1 are from a randomised trial to evaluate the use of an integrated care scheme for asthma patients, in which care is shared between the GP and a specialist chest physician (Grampian Asthma Study 1994). The treatment group patients each received this integrated care, the control group received conventional care from their GP only. The researchers were interested in the differences between the groups, if any, in a number of outcomes, shown in the figure (ignore the last column for now). The target population they have in mind is, perhaps, all asthma patients in the UK.

Table 9.1 Means and 95 per cent confidence intervals for a number of clinical outcomes over 12 months, for asthma patients. The treatment group patients received integrated care, the control group conventional GP care. Reproduced from *BMJ*, **308**, 559–64, courtesy of BMJ Publishing Group

Clinical outcome	Integrated care (n ≥ 296)	Conventional care (n ≥ 277)	Ratio of means
No of bronchodilators prescribed	10.1 (9.2 to 11.1)	10.6 (9.7 to 11.7)	0.95 (0.83 to 1.09)
No of inhaled steroids prescribed	6.4 (5.9 to 6.9)	6.5 (6.1 to 7.1)	0.98 (0.88 to 1.09)
No of courses of oral steroids used	1.6 (1.4 to 1.8)	1.6 (1.4 to 1.9)	0.97 (0.79 to 1.20)
No of general practice asthma consultations	2.7 (2.4 to 3.1)	2.5 (2.2 to 2.8)	1.11 (0.95 to 1.31)
No of hospital admissions for asthma	0.15 (0.11 to 0.19)	0.11 (0.08 to 0.15)	1.31 (0.87 to 1.96)

Means and 95% confidence interval are estimated from Poisson regression models after controlling for initial peak flow, forced expiratory volume (as % of predicted), and duration of asthma.

You can see that in the integrated care group of 296 subjects, the *sample* mean number of bronchodilators prescribed over 12 months was 10.1, with a 95 per cent confidence interval for the *population* mean of (9.2 to 11.1). So you can be 95 per cent confident that the population mean number of bronchodilators prescribed for this group is somewhere between 9.2 and 11.1. In the control group, the sample mean is 10.6 with a 95 per cent confidence interval for the population mean (9.7 to 11.7), which can be similarly interpreted.

Exercise 9.3 Interpret and compare the sample mean number of hospital admissions, and their corresponding confidence intervals, for the two groups in Table 9.1.

Confidence intervals as described above can also be applied to a population *percentage*, provided that the values are percentages of a metric variable, for example percentage mortality across a number of hospitals following some procedure (see, for example, Table 2.7). However, if the data is a proportion or percentage of a nominal or ordinal variable, say the proportion of patients with a pressure sore, or the proportion of mothers with an Edinburgh Maternal Depression Scale score of more than 8, then a different approach, described next, is needed.

Confidence interval for a population proportion

We start with an expression for the standard error of the sample proportion:

$$\text{s.e.} = (p)\sqrt{\frac{p(1-p)}{n}}$$

where p is the sample proportion, and n is sample size. Incidentally, the sampling distribution of sample proportions has a binomial distribution, which is quite different from the Normal distribution if the sample is small, but becomes more Normal as sample size increases. The 95 per cent confidence interval for the population proportion is equal to the sample proportion plus or minus 1.96^4 standard errors:

$$\{[p - 1.96 \times \text{s.e.}(p)] \text{ to } [p + 1.96 \times \text{s.e.}(p)]\}$$

For example, from Table 1.6, 14 of the 106 women with a malignant diagnosis are premenopausal giving a sample proportion p of 14/106 or 0.13. The standard error of p is thus:

$$\text{s.e.}(p) = \sqrt{\frac{0.13(1-0.13)}{106}} = 0.033$$

Therefore the 95 per cent confidence interval for the population proportion who are

⁴ When we are dealing with proportions, we use, not the t distribution, but the z , or *Standard Normal*, distribution. The 95 per cent value for z is 1.96.

pre-menopausal is:

$$(0.13 - 1.96 \times 0.033 \text{ to } 0.13 + 1.96 \times 0.033) = (0.065 \text{ to } 0.195)$$

In other words you can be 95 per cent confident that the proportion of cases in this population who are pre-menopausal lies somewhere between 0.065 to 0.195. Or alternatively, that this interval represents a plausible range of values for the population proportion who are menopausal.

Exercise 9.4 Calculate the standard error for the sample proportion of controls in Table 1.6 who are pre-menopausal, and hence calculate the 95 per cent confidence interval for the corresponding population proportion. Interpret your result.

Estimating a confidence interval for the median of a single population

If your data is ordinal then the median rather than the mean is the appropriate measure of location (review Chapter 5 if you're not sure why). Alternatively, if your data is metric but skewed (or your sample is too small to check the distributional shape), you might also prefer the median as a more representative measure. Either way a confidence interval will enable you to assess the likely range of values for the population median. As far as I know, SPSS does not calculate a confidence interval for a single median, but Minitab does, and bases its calculation on the *Wilcoxon signed-rank test*⁵ (I'll discuss this in Chapter 12).

Table 9.2 Sample median pain levels, and 95 per cent confidence intervals for the difference between the two groups, at three time periods, in the analgesics/stump pain study. Reproduced courtesy of Elsevier (*The Lancet*, 1994, Vol No. **344**, page 1724–6)

	Median (IQR) pain		
	Blockade group (n = 27)	Control group (n = 29)	95% CI for difference (p)
After epidural bolus	0 (0–0)	38 (17–67)	24 to 43 (p < 0.0001)
After continuous epidural infusion	0 (0–0)	31 (20–51)	24 to 43 (p < 0.0001)
After epidural bolus in operating theatre	0 (0–0)	35 (16–64)	19 to 42 (p < 0.0001)

Pain assessed by visual analogue scale (0–100 mm).

⁵ We won't deal with tests (i.e. *hypothesis tests*) until we get to Chapter 12, but the confidence intervals that I discuss in this and in the next chapter are based on a number of different hypothesis tests. The alternative would have been for me to introduce hypothesis tests before I dealt with confidence intervals. However, for various pedagogic reasons I didn't think this was appropriate.

An example from practice

Table 9.2 is from the analgesics and stump pain study referred to in Table 5.3, and shows the sample median pain levels and their 95 per cent confidence intervals (assessed using a visual analogue scale), for the treatment and control groups, at three time periods.

Exercise 9.5 In Table 9.2, interpret and compare the differences in median pain levels and their 95 per cent confidence intervals for each of the three time periods.

10

Estimating the difference between two population parameters

Learning objectives

When you have finished this chapter you should be able to:

- Give some examples of situations where there is a need to estimate the difference between two population parameters.
- Very briefly outline the basis of estimation of the difference between two population means using methods based on the two-sample t test¹ (for independent populations) and the matched-pairs t test (for matched populations).
- Very briefly outline the basis of estimation of the difference between two population medians using methods based on the Mann-Whitney test (for independent populations) and the Wilcoxon test (for matched populations).
- Interpret results from studies that estimate the difference between two population means, two percentages or two medians.
- Demonstrate an awareness of any assumptions that must be satisfied when estimating the difference between two population parameters.

¹ Throughout this chapter we will be looking at methods of estimation based on various *hypothesis tests*. I will begin to discuss hypothesis tests properly in Chapter 12.

What's the difference?

As you have just seen, it's possible to determine a confidence interval for any single population parameter – a population mean, a median, a percentage and so on. However, by far the most common application of confidence intervals is the *comparison* of *two* population parameters, for example between the means of two populations, such as the mean age of a population of women and the mean age of a population of men; I'll start with this.

Estimating the difference between the means of two independent populations – using a method based on the two-sample *t* test

The procedure here, like that for the single mean (see Chapter 9), is based on the *t* distribution (see the footnote on p. 114). However, with two populations, you need to know if they are *independent* or *matched* (see p. 81 to review matching). I'll start with estimating the difference in the means of two *independent* populations, since this is by far the most common in practice. For this we use a method based on the *two-sample t test*. First, there are a number of pre-requisites that need to be met:

- Data for both groups must be *metric*. As you know from Chapter 5 the mean is only appropriate with metric data anyway.
- The distribution of the relevant variable in *each* population must be reasonably *Normal*. You can check this assumption from the sample data using a histogram, although with small sample sizes this can be difficult.
- The population standard deviations of the two variables concerned should be *approximately* the same, but this requirement becomes less important as sample sizes get larger. You can check this by examining the two sample standard deviations.²

An example using birthweights

Suppose you want to compare (by estimating the difference between them), the population mean birthweights of infants born in a maternity unit with that of infants born at home (sample data in Table 10.1). The two samples were selected independently with no attempt at matching.

Both SPSS and Minitab compute the sample mean birthweight of the home-born infants to be 3726.5 g, with a standard deviation of 385.7 g. Recall that for the infants born in the maternity units, sample mean birthweight was 3644.4 g with a standard deviation of 376.8 g (see p. 112). So there *is* a difference in the *sample* mean birthweights of 82.1 g, (3726.5 g – 3644.4 g), but this does *not* mean that there is a difference in the *population* mean birthweights.

² This condition is usually stated in terms of the two *variances* being approximately the same. Variance is standard deviation squared.

Table 10.1 Sample data for birthweight (g), Apgar scores and whether mother smoked during pregnancy for 30 infants born in a maternity unit and 30 born at home

Infant	Birthweight (g)		Mother smoked		Apgar score	
	Hospital birth ^a	Home birth	Hospital birth	Home birth	Hospital birth	Home birth
1	3710	3810	0	0	8	10
2	3650	3865	0	0	7	8
3	4490	4578	0	0	8	9
4	3421	3522	1	0	6	6
5	3399	3400	0	1	6	7
6	4094	4156	0	0	9	10
7	4006	4200	0	0	8	9
8	3287	3265	1	0	5	6
9	3594	3599	0	1	7	8
10	4206	4215	0	0	9	10
11	3508	3697	0	0	7	8
12	4010	4209	0	0	8	9
13	3896	3911	0	0	8	8
14	3800	3943	0	0	8	9
15	2860	3000	0	1	4	3
16	3798	3802	0	0	8	9
17	3666	3654	0	0	7	8
18	4200	4295	1	0	9	10
19	3615	3732	0	0	7	8
20	3193	3098	1	1	4	5
21	2994	3105	1	1	5	5
22	3266	3455	1	0	5	6
23	3400	3507	0	0	6	7
24	4090	4103	0	0	8	9
25	3303	3456	1	0	6	7
26	3447	3538	1	0	6	7
27	3388	3400	1	1	6	7
28	3613	3715	0	0	7	7
29	3541	3566	0	0	7	8
30	3886	4000	1	0	8	6

^aThis is the data from Table 2.5.

It is important to remember that a difference between two sample values does not necessarily mean that there is a difference in the two population values. Any difference in these sample birthweight means might simply be due to chance. Now we come to an important point:

If the 95 per cent confidence interval for the difference between two population parameters includes zero, then you can be 95 per cent confident that there is *no* difference in the two parameter values. If the interval *doesn't* contain zero, then you can be 95 per cent confident that there *is* a statistically significant difference in the means.

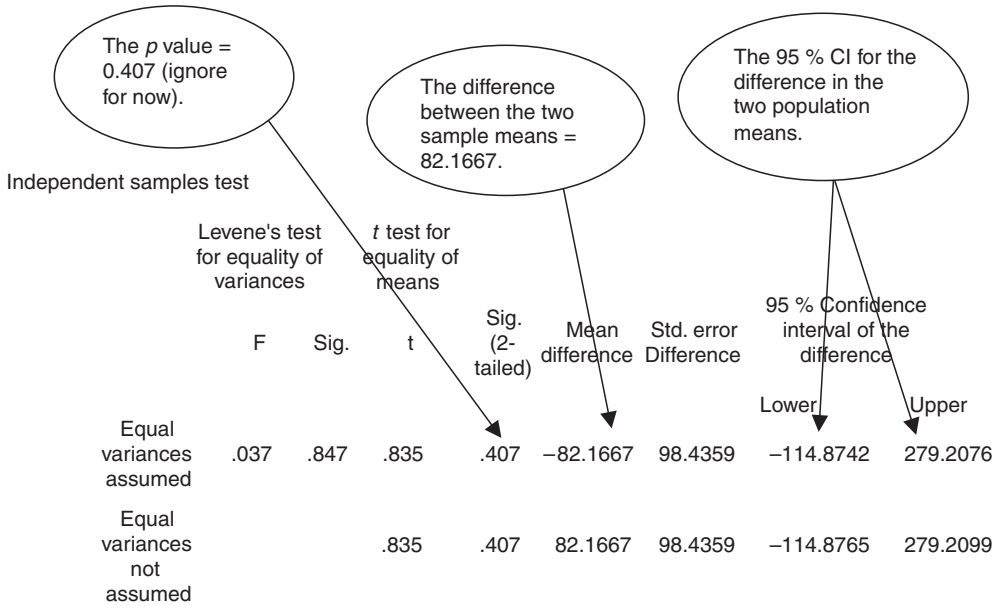


Figure 10.1 SPSS output (abridged) for 95 per cent confidence interval (last two columns) for the difference between two independent population mean birthweights, using samples of 30 infants born in maternity units and 30 at home (data in Table 10.1)

In other words, if you want to know if there is a statistically significant difference between two population means, calculate the 95 per cent confidence interval for the difference and see if it contains zero.

It is possible to calculate these confidence intervals by hand, but the process is time-consuming and tedious. Fortunately, most statistics programs will do it for you. Since difference between independent population means is one of the most commonly used approaches in clinical research, you might find it helpful to see some of the output from SPSS and Minitab for this procedure.

With SPSS

Using the birthweight data in Table 10.1, SPSS produces the results (abridged³) shown in Figure 10.1. These tell us that the difference in the two *sample* mean birthweights is -82.17 g. The sign in front of this value depends on which variable you select first in the SPSS dialogue box. SPSS subtracts the second variable selected (home births in this case) from the first (maternity unit births). This result means that the sample mean birthweight was 82.17 g higher in the home birth infants.

SPSS calculates two confidence intervals, one with standard deviations⁴ assumed to be equal, and one with them not equal. The 95 per cent confidence interval shown in the last two columns

³ I've removed material that is not relevant.

⁴ Both Minitab and SPSS refer to equality of *variances*.

is $(-114.9$ to $279.2)$ g, the same in both cases. SPSS tests for equality of the standard deviations (or *variances*), using Levene’s test. The assumption is that they are the same. We will discuss tests in Chapter 12.

Since this confidence interval includes zero, you can conclude that there is no statistically significant difference in *population* mean birthweights of infants born in a maternity unit and infants born at home.



With Minitab The Minitab output, which confirms that from SPSS, is shown in Figure 10.2. The 95 per cent confidence interval is in the second row up.

An example from practice

Table 10.2 is from a cohort study of maternal smoking during pregnancy and infant growth after birth (Conter *et al* 1995). The subjects were 12 987 babies who were followed up for three years after birth. Of these, 10 238 had non-smoking mothers, 2276 had mothers who had

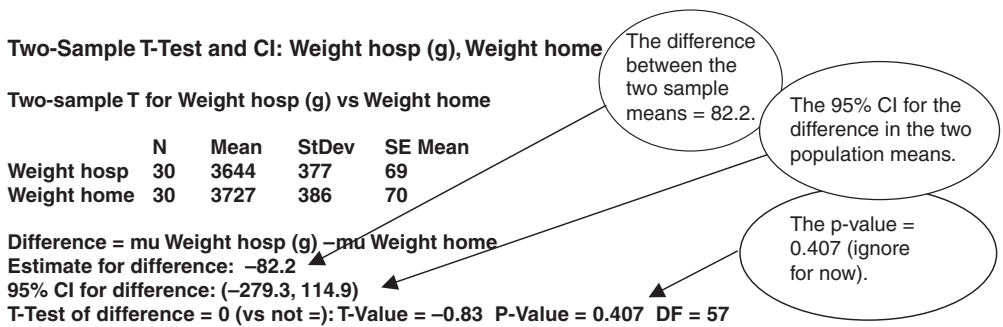


Figure 10.2 Minitab output for 95 per cent confidence interval for the difference between two independent population mean birthweights, using samples of 30 infants born in maternity units and 30 at home. Note that Minitab uses the word ‘mu’ to denote the population mean, normally designated as Greek μ

Table 10.2 95 per cent confidence intervals for difference in weights according to sex and smoking habits of mothers between independent groups of babies. Reproduced from *BMJ*, 310, 768–71, courtesy of BMJ Publishing Group

Mother's smoking habit	At birth			At 3 months			At 6 months		
	No. children	Weight (g)	95%CI for difference	No. children	Weight (g)	95%CI for difference	No. children	Weight (g)	95%CI for difference
Girls									
Non-smokers	4904	3220		4904	5584		4895	7462	
1–9 cigs per day	1072	3132	(–121 to –55)	1071	5550	(–77 to 9)	1072	7471	(–47 to 65)
≥10 per day	228	3052	(–234 to –102)	228	5519	(–152 to 22)	227	7434	(–141 to 85)
Boys									
Non-smokers	5334	3373		5332	6026		5330	8038	
1–9 cigs per day	1204	3266	(–139 to –75)	1204	5958	(–113 to –23)	1204	7974	(–118 to –10)
≥10 cigs per day	245	3126	(–312 to –181)	245	5907	(–212 to –26)	245	8014	(–136 to 88)

smoked one to nine cigarettes a day, and 473 had mothers who had smoked 10 or more cigarettes a day. The figure shows the 95 per cent confidence intervals for differences in mean weight according to sex of baby and smoking habits of mothers: at birth, and at three and six months.

The results show, for example, that at birth, the difference between the sample mean weight of female babies born to non-smoking mothers and those born to mothers smoking 10 or more cigarettes a day, was $(3220 - 3052) = 168$ g. That is, the infants of smoking mothers are on average lighter by 168 g. Is this difference statistically significant in the population, or due simply to chance? The 95 per cent confidence interval of $(-234$ to $-102)$ g, does *not* include zero, so you can be 95 per cent confident that the difference is real, i.e. is statistically significant.

Exercise 10.1 Interpret the sample mean and confidence intervals shown in Table 10.2 for all four differences in weights at six months.

Estimating the difference between two matched population means – using a method based on the matched-pairs *t* test

If the data within each of the two groups whose means you are comparing is widely spread compared to the difference in the spreads between the groups,⁵ this can make it more difficult to detect any difference in their means. When data is matched (see Chapter 7 for an explanation of matching), this reduces much of the within-group variation, and, for a given sample size, makes it easier to detect any differences between groups. As a consequence, you can achieve better precision (narrower confidence intervals), without having to increase sample size. The disadvantage of matching is that it is sometimes difficult to find a sufficiently large number of matches (as you saw in the case-control discussion earlier).

In the independent groups case, the mean of each group is computed separately, and then a confidence interval for the difference in these means is calculated. In the matched groups case, we use a method based on the *matched-pairs t test*, in which the *difference* between each pair of values is computed first and then a confidence interval for the mean of these differences is calculated.

An example from practice

Table 10.3 shows the 95 per cent confidence intervals for the difference in bone mineral density in two matched groups of women, one group depressed and one ‘normal’ (Michelson *et al.* 1995). (Ignore the ‘SD from expected peak’ rows.) Only one of the confidence intervals contains zero, indicating that there is no difference in population mean bone mineral density at the radius, but there is at all of the other five sites.

⁵ Called ‘between-group’ variation.

Table 10.3 Confidence intervals for the differences between the population mean bone mineral densities in two individually *matched* groups of women, one group depressed, the other ‘normal’, using a method based on the matched-pairs t test. Reproduced from *NEJM*, **335**, 1176–81, by permission of Massachusetts Medical Society

Bone Measured [†]	Depressed Women	Normal Women	Mean Difference (95% CI)	P Value
Lumbar spine (anteroposterior)				
Density (g/cm ²)	1.00 ± 0.15	1.07 ± 0.09	0.08 (0.02 to 0.14)	0.02
SD from expected peak	−0.42 ± 1.28	0.26 ± 0.82	0.68 (0.13 to 1.33)	
Lumbar spine (lateral) [‡]				
Density (g/cm ²)	0.74 ± 0.09	0.79 ± 0.07	0.05 (0.00 to 0.09)	0.03
SD from expected peak	−0.88 ± 1.07	−0.36 ± 0.80	0.50 (0.04 to 1.03)	
Femoral neck				
Density (g/cm ²)	0.76 ± 0.11	0.88 ± 0.11	0.11 (0.06 to 0.17)	<0.00
SD from expected peak	−1.30 ± 1.07	−0.22 ± 0.99	1.08 (0.55 to 1.61)	
Ward’s triangle				
Density (g/cm ²)	0.70 ± 0.14	0.81 ± 0.13	0.11 (0.06 to 0.17)	<0.00
SD from expected peak	−0.93 ± 1.24	0.18 ± 1.22	1.11 (0.60 to 1.62)	
Trochanter				
Density (g/cm ²)	0.66 ± 0.11	0.74 ± 0.08	0.08 (0.04 to 0.13)	<0.001
SD from expected peak	−0.70 ± 1.22	0.26 ± 0.91	0.97 (0.46 to 1.47)	
Radius				
Density (g/cm ²)	0.68 ± 0.04	0.70 ± 0.04	0.01 (−0.01 to 0.04)	0.25
SD from expected peak	−0.19 ± 0.67	0.03 ± 0.67	0.21 (−0.21 to 0.64)	

*Plus-minus values are means ± SD. CI denotes confidence interval.

[†]Values for “SD from expected peak” are the numbers of standard deviations from the expected peak density derived from a population-based study of normal white women.³

[‡]This measurement was made in 23 depressed women and 23 normal women.

Exercise 10.2 In Table 10.3, which population difference in bone mineral density is estimated with the greatest precision?

You can also calculate a confidence interval for the difference in two population *percentages* provided they derive from two metric variables. For the difference between two population proportions, however, a different approach is needed. This is an extension of the single proportion case discussed in Chapter 9, as you will now see.

Estimating the difference between two independent population proportions

Suppose you want to calculate a 95 per cent confidence interval for the difference between the population proportion of women having maternity unit births who smoked during pregnancy and the proportion having home births who smoked. The sample data on smoking status for the sample of 60 mothers is shown in Table 10.1.

There are 10 mothers who smoked among the 30 giving birth in the maternity unit and six among the 30 giving birth at home. This gives sample proportions of $10/30 = 0.3333$, and $6/30 = 0.2000$, respectively. You can check whether this difference is statistically significant or likely to be due to chance alone, by calculating a 95 per cent confidence interval for the difference in the corresponding population proportions.⁶ To do this by hand is a bit long-winded and you would want to use a computer program to do the calculation for you.

An example from practice

If you look back at Table 9.1, the randomised trial of integrated versus conventional care for asthma patients, the last column shows the 95 per cent confidence intervals for the difference in population percentages between the two groups, for a number of patient perceptions of the scheme. As you can see, none of the confidence intervals include zero, so you can be 95 per cent confident that the difference in population percentages between the groups of patients is statistically significant in each case.

Estimating the difference between two independent population medians – the Mann–Whitney rank-sums method

As you know from Chapter 5, the mean may not be the most representative measure of location if the data is skewed, and is not appropriate anyway if the data is ordinal. In these circumstances, you can compare the population *medians* rather than the means, and in place of the 2-sample *t* test (a parametric procedure), use a method based on the *Mann–Whitney* test (a non-parametric procedure).

Parametric versus non-parametric methods

A *parametric* procedure can be applied to data which is metric, and also has some particular distribution, most commonly the Normal distribution. A *non-parametric* procedure does not make these distributional requirements. So if you are analysing data that is either metric but not Normal, or is ordinal, then you need to use a non-parametric approach. The Mann–Whitney procedure only requires that the two population distributions have the same approximate shape, but does not require either to be Normal. It is the non-parametric equivalent of the two-sample *t* test.

Briefly, the Mann–Whitney method starts by combining the data from both groups, which are then ranked. The rank values for each group are then separated and summed. If the medians of the two groups are the same, then the sums of the ranks of the two groups should be

⁶ The 95 per cent confidence interval is $(-0.088$ to $0.355)$. Since this interval includes 0, we conclude that there is no difference in the proportion of mothers who smoked at home and in the maternity unit.

Mann-Whitney Test and CI: Apgar matn, Apgar home

```

Apgar ma  N = 30      Median =      7.000
Apgar ho  N = 30      Median =      8.000
Point estimate for ETA1-ETA2 is      -1.000
95.2 Percent CI for ETA1-ETA2 is (-2.000,0.000)

W = 790.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0668
The test is significant at 0.0616 (adjusted for ties)

Cannot reject at alpha = 0.05

```

Confidence interval for the difference in the two medians.

Figure 10.3 Minitab's Mann-Whitney output for a 95 per cent confidence interval for the difference between two independent median Apgar scores – for infants born in maternity units and at home (raw data in Table 10.1). Note that Minitab uses Greek 'ETA' to denote the population median

similar. However, if the rank sums are different, you need to know whether this difference could simply be due to chance, or is because there really is a statistically significant difference in the population medians. A Mann-Whitney confidence interval for the difference will help you decide between these alternatives.

As an illustration, let's compare the difference in the population median Apgar scores for the maternity unit and home birth infants, using the sample data in Table 10.1. These are independent groups, but since this data is ordinal, we cannot use the two-sample *t* test, but we can use the Mann-Whitney test of medians. The output from Minitab is shown in Figure 10.3, with the 95 per cent confidence interval in the fourth row.⁷ Since the confidence interval of (–2 to 0) contains zero, you must conclude that the difference in the population median Apgar scores is not statistically significant. Notice that the confidence level is given as 95.2 per cent, not 95 per cent. Confidence intervals for medians cannot always achieve the precise confidence level you asked for, because of the way in which a median is calculated.

An example from practice

Table 10.4 is from a randomised controlled double-blind trial to compare the cost effectiveness of two treatments in relieving pain after blunt instrument injury in an A&E department (Rainer *et al.* 2000). It shows the median times spent by two groups of patients in various clinical situations. One group received ketorolac, the other group morphine. The penultimate column contains the 95 per cent confidence intervals for the difference in various median treatment times (minutes), between the groups (ignore the last column). As the footnote to the table indicates, these results were obtained using the Mann-Whitney method.

The only confidence interval not containing zero is that for the difference in median 'time between receiving analgesia and leaving A&E', for which the difference in the sample medians is 20.0 minutes. So this is the only treatment time for which the difference in population median

⁷ As far as I am aware, SPSS does not appear to calculate a confidence interval for two independent medians.

Table 10.4 Mann–Whitney confidence intervals for the difference between two *independent* groups of patients in their median times spent in several clinical situations. One group received ketorolac, the other morphine median number (interquartile range) of minutes relating to participants treatment. Reproduced from *BMJ*, 321, 1247–51, courtesy of BMJ Publishing Group

Variable	Ketorolac group (n = 75)	Morphine group (n = 73)	Median difference (95% confidence interval)	P value*
Interval between arrival in emergency department and doctor prescribing analgesia	38.0 (30.0 to 54.0)	39.0 (29.0 to 53.0)	1.0 (–5.0 to 7.0)	0.72
Preparation for analgesia	5.0 (5.0 to 10.0)	10.0 (5.5 to 12.5)	2.0 (0 to 5.0)	0.0002
Undergoing radiography	5.0 (5.0 to 10.0)	5.0 (4.0 to 10.0)	0 (–1.0 to 0)	0.75
Total time spent in emergency department	155.0 (112.0 to 198.0)	171.0 (126.0 to 208.5)	15.0 (–4.0 to 33.0)	0.11
Interval between receiving analgesia and leaving emergency department	115.0 (75.0 to 149.0)	130.0 (95.0 to 170.0)	20.0 (4.0 to 39.0)	0.02

*Mann–Whitney U test.

Table 10.5 Confidence interval estimates from the Wilcoxon signed-ranks method for the difference in population food intakes per day, for a number of substances, from a study of the dietary habits of schizophrenics. Values are median (range). Reproduced from *BMJ*, 317, 784–5, courtesy of BMJ Publishing Group

Intake/day	Men		Women		All		Wilcoxon signed ranks test	
	Patients (n = 17)	Controls (n = 17)	Patients (n = 13)	Controls (n = 13)	Patients (n = 30)	Controls (n = 30)	Median difference (95% CI)	P
Energy (MJ)	11.84 (7.67–17.93)	14.19 (6.94–23.22)	8.87 (5.07–13.02)	9.99 (5.25–16.25)	9.71 (5.07–17.94)	11.98 (5.25–23.22)	2.06 (0.26–4.23)	0.04
Protein (g)	92.5 (65.1–157.4)	114.2 (74–633)	68.7 (38.4–104.2)	82.5 (40.5–142.7)	84.5 (38.4–157.4)	96.0 (40.5 to 633.0)	15.9 (–1.1 to 32.8)	0.07
Total fibre (g)	13.0 (8.5–20.8)	22.0 (8.7–86.2)	10.7 (7.3–18.0)	15.5 (10.7–22.9)	12.6 (7.3–20.8)	18.9 (8.7–86.2)	7.0 (3.6 to 10.6)	0.0001
Retinol (μ g)	647 (294–1498)	817 (134–12341)	533 (288–7556)	817 (201–11585)	590 (288–7556)	817 (134–12341)	310 (93 to 1269)	0.02
Carotene (μ g)	783 (219–3638)	2510 (523–11313)	2048 (550–4657)	3079 (956–6188)	1443 (219–4657)	2798 (523–11313)	1376 (549 to 2452)	0.004
Vitamin C (mg)	41.0 (4.0–204)	81.0 (14.0–262)	40.0 (3–165)	61.0 (27.0–291.0)	40.5 (3.0–204)	80.5 (14.0–219)	33.5 (2.0 to 64.0)	0.03
Vitamin E (mg)	4.8 (3.4–18.0)	10.26 (2.23–32.0)	4.5 (2.3–6.0)	5.38 (3.6–14.7)	4.7 (2.3–18.0)	7.8 (2.2–32.0)	2.9 (1.45 to 5.35)	0.0002
Alcohol (g)	3.8 (0–19.4)	11.7 (0–80)	0 (0–5.6)	1.8 (0–12)	0 (0–19.4)	5.7 (0–80)	5.4 (1.2 to 9.9)	0.009

times is statistically significant, and you can be 95 per cent confident that this difference is between 4 and 39 minutes.

Exercise 10.3 Table 10.4 includes the sample median times and their 95 per cent confidence intervals for each time interval, for both groups separately. Only one pair of confidence intervals don't overlap, those for the only time difference which is statistically significant. Why aren't you surprised by this?

Estimating the difference between two matched population medians – Wilcoxon signed-ranks method

When two groups are *matched*, but either the data is ordinal, or if metric is noticeably skewed, you can obtain confidence intervals for differences in population medians, based on the non-parametric *Wilcoxon test*. The two population distributions, regardless of shape, should be symmetric. This is the non-parametric equivalent of the parametric matched-pairs *t* test, described above. The matching will again reduce the variation within groups, so narrower, and therefore more precise, confidence intervals are available for a given sample size.

Briefly the Wilcoxon method starts by calculating the difference between each pair of values, and these differences are then ranked (ignoring any minus signs). Any negative signs are then restored to the rank values, and the negative and positive ranks are separately summed. If the medians in the two groups are the same, then these two rank sums should be similar. If different, the Wilcoxon method provides a way of determining whether this is due to chance, or represents a statistically significant difference in the population medians.

An example from practice

Table 10.5 contains the results of a case-control study into the dietary intake of schizophrenic patients living in the community in Scotland (McCreadie *et al.* (1998). It shows the daily energy intake of eight dietary substances for the cases (17 men and 13 women diagnosed with schizophrenia), and the controls, each individually matched on sex, age, smoking status and employment status.

If you focus on the penultimate column, in which data for men and women is combined, you can see that only the confidence interval for daily protein intake, (–1.1 to 32.8) g, contains zero, which implies that there is no difference in population median protein intake between schizophrenics and normal individuals. For all other substances, the difference is statistically significant.

Exercise 10.4 Explain the meaning of the 95 per cent confidence interval for difference in median alcohol intake of the two groups in Table 10.5.

11

Estimating the *ratio* of two population parameters

Learning objectives

When you have finished this chapter you should be able to:

- Explain what is meant by the ratio of two population parameters and give some examples of situations where there is a need to estimate such a ratio.
- Explain and interpret a confidence interval for a risk ratio.
- Explain and interpret a confidence interval for an odds ratio.
- Explain the difference between crude and adjusted risk and odds ratios.

Estimating ratios

Estimating the ratio of two independent population means

When you compare two population means you usually want to know if they're the same or not, and if not, how big the difference between them is. Sometimes though, you might want to know *how many times bigger* one population mean is than another. The *ratio* of the two means will tell us that.

If two sample means have a ratio of 1, this tells us *only* that the means are the same size in the *sample*. If the sample ratio *is* different from 1, you need to check whether this is simply due to chance, or if the difference is statistically significant – one mean *is* bigger than the other. You can do this with a 95 per cent confidence interval for the ratio of population means. And here's the rule:

If the confidence interval for the *ratio* of two population parameters does *not* contain the value 1, then you can be 95 per cent confident that any difference in the size of the two measures is statistically significant.

Compare this with the rule for the *difference* between two population parameters, where that rule is that if the confidence interval does *not* contain zero, then any difference between the two parameters *is* statistically significant.

An example from practice

Look again at the last column in Table 9.1, which shows a number of outcomes from a randomised trial to compare integrated versus conventional care for asthma patients. The last column contains the 95 per cent confidence intervals for the ratio of population means for the treatment and control groups. You will see that *all* of the confidence intervals contain 1, indicating that the population mean number of bronchodilators used, the number of inhaled steroids prescribed and so on, was no larger (or smaller) in one population than in the other.

The *sample* ratio furthest away from 1 is 1.31, for the ratio of mean number of hospital admissions, i.e. the *sample* of integrated care group patients had 31 per cent more admissions than the conventionally treated control group patients. However, the 95 per cent confidence interval of (0.87 to 1.96) includes 1, which implies that this is generally *not* the case in the populations.

Confidence interval for a population risk ratio

Table 6.1 showed the contingency table for a cohort study into the risk of coronary heart disease (CHD) as an adult, among men who weighed 18 lbs or less at 12 months old (the risk factor). On p. 104 we derived a risk ratio of 1.93 from this sample cohort. In other words, men who weighed 18 lbs or less at one year, appear to have nearly twice the risk of CHD when an adult, as men who weighed more than 18 lbs at one year. But is this true in the *population* of such men, or no more than a *chance* departure from a population ratio of 1? You now know that you can answer this question by examining the 95 per cent confidence interval for this risk ratio.

The 95 per cent confidence interval for the CHD risk ratio turns out to be (0.793 to 4.697).¹ Since this interval contains 1, you can conclude, that despite a *sample* risk ratio of nearly 2, that

¹ The calculation of confidence intervals for risk ratios and odds ratios is a step too far for this book. Those interested in doing the calculation by hand can consult Altman (1991) who gives the necessary formulae.

weighing 18 lbs or less at one year is *not* a significant risk factor for coronary heart disease in adult life in the sampled *population*. Notice that, in general, the value of a sample risk or odds ratio, as in this example, does *not* lie in the centre of its confidence interval, but is usually closer to the lower value.

An example from practice

Table 11.1 is from a cohort study of 552 men surviving acute myocardial infarction, in which each subject was assessed for depression at the beginning of the study (Ladwig *et al.* 1994). 14.5 per cent were identified as severely depressed, 2.3 per cent as moderately depressed, and 63.2 per cent had low levels of depression. The subjects were followed up at 6 months, and a number of outcomes measured, including: suffering angina, returning to work, emotional stability and smoking. The researchers were interested in examining the role of moderate and of severe depression (compared to low depression), as risk factors for each of these outcomes.

The results show the crude and adjusted risk ratios (labelled ‘relative risks’ by the authors) for each outcome. The crude risk ratios are *not* adjusted for any confounding factors, whereas the adjusted risk ratios *are* adjusted for the factors listed in the table footnote (review the material on confounding and adjustment in Chapter 7 if necessary).

Let’s interpret the 95 per cent risk ratios for ‘return to work’. The *crude* risk ratios for a return to work indicate lower rates of return to work for men both moderately depressed (risk

Table 11.1 The crude and adjusted risk ratios (labelled relative risk by the authors), for a number of outcomes related to the risk factor of experiencing moderate and severe levels of depression compared to low depression. Reprinted courtesy of Elsevier (*The Lancet*, 1994, Vol No. 343, page 20–3)

Depression level	Relative risk (95% CI)	
	Crude	Adjusted*
Angina pectoris		
Moderate	1.36 (0.83 to 2.23)	0.97 (0.55 to 1.70)
Severe	3.12 (1.58 to 6.16)	2.31 (1.11 to 4.80)
Return to work		
Moderate	0.41 (0.22 to 0.77)	0.58 (0.28 to 1.17)
Severe	0.39 (0.18 to 0.88)	0.54 (0.22 to 1.31)
Emotional Instability		
Moderate	2.21 (1.33 to 3.69)	1.87 (1.07 to 3.27)
Severe	5.55 (2.87 to 10.71)	4.61 (2.32 to 9.18)
Smoking		
Moderate	1.39 (0.71 to 2.73)	1.19 (0.56 to 2.51)
Severe	2.63 (1.23 to 5.60)	2.84 (1.22 to 6.63)
Late potentials		
Moderate	1.30 (0.76 to 2.22)	1.54 (0.86 to 2.74)
Severe	0.70 (0.33 to 1.47)	0.75 (0.35 to 2.17)

* Adjusted for age, social class, recurrent infarction, rehabilitation, cardiac events and helplessness

ratio = 0.41), and severely depressed (risk ratio = 0.39), compared to men with low levels of depression. Neither of the confidence intervals, (0.22 to 0.77) and (0.18 to 0.88), includes 1, indicating statistical significance. However, after adjusting for possible confounding variables, the *adjusted* risk ratios are 0.58 and 0.54, and are no longer statistically significant, because the confidence intervals for both risk ratios, for moderate depression (0.28 to 1.17), and severe depression (0.22 to 1.31), now include 1.

Exercise 11.1 Table 11.2 is from the same cohort study referred to in Exercise 8.9, to investigate dental disease, and risk of coronary heart disease (CHD) and mortality, involving over 20 000 men and women aged 25–74, who were followed up between 1971–4 and 1986–7 (DeStefano *et al.* 1993).

The results give the risk ratios (called relative risks here) for CHD and mortality in those with a number of dental diseases compared to those without (the referent group), adjusted for a number of possible confounding variables (see table footnote for a list of the variables adjusted for).

Briefly summarise what the results show about dental disease as a risk factor for CHD and mortality. Note: the periodontal index (range from 0–8, higher is worse) measures the average degree of periodontal disease in all teeth present, and the oral hygiene index (range 0–6, higher is worse) measures the average degree of debris and calculus on the surfaces of six selected teeth.

Confidence intervals for a population odds ratio

Table 6.2 showed the data for the case-control study into exercise between the ages of 15 and 25, and stroke later in life. The risk factor was ‘not exercising’, and you calculated the *sample* crude odds ratio of 0.411 for a stroke, in those who hadn’t exercised compared to those who

Table 11.2 Adjusted risk ratios for CHD and mortality among those with dental disease compared to those without dental disease*. Reproduced by permission of BMJ Publishing Group. (*BMJ*, 1993, Vol. **306**, pages 688–691)

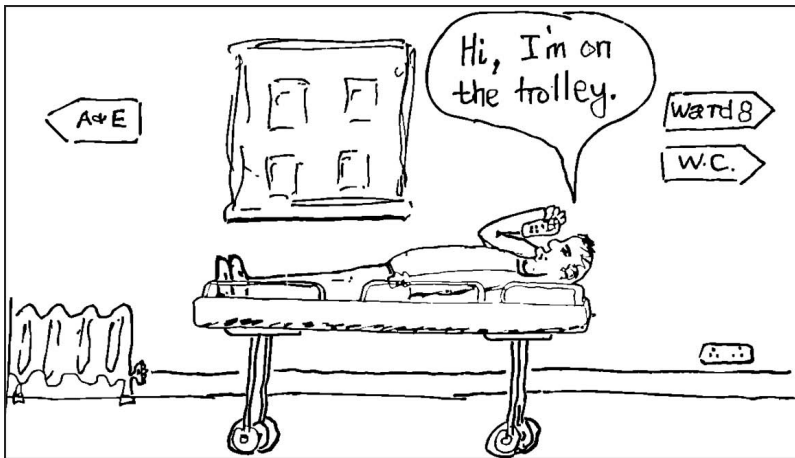
Indicator	No of subjects [†]	Coronary heart disease	Total mortality
Periodontal class:			
No disease	673	1.00	1.00
Gingivitis	529	0.98 (0.63 to 1.54)	1.42 (0.84 to 2.42)
Periodontitis	300	1.72 (1.10 to 2.68)	2.12 (1.24 to 3.62)
No teeth	92	1.71 (0.93 to 3.15)	2.60 (1.33 to 5.07)
Periodontal index (per unit)	1502	1.09 (1.00 to 1.19)	1.11 (1.01 to 1.22)
Oral hygiene index (per unit)	1436	1.11 (0.96 to 1.27)	1.23 (1.06 to 1.43)

* Adjusted for age, sex, race, education, poverty index, marital state, systolic blood pressure, total cholesterol concentration, diabetes, body mass index, physical activity, alcohol consumption, and cigarette smoking.

[†] Excluding those with missing data for any variable and, for periodontal index and hygiene index, those who had no teeth.

had (see p. 105). So the exercising group appear to have under half the odds for a stroke as the non-exercising group. However, you need to examine the confidence interval for this odds ratio to see if it contains 1 or not, before you can come to a conclusion about the statistical significance of the *population* odds ratio.

SPSS produces an odds ratio of 0.411, with a 95 per cent confidence interval of (0.260 to 0.650). This does not contain 1, so you can be 95 per cent confident that the odds ratio for a stroke in the *population* of those who did exercise compared to the population of those who didn't exercise is somewhere between 0.260 and 0.650. So early-life exercise does seem to reduce the odds for a stroke later on. Of course this is a crude, unadjusted odds ratio, which takes no account of the contribution, positive or negative, of any other relevant variables.



An example from practice

Table 11.3 shows the results from this same exercise/stroke study, where the authors provide both crude odds ratios and ratios *adjusted* for a number of different variables (Shinton and Sagar 1993).

We have been looking at exercise between the ages of 15 and 25, the first row of the table. Compared to the *crude* odds ratio calculated above of 0.411, the authors report an odds ratio for stroke, *adjusted* for age and sex, among those who exercised compared to those who didn't exercise, as 0.33, with a 95 per cent confidence interval of (0.20 to 0.60). So even after the effects of any differences in age and sex between the two groups has been adjusted for, exercising remains a statistically significant 'risk' factor for stroke (although *beneficial* in this case). Adjustment for possible confounders is crucial if your results are to be of any use, and I will return to adjustment and how it can be achieved in Chapter 18.

Table 11.3 Odds ratios for stroke*, according to whether, and at what age, exercise was undertaken by patients, compared to controls without stroke. Reproduced by permission of BMJ Publishing Group. (*BMJ*, 1993, Vol. **307**, pages 231–234)

	Exercise not undertaken		Exercise undertaken	
	Odds ratio	No of cases: no of controls	Odds ratio (95% confidence interval)	No of cases: no of controls
Age when exercise undertaken (years):				
15–25	1.0	70:68	0.33 (0.2 to 0.6)	55:130
25–40	1.0	103:136	0.43 (0.2 to 0.8)	21:57
40–55	1.0	101:139	0.63 (0.3 to 1.5)	10:22

* Adjusted for age and sex

Exercise 11.2. (a) Explain briefly why, in Table 11.3, age and sex differences between the groups have to be adjusted for. (b) What do the results indicate about exercise as a risk factor for stroke among the 25–40 years and 40–55 years groups?

Exercise 11.3. Refer back to Table 1.7, the results from a cross-section study into thrombotic risk during pregnancy. Identify and interpret any statistically significant odds ratios.

VI

Putting it to the Test

12

Testing hypotheses about the *difference* between two population parameters

Learning objectives

When you have finished this chapter you should be able to:

- Explain how a research question can be expressed in the form of a testable hypothesis.
- Explain what a null hypothesis is.
- Summarise the hypothesis test procedure.
- Explain what a p -value is.
- Use the p -value to appropriately reject or not reject a null hypothesis.
- Summarise the principal tests described in this chapter, along with their most appropriate application, and any distributional and other requirements.
- Interpret SPSS and Minitab results from a hypothesis test.
- Interpret published results of hypothesis tests.
- Point out the advantages of confidence intervals over hypothesis tests.
- Describe type I and type II errors, and their probabilities.

- Explain the power of a test and how it is calculated.
- Explain the connection between power and sample size.
- Calculate sample size required in some common situations.

The research question and the hypothesis test

The procedures discussed in the preceding three chapters have one primary aim: to use confidence intervals to estimate population parameter values, and their differences and ratios. We were able to make statements like, ‘We are 95 per cent confident that the range of values defined by the confidence interval will include the value of the population parameter,’ or, ‘The confidence interval represents a plausible range of values for the population parameter.’

There is, however, an alternative approach called *hypothesis testing*, which uses exactly the same sample data as the confidence interval approach, but focuses not on *estimating* a parameter value, but on *testing* whether its value is the same as a previously specified or *hypothesised* value. In recent years, the estimation approach has become more generally favoured, primarily because the results from a confidence interval provides *more information* than the results of a hypothesis test (as you will see a bit later). However, hypothesis testing is still very common in research publications, and so I will describe a few of the more common tests.¹ Let’s first establish some basic concepts.

The null hypothesis

As we have seen, almost all clinical research begins with a question. For example, is Malathion a more effective drug for treating head lice than *d*-phenothrin? Is stress a risk factor for breast cancer? To answer questions like this you have to transform the *research question* into a *testable hypothesis* called the *null hypothesis*, conventionally labelled H_0 . This usually takes the following form:

H_0 : Malathion is *NOT* a more effective drug for treating head lice than *d*-phenothrin.

H_0 : Stress is *NOT* a risk factor for breast cancer.

Notice that both of these null hypotheses reflect the conservative position of *no* difference, *no* risk, *no* effect, etc., hence the name, ‘*null*’ hypothesis. To *test* this null hypothesis, researchers will take samples and measure outcomes, and decide whether the data from the sample provides strong enough evidence to be able to refute or *reject* the null hypothesis or not. If evidence against the null hypothesis is strong enough for us to be able to reject it, then we are implicitly accepting that some specified alternative hypothesis, usually labelled H_1 , is probably true.

¹ And there are some situations where there is no reasonable alternative to a hypothesis test.

The hypothesis testing process

The hypothesis testing process can be summarised thus:

- Select a suitable outcome variable.
- Use your research question to define an appropriate and testable null hypothesis involving this outcome variable.
- Collect the appropriate sample data and determine the relevant sample statistic, e.g. sample mean, sample proportion, sample median, (or their difference or ratio), etc.
- Use a decision rule that will enable you to judge whether the sample evidence supports or does not support your null hypothesis.
- Thus, on the strength of this evidence, either reject or do not reject your null hypothesis.

Let's take a simple example. Suppose you want to test whether a coin is fair, i.e. not weighted to produce more heads or more tails than it should. Your null hypothesis is that the coin is fair, i.e. will produce as many heads as tails, so that the population proportion π , equals 0.5. Your outcome variable is the sample proportion of heads, p . You toss the coin 100 times, and get 42 heads, so $p = 0.42$. Is this outcome compatible with your hypothesised value of 0.5? Is the difference between 0.5 and 0.42 statistically significant or could it be due to chance?

You can probably see the problem. How do we decide *what* proportion of heads we might expect to get if the coin is fair? As it happens, there is a generally accepted rule, which involves something known as the *p-value*.

The p-value and the decision rule

The hypothesis test decision rule is: *If the probability of getting the number of heads you get (or even fewer) is less than 0.05,² when the null hypothesis is true, then this is strong enough evidence against the null hypothesis and it can be rejected.* The beauty of this rule is that you can apply it to any situation where the probability of an outcome can be calculated, not just to coin tossing.

As a matter of interest, the probability of getting say 42 or fewer heads if the coin is fair is 0.0666, which is *not* less than 0.05. This is *not* strong enough evidence against the null hypothesis. However, if you had got 41 heads or fewer, the probability of which is 0.0443, this *is* less than 0.05, now the evidence against H_0 is strong enough and it can be rejected. The coin is not fair. This crucial threshold outcome probability (0.0443 in this example), is called the *p-value*, and defined thus:

A *p-value* is the probability of getting the outcome observed (or one more extreme), assuming the null hypothesis to be true.

²Or 0.01. There is nothing magical about these values, they are quite arbitrary.

So, in the end, the decision rule is simple:

- Determine the *p-value* for the output you have obtained (using a computer).
- Compare it with the *critical value*, usually 0.05.
- If the *p-value* is *less* than the critical value, reject the null hypothesis; otherwise do not reject it.

When you reject a null hypothesis, it's worth remembering that although there is a probability of 0.95 that you are making the correct decision, there is a corresponding probability of 0.05 that your decision is incorrect. In fact, you *never* know whether your decision is correct or not,³ but there are 95 chances in 100 that it is. Compare this with the conclusion from a confidence interval where you can be 95 per cent confident that a confidence interval will include the population parameter, but there's still a 5 per cent chance that it will not.

It's important to stress that the *p-value* is *not* the probability that the null hypothesis is true (or not true). It's a measure of the *strength of the evidence against* the null hypothesis. The smaller the *p-value*, the stronger the evidence (the less likely it is that the outcome you got occurred by chance). Note that the critical value, usually 0.05 or 0.01, is called the *significance level* of the hypothesis test and denoted α (alpha). We'll return to alpha again shortly.

Exercise 12.1 Suppose you want to check your belief that as many males as females use your genito-urinary clinic. (a) Frame your belief as a research question. (b) Write down an appropriate null hypothesis. (c) You take a sample of 100 patients on Monday and find that 40 are male. The *p-value* for 40 or fewer males from a sample of 100 individuals is 0.028. Do you reject the null hypothesis? (d) Your colleague takes a sample of 100 patients on the following Friday and gets 43 males, the *p-value* for which is 0.097. Does your colleague come to the same decision as you did? Explain your answer.

A brief summary of a few of the commonest tests

Some hypothesis tests are suitable only for metric data, some for metric and ordinal data, and some for ordinal and nominal data. Some require data to have a particular distribution (often Normal); these are *parametric* tests. Some have no or less strict distributional requirements; the *non-parametric* tests. Before I discuss a few tests in any detail, I have listed in Table 12.1 a brief summary of the more commonly used tests, along with their data and distributional requirements, if any. I am ignoring tests of single population parameters since these are not required often enough to justify any discussion.

³ Because you'll never know what the value of any population parameter is.

Table 12.1 Some of the more common hypothesis tests

Two-sample t test. Used to test whether or not the difference between two *independent* population means is zero (i.e. the two means are equal). The null assumption is that it is. Both variables must be metric and Normally distributed (this is a parametric test). In addition the two population standard deviations should be similar (but for larger sample sizes this becomes less important).

Matched-pairs t test. Used to test whether or not the difference between two *paired* population means is zero. The null assumption is that it is, i.e. the two means are equal. Both variables must be metric, and the *differences* between the two must be Normally distributed (this is a parametric test).

Mann-Whitney test. Used to test whether or not the difference between two *independent* population medians is zero. The null assumption is that it is, i.e. the two medians are equal. Variables can be either metric or ordinal. No requirement as to shape of the distributions, but they need to be similar. This is the non-parametric equivalent of the two-sample t test.

Kruskal-Wallis test. Used to test whether the medians of three or more *independent* groups are the same. Variables can be either ordinal or metric. Distributions any shape, but all need to be similar. This non-parametric test is an extension of the Mann-Whitney test.

Wilcoxon test. Used to test whether or not the difference between two *paired* population medians is zero. The null assumption is that it is, i.e. the two medians are equal. Variables can be either metric or ordinal. Distributions any shape, but the *differences* should be distributed symmetrically. This is the non-parametric equivalent of the matched-pairs t test.

Chi-squared test. (χ^2). Used to test whether the proportions across a number of categories of two or more *independent* groups is the same. The null hypothesis is that they are. Variables must be categorical.^a The chi-squared test is also a test of the independence of the two variables (and has a number of other applications). We will deal with the chi-squared test in Chapter 14.

Fisher's Exact test. Used to test whether the proportions in two categories of two *independent* groups is the same. The null hypothesis is that they are. Variables must be categorical. This test is an alternative to the 2×2 chi-squared test, when cell sizes are too small (I'll explain this later).

McNemar's test. Used to test whether the proportions in two categories of two *matched* groups is the same. The null hypothesis is that they are. Variables must be categorical.

^aCategorical will normally be nominal or ordinal, but metric discrete or grouped metric continuous might be used provided the number of values or groups is small.

Interpreting computer hypothesis test results for the difference in two independent population means – the two-sample t test

Since the two-sample t test is one of the more commonly used hypothesis tests, it will be helpful to have a look at the computer output. For example, let's apply the two-sample t test to test the null hypothesis of no difference in the population mean birthweight of maternity-unit-born infants and the mean birthweight of home-born infants (data in Table 10.1). The null hypothesis is:

$$H_0: \mu_M = \mu_H$$

Where, μ_M = population mean birthweight of maternity-unit-born infants, and μ_H = the population mean birthweight of home-born infants.⁴

With SPSS

Look back at Figure 10.1, which shows the output from SPSS, which, in addition to the 95 per cent confidence interval, gives the result of the two-sample t test of the equality of the two population mean birthweights. The test results are given in columns five, six and seven. The column headed 'Sig. (2-tailed)' gives the p -value of 0.407. Since this is not less than 0.05, you cannot reject the null hypothesis. You thus conclude that there is no difference in the two population mean birthweights.

With Minitab

The Minitab output in Figure 10.2 gives the same p -value value as SPSS (0.407), confirming that the two population means are not significantly different.

Some examples of hypothesis tests from practice

Two independent means – the two-sample t test

Table 12.2 shows the baseline characteristics of two independent groups in a randomised controlled trial to compare conventional blood pressure measurement (CBP) and ambulatory blood pressure measurement (ABP) in the treatment of hypertension (Staessen *et al.* 1997). p -values for the differences in the basic characteristics of the two groups are shown in the last column.

The authors used a variety of tests to assess the difference between several parameters for these independent groups (although these are referred to in the text, this information should have been available somewhere in the table itself). To assess the difference in population mean age, and mean body mass index, they used a two-sample t test. For age, the p -value is 0.03, so you can reject the null hypothesis of equal mean ages and conclude that the difference is statistically significant. The p -value for the difference in mean body mass index is 0.39, so you can conclude that the mean body mass index in the two populations is the same.

Exercise 12.2 Comment on what the results in Table 12.2 indicate about the difference between the two populations in terms of their mean serum creatinine and serum total cholesterol levels.

Exercise 12.3 Refer back to Table 1.6, showing the basic characteristics of women in the breast cancer and stressful life events case-control study. Comment on what the p -values tell you about the equality or otherwise, between cases and controls, of the means of the seven metric variables (shown with an * – see table footnote).

⁴Note that *differences in independent percentages* can also be tested with the two-sample t test.

Table 12.2 Baseline characteristics of two *independent* groups, from a randomised controlled trial to compare conventional blood pressure measurement (CBP) and ambulatory blood pressure measurement (ABP) in the treatment of hypertension. Reproduced from *JAMA*, **278**, 1065–72, courtesy of the American Medical Association

Characteristics	CBP Group (<i>n</i> = 206)	ABP Group (<i>n</i> = 213)	<i>P</i>
Age, mean (SD), y	51.3 (11.9)	53.8 (10.8)	.03
Body mass index, mean (SD), kg/m ²	28.5 (4.8)	28.2 (4.4)	.39
Women, No. (%)	102 (49.5)	124 (58.2)	.07
Receiving oral contraceptives, No. (%)*	14 (13.7)	10 (8.1)	.17
Receiving hormonal substitution, No. (%)*	19 (18.6)	19 (15.3)	.51
Previous antihypertensive treatment, No. (%) [†]	134 (65.0)	139 (65.3)	.95
Diuretics, No. (%)*	47 (35.1)	59 (42.4)	.26
β-Blockers, No. (%)*	65 (48.5)	80 (57.6)	.17
Calcium channel blockers, No. (%)*	45 (33.6)	38 (27.3)	.32
Angiotensin-converting enzyme inhibitors, No. (%)*	50 (37.3)	48 (34.5)	.72
Multiple-drug treatment, No. (%)*	62 (46.3)	65 (46.8)	.97
Smokers, No. (%)	42 (20.5)	35 (16.4)	.29
Alcohol use, No. (%)	115 (55.8)	102 (47.9)	.10
Serum creatinine, mean (SD), μmol/L [‡]	85.75 (15.91)	88.4 (16.80)	.25
Serum total cholesterol, mean (SD), mmol/L [‡]	6.00 (1.03)	6.10 (1.19)	.32

*Percentages and values of *P* computed considering only women receiving antihypertensive drug treatment before their enrollment.

[†]Defined as antihypertensive drug treatment within 6 months before the screening visit.

[‡]Divide creatinine by 88.4 and cholesterol by 0.02586 to convert milligrams per deciliter.

Two matched means – the matched-pairs *t* test

Table 10.3 provides an example from practice, and shows the *p*-values for the differences in population mean bone mineral densities between two individually matched groups of depressed and normal women (which we have already discussed in confidence interval terms). As you can see, only at the radius are the population mean bone mineral densities the same, indicated by a *p*-value of 0.25. All the other *p*-values are less than 0.05. Notice that this confirms the confidence interval results.⁵

Two independent medians – the Mann-Whitney test

With two independent groups, and when the data is ordinal or skewed metric, the median is the preferred measure of location. In these circumstances, the Mann-Whitney test can be used to test the null hypothesis that the two population medians are the same.

Recall that in Chapter 10, I introduced the Mann-Whitney procedure to calculate confidence intervals for the difference between two independent population median treatment times. These

⁵ Note that differences in matched percentages can also be tested with the matched-pairs *t* test.

were from a study of the use of ketorolac versus morphine to treat limb injury pain. Table 10.4 contains both 95 per cent confidence intervals and p -values from this study. Only one confidence interval does not include zero, that for the time between receiving analgesia and leaving A&E (4.0 to 39.0). This outcome has a p -value of 0.02, less than 0.05, which confirms the fact that the difference in treatment time between the two population median times is statistically significant.

However there is a problem with the time for preparation of the analgesia. Table 10.4 shows this has a 95 per cent confidence interval of (0 to 5.0), which includes zero, implying no significant difference in treatment times. But the p -value is given as 0.0002, which suggests a highly significant difference in population medians. In the accompanying text the authors indicate that this difference is significant and quote the low p -value, so I can only assume a typographical error in the confidence interval.

Interpreting computer output for the Mann-Whitney test

In view of the widespread use of the Mann-Whitney test you might find it helpful to see the output for this procedure from both SPSS and Minitab.

With SPSS

With the Apgar scores in Table 10.1, you can use the Mann-Whitney test to check if the population median Apgar scores for infants born in a maternity unit and those born at home are the same and differ in the sample only by chance. The null hypothesis is that these medians are equal. The output from SPSS is shown in Figure 12.1. The p -value of 0.061 is labelled 'Asymp. Sig. (2-tailed)'. Since this is not less than 0.05 you *cannot* reject the null hypothesis of no difference in population median Apgar scores between the two groups.

Test Statistics	
	APGARALL
Mann-Whitney U	325.500
Wilcoxon W	790.500
Z	-1.876
Asymp. Sig. (2-tailed)	.061

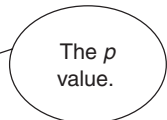


Figure 12.1 Output from SPSS for the Mann-Whitney test of the difference between population medians of the two independent Apgar scores (raw data in Table 10.1)

With Minitab

If you refer back to Figure 10.3, you will see the results of Minitab's Mann-Whitney test three rows from the bottom.⁶ The p -value is given in the second row up as 0.0616 and since this is

⁶ 'ETA' is Minitab's word for the population median.

not less than 0.05 you *cannot* reject the null hypothesis. This is confirmed in the bottom row of the table, and enables you to conclude that the population median Apgar scores are the same in both groups of infants.

Two matched medians – the Wilcoxon test

In the same circumstances as for the Mann-Whitney test described above, but with *matched* populations, the Wilcoxon test is appropriate. Look back at Table 10.5, which was from a matched case-control study into the dietary intake of schizophrenic patients living in the community in Scotland. Here the authors have used the Wilcoxon matched-pairs test to test for differences in the population median daily intakes of a number of substances between ‘All Patients’ and ‘All Controls’. The *p*-values are in the column headed ‘P’. As you can see, the only *p* value *not* less than 0.05 is that for protein (*p*-value = 0.07), so this is the only substance whose median daily intake does *not* differ between the two populations. Once again this confirms the confidence interval results.

Confidence intervals versus hypothesis testing

I said at the beginning of this chapter that where possible, confidence intervals are preferred to hypothesis tests because the confidence intervals are more informative. How so? Have another look at Table 10.4, from the study comparing ketorolac and morphine for limb injury pain. The authors give both 95 per cent confidence intervals and *p*-values for differences in a number of different treatment times, between two groups of limb injury patients. Let’s take the last of these. For the ‘interval between receiving analgesia and leaving A&E’, the *p*-value of 0.02 enables us to reject the null hypothesis, and you would conclude that the difference between the two population median treatment times is statistically significant.

The 95 per cent confidence interval of (4.0 to 39.0) minutes, tells us, *not only* that the difference between the population medians is statistically significant – because the confidence interval does not contain zero – but in *addition*, that the value of this difference in population medians is likely to be somewhere between 4.0 minutes and 39 minutes. So the confidence interval does everything that the hypothesis test does – it tells us if the medians are equal or not, but it *also* gives us extra information – on the likely range of values for this difference. Moreover, unlike a *p*-value, the confidence interval is in *clinically meaningful* units, which helps with the interpretation. So whenever possible, it is good practice to use confidence intervals in preference to *p*-values.

Nobody’s perfect – types of error

Suppose you are investigating a new drug for the treatment of hypertension. Your null hypothesis is that the drug has no effect. Let’s suppose that the drug *does* actually reduce mean systolic blood pressure, but, on average, by only 5 mmHg. However, the hypothesis test you use can only detect a change of 10 mmHg or more. As a consequence, you will not find strong enough

evidence to reject the null hypothesis, and you'll conclude, mistakenly, that the new drug is *not* effective. But the effect is there, it's just that your test does not have enough *power* to detect it.

There are three questions here. First, what exactly is the power of a test and how can we measure it? Second, how can we increase the power of the test we are using? Third, is there a more powerful test that we can use instead? Before I address these questions, a few words on *types of error*.

Whenever you decide either to reject or not reject a null hypothesis, you could be making a mistake. After all, you are basing your decision on *sample* evidence. Even if you have done everything right, your sample could still, by chance, not be very representative of the population. Moreover, your test might not be powerful enough to detect an effect if there is one. There are two possible errors:

Type I error: Rejecting a null hypothesis when it is true. Also known as a *false positive*. In other words, concluding there *is* an effect when there isn't. The probability of committing a type I error is denoted α (alpha), and is the *same* alpha as the significance level of a test.

Type II error: Not rejecting a null hypothesis when it is false. Also known as a *false negative*. That is, concluding there is *no* effect when there is. The probability of committing a type II error is denoted β (beta).

Ideally, you would like a test procedure which minimised the probability of a type I error, because in many clinical situations such an error is potentially serious – judging some procedure to be effective when it is not. When you set the significance level of a test to $\alpha = 0.05$, it's because you want the probability of a type I error to be no more than 0.05. Nonetheless, if there *is* a real effect you would certainly like to detect it, so you also want to minimise the probability of β , a type II error, or put another way, you want to make $(1 - \beta)$ as large as possible.

Exercise 12.4 Explain, with examples, what is meant in hypothesis testing by: (a) a false positive; (b) a false negative.



The power of a test

We can now come back to the three questions above. To answer the first question – the *power* of a test is defined to be $(1 - \beta)$; it is a measure of its capacity to reject the null hypothesis when it is false. In other words, to detect an effect if one is present. In practice, β is typically set at 0.2 or 0.1. This provides power values of 0.80 (or 80 per cent), and 0.90 (or 90 per cent) respectively. So if there *is* an effect, then the probability of the test detecting it is 0.80 or 0.90.

The *power* of a test is a measure of its capacity to reject the null hypothesis when it is false. In other words, its capacity to detect an effect if one is present.

Although you would like to minimise both α and β , unfortunately they are, for a given sample size, linked. You can't make β smaller without making α larger, and vice versa. Thus when you decide a value for α , you are also inevitably fixing the value of β . To answer the second question – the only way to reduce both simultaneously (and increase the power of a test) is to increase the sample size.

To answer the third question, is there a more powerful test? Briefly, parametric tests are more powerful than non-parametric tests (see p. 127 on the meaning of these terms). For example, a Mann-Whitney test has 95 percent of the power of the two-sample t test.⁷ The Wilcoxon matched-pairs test similarly has 95 per cent of the power of the matched-pairs t test. As for the chi-squared test, there is usually no obvious alternative when used for categorical data, so comparisons of power are less relevant, but it is known to be a powerful test. Generally you should of course use the most powerful test that the type of data, and its distributional shape, will allow.

An example from practice

The following is an extract from the RCT of epidural analgesic in the prevention of stump and phantom pain after amputation, referred to in Table 5.3. The authors of the study outline their thinking on power thus:

The natural history of phantom pain after amputation shows rates of about 70%, and in most patients the pain is not severe. Since epidural treatment is an invasive procedure, we decided that a clinically relevant treatment should reduce the incidence of phantom pain to less than 30% at week 1 and then at 3, 6, and 12 months after amputation. Before the start of the study, we estimated that a sample size of 27 patients per group would be required to detect a between-group difference of 40% in the rate of phantom pain (type I error rate 0.05; type II error rate 0.2; power = 0.8).

⁷ In view of the restrictions associated with the two-sample t test, the Mann-Whitney test seems an excellent alternative!

Exercise 12.5 a) Explain, with the help of a few clinical examples, why you would normally want to minimise α , when testing a hypothesis. (b) α is conventionally set to 0.05, or 0.01. Why, if you want to minimise it, don't you set it at 0.001 or 0.000001, or even 0?

Maximising power – calculating sample size

Generally, the bigger the sample, the more powerful the test.⁸ The minimum size of a sample for a given power is determined both by the chosen level of alpha, as well as the power required. The sample size calculation can be summarised thus:

- Decide on the minimum size of the *effect* that would be clinically useful (or otherwise of interest).
- Decide the significance level α , usually 0.05.
- Decide the power required, usually 80 per cent.
- Do the sample size calculation, using some appropriate software, or the rule of thumb described below.

Minitab has an easy to use sample size calculator for the most commonly used tests. Machin, *et al.* (1987) is a comprehensive collection of sample size calculations for a large number of different test situations.

Rules of thumb⁹

Comparing the means of two independent populations (metric data)

The required sample size n is given by the following expression:

$$n = \frac{2 \times \text{s.d.}^2}{E^2} \times k$$

Where s.d. is the population standard deviation (assumed equal in both populations). This can be estimated using the sample standard deviations, if they are available from a pilot study, say. Otherwise the s.d. will have to be guessed using whatever information is available. E is the minimum change in the mean that would be clinically useful or otherwise interesting. k is a magic number which depends on the power and significance levels required, and is obtained from Table 12.3.

⁸ These sample size calculations also apply if you are calculating confidence intervals. Samples that are too small produce wide confidence intervals, sometimes too wide to enable a real effect to be identified.

⁹ I am indebted to Andy Vail for this material.

Table 12.3 Table of magic numbers for sample size calculations

		Power, (1 - β)			
		70 %	80 %	90 %	95 %
Significance level, α	0.05	6.2	7.8	10.5	13.0
	0.01	9.6	11.7	14.9	17.8

For example, suppose you propose to use a case-control study to examine the efficacy of a program of regular exercise, as an alternative to your current drug of choice, in treating moderately hypertensive patients. The minimal difference in mean systolic blood pressures between the cases (given the exercise program), and the controls (given the existing drug), that you think clinically worthwhile is 10 mmHg. You will have to make an intelligent guess as to the standard deviation of systolic blood pressure (assumed the same in both groups – see above). Information on this, and many other measures, is likely to be available from reference sources, from the research literature, from colleagues, etc. Let’s assume systolic blood pressure s.d. = 12 mmHg. If power required is 80 per cent, with a significance level of 0.05, then from Table 12.3, $k = 7.8$, and the sample size required per group is:

$$n = \frac{2 \times 12^2}{10^2} \times 7.8 = 22.5$$

So you will need at least 23 subjects in each of the two groups (always round up to next highest integer) to detect a difference between the means of 10 mmHg. Note that these sample sizes will also be large enough for two matched populations since these require smaller sample sizes for the same power.

Comparing the proportions in two independent populations (binary data)

The required sample size, n , is given by:

$$n = \frac{[P_a \times (1 - P_a)] + [P_b \times (1 - P_b)]}{(P_a - P_b)^2} \times k$$

Where P_a is the proportion with treatment a , P_b is proportion with treatment b , so $(P_a - P_b)$ is the effect size; and k is the magic number from Table 12.3.

For example, suppose the percentage of elderly patients in a large district hospital with pressure sores is currently around 40 per cent, or 0.40. You want to test a new pressure-sore-reducing mattress, and you would like the percentage with pressure sores to decrease to at least 20 per cent, or 0.20. So $P_a = 0.40$, and $(1 - P_a) = 0.60$; $P_b = 0.20$, and $(1 - P_b) = 0.80$; therefore $(P_a - P_b) = (0.40 - 0.20) = 0.20$. If power required is 80 per cent and significance

level $\alpha = 0.05$, then required sample size per group is:

$$n = \frac{(0.40 \times 0.60) + (0.20 \times 0.80)}{0.20^2} \times 7.8 = 78.0$$

Thus you would need at least 78 subjects in each group, which would also be big enough for matched proportions.

Exercise 12.6 In the above examples for: (a) hypertension and (b) the pressure sore example; what sample sizes would be required if power and significance levels were respectively: (i) 90 per cent and 0.05; (ii) 90 per cent and 0.01; (iii) 80 per cent and 0.01?

Exercise 12.7 Suppose you are proposing to use a randomised controlled trial to study the effectiveness of St John's Wort, as an alternative to an existing drug for the treatment of mild to moderate depression. The percentage of patients reporting an improvement in mood three months after existing drug treatment is 70 per cent. You would be satisfied if the percentage reporting mood improvement after three months of St John's Wort was 80 per cent. How big a sample would you require to detect this improvement if you wanted your test to have, (a) 80 per cent power and an α of 0.05; (b) 90 per cent power and an α of 0.01?

13

Testing hypotheses about the *ratio* of two population parameters

Learning objectives

When you have finished this chapter you should be able to:

- Describe the usual form of the null hypothesis in the context of testing the ratio of two population parameters
- Outline the differences between tests of ratios and tests of differences.
- Interpret published results on tests of risk and odds ratios.

Testing the risk ratio

In Chapter 11 you saw that if the confidence interval for a sample risk ratio contains 1, then the population risk ratio is most probably not statistically significant, i.e. not significantly different from 1. This in turn means that the risk factor in question is *not* a statistically significant risk. You can also use the *hypothesis* test approach to find out whether any departure in the sample risk ratio from 1 is statistically significant, or is more likely due to chance. The null hypothesis

is that the population risk ratio equals 1, the alternate hypothesis is that it isn't equal to 1. That is:

$$H_0: \text{population risk ratio} = 1$$

$$H_1: \text{population risk ratio} \neq 1$$

In other words, if H_0 is true, the risk factor in question does not significantly increase or decrease the risk for the condition or disease. If the associated p value is less than 0.05 (or 0.01), you can reject the null hypothesis H_0 , and conclude that the population risk ratio in question is statistically significant, and the risk factor in question is a *statistically significant* risk.

An example from practice

Table 13.1 is from a randomised trial into the efficacy of long-term treatment with subcutaneous heparin in unstable coronary-artery disease (FRISC II Investigators 1999), and shows the risk ratios, 95 per cent confidence intervals and p values for a number of clinical outcomes, in two independent groups, one group given heparin, the other a placebo.

As you can see from the p values in the last column, three out of the six risk ratios were statistically significant: death, myocardial infarction or both, at one month (p value = 0.048); death, myocardial infarction or revascularisation, at one month (p value = 0.001); and death, myocardial infarction or revascularisation, at three months (p value = 0.031). All three of these p values are less than 0.05, the remaining three are all greater than 0.05. Notice that these results are confirmed by the corresponding 95 per cent confidence intervals.

Table 13.1 Risk ratios, 95 per cent confidence intervals, and p -values, for a number of clinical outcomes, at 1 month, 3 months and 6 months, in two independent groups, one group given heparin and the other group a placebo. Reprinted courtesy of Elsevier (*The Lancet*, 1999, Vol No. 354, page 701–7)

Variable	Dalteparin ($n = 1129$)	Placebo ($n = 1121$)	Risk ratio (95% CI)	p
1 month				
Death, MI, or both	70 (6.2%)	95 (8.4%)	0.73 (0.54–0.99)	0.048
Death, MI, or revascularisation	220 (19.5%)	288 (25.7%)	0.76 (0.65–0.89)	0.001
3 months				
Death, MI, or both	113 (10.0%)	126 (11.2%)	0.89 (0.70–1.13)	0.34
Death, MI, or revascularisation	328 (29.1%)	374 (33.4%)	0.87 (0.77–0.99)	0.031
6 months*				
Death, MI, or both	148 (13.3%)	145 (13.1%)	1.01 (0.82–1.25)	0.93
Death, MI, or revascularisation	428 (38.4%)	440 (39.9%)	0.96 (0.87–1.07)	0.50

MI = myocardial infarction.

*Dalteparin ($n = 1115$), placebo ($n = 1103$).

Table 13.2 Relative risk for a number of non-cerebral bleeding complications in patients receiving tenecteplase compared to those receiving alteplase, in the treatment of acute myocardial infarction. Reprinted courtesy of Elsevier (*The Lancet*, 1999, Vol No. **354**, page 716–21)

Complication	Frequency (%)		Relative risk (95% CI)	p
	Tenecteplase (n = 8461)	Alteplase (n = 8488)		
Reinfarction	4.1	3.8	1.078 (0.929–1.250)	0.325
Recurrent angina	19.4	19.5	0.995 (0.935–1.058)	0.877
Sustained hypotension	15.9	16.1	0.988 (0.921–1.058)	0.737
Cardiogenic shock	3.9	4.0	0.965 (0.832–1.119)	0.664
Major arrhythmias	20.5	21.2	0.968 (0.913–1.027)	0.281
Pericarditis	3.0	2.6	1.124 (0.941–1.343)	0.209
Invasive cardiac procedures				
PTCA	24.0	23.9	1.006 (0.953–1.061)	0.843
Stent placement	19.0	19.7	0.968 (0.910–1.029)	0.302
CABG	5.5	6.2	0.884 (0.783–0.999)	0.049
IABP	2.6	2.7	0.968 (0.805–1.163)	0.736
Killip class >I	6.1	7.0	0.991 (0.982–0.999)	0.026
Tamponade or cardiac rupture	0.6	0.7	0.816 (0.558–1.193)	0.332
Acute mitral regurgitation	0.6	0.7	0.886 (0.613–1.281)	0.571
Ventricular septum defect	0.3	0.3	0.817 (0.466–1.434)	0.568
Anaphylaxis	0.1	0.2	0.376 (0.147–0.961)	0.052
Pulmonary embolism	0.09	0.04	2.675 (0.710–10.080)	0.145

PTCA = Percutaneous transluminal coronary angioplasty; CABG = coronary-artery bypass graft; IABP = Intra-aortic balloon pump.

Exercise 13.1 Table 13.2 is from a double blind RCT to assess the efficacy of tenecteplase as a possible alternative to alteplase in the treatment of acute myocardial infarction (ASSENT-2 Investigators 1999). The table contains the risk ratios (relative risks) for a number of in-hospital cardiac events and procedures, for patients receiving tenecteplase, compared to those receiving alteplase.¹

Identify and comment on those cardiac events and procedures for which patients on alteplase had a significant higher risk of experiencing than those on tenecteplase. Note: the key to the cardiac procedures is given in the table footnote. The Killip scale is a classification system for heart failure in patients with acute myocardial infarction, and varies from I (least serious, no heart failure, 5 per cent expected mortality), to IV (most serious, cardiogenic shock, 90 per cent expected mortality).

¹ As a background note: rapid infusion of alteplase, with aspirin and heparin, is the current gold standard for pharmacological reperfusion in acute myocardial infarction. Tenecteplase is a mutant of alteplase with fewer of the limitations of alteplase.

Testing the odds ratio

Here the null hypothesis is that the population odds ratio is not significantly different from 1. That is:

$$H_0: \text{population odds ratio} = 1$$

$$H_1: \text{population odds ratio} \neq 1$$

In other words, in the population, if H_0 is true the risk factor in question does not significantly increase or decrease the odds for the condition or disease. Only if the p value for the sample odds ratio is less than 0.05, can you reject the null hypothesis, and conclude that the risk factor is statistically significant.

An example from practice

Table 13.3 is from an unmatched case-control study into the effect of passive smoking as a risk factor for coronary heart disease (CHD), in Chinese women who had never smoked (He *et al.* 1994). The cases were patients with CHD, the controls women without CHD. The study looked at both passive smoking at home from husbands who smoked, and at work from smoking co-workers. The null hypotheses were that the population odds ratio was equal to 1, both at home and at work, i.e. passive smoking has no effect on the odds for CHD. The

Table 13.3 Odds ratios, 95 per cent confidence intervals and p values, from an unmatched case-control study into the effect of passive smoking as a risk factor for coronary heart disease. The cases were patients with coronary heart disease, the controls individuals without coronary heart disease. Reproduced from *BMJ*, **308**, 380–4, courtesy of BMJ Publishing Group

	Adjusted odds ratio (95% confidence interval)*	P value
Final model (factors 1 to 7):		
1 Age (years)	1.13 (1.04 to 1.22)	0.003
2 History of hypertension	2.47 (1.14 to 5.36)	0.022
3 Type A personality	2.83 (1.31 to 6.37)	0.008
4 Total cholesterol (mg/dl)	1.02 (1.01 to 1.03)	0.0006
5 High density lipoprotein cholesterol (mg/dl)	0.94 (0.90 to 0.98)	0.003
6 Passive smoking from husband	1.24 (0.56 to 2.72)	0.60
7 Passive smoking at work	1.85 (0.86 to 4.00)	0.12
Other model (factors 1 to 5 and passive smoking at work)	1.95 (0.90 to 4.10) [†]	0.087
Other model (factors 1 to 5 and passive smoking from husband or at work, or both)	2.36 (1.01 to 5.55) [†]	0.049

*Adjusted for the other variables in the final model.

[†]Adjusted for the first five variables above; odds ratios for these variables in the other models were essentially the same as those shown above and are not shown.

table contains the adjusted odds ratios for CHD for a number of risk factors, with 95 per cent confidence intervals and p values.

As you can see, the adjusted odds ratio for CHD because of passive smoking from the husband was 1.24, with a p value of 0.60, so you *cannot* reject the null hypothesis. You conclude that passive smoking from husbands is not a statistically significant risk factor for CHD in wives. The same conclusions can be drawn for the odds ratio of 1.85 for passive smoking at work, p value equals 0.12.

Exercise 13.2 In Table 13.3, identify those risk factors which are statistically significant for CHD in the population from whom this sample of women was drawn.

14

Testing hypotheses about the equality of population proportions: the chi-squared test

Learning objectives

When you have finished this chapter you should be able to:

- Describe the rationale underlying the chi-squared hypothesis test.
- Explain the difference between observed and expected values.
- Calculate expected values and the test statistic.
- Outline the procedure for the chi-squared test for the independence of two variables in a population.
- Outline the procedure for the chi-squared test for the equality of two population proportions, and show this is equivalent to the test of the independence of two variables.
- Perform a chi-squared test in 2×2 , 2×3 , 2×4 and 3×4 cases.
- Interpret SPSS and Minitab chi-squared test results.
- Interpret published results of chi-squared tests.
- Outline the procedure for the chi-squared test for trend.

Table 14.1 Observed values in the sample of mothers giving birth in maternity units and at home, and whether they smoked during their pregnancy (raw data in Table 10.1)

		Smoked?		Totals
		Yes	No	
Birthing place	Maternity unit	10	20	30
	Home	6	24	30
	Totals	16	44	60

If the two variables, smoking and birthing place, are unrelated in the population, then there is no reason why this proportion of smokers should be any different ...

... to this proportion. (see text below and footnote).

Of all the tests in all the world . . . the chi-squared (χ^2) test

Two hypothesis tests are prominent in general clinical research. One is the two-sample *t* test, (see Chapter 12), which, as you have seen, is used with metric data to test the equality of two independent population means. The second is the *chi-squared test*¹ (denoted χ^2).² This has two common applications: first as a test of whether two *categorical* variables are independent or not; second, as a test of whether two proportions are equal or not. As you will see, these tests are in fact equivalent.

The chi-squared test is applied to frequency data³ in the form of a contingency table (i.e. a table of cross-tabulations), with the rows representing categories of one variable and the columns categories of a second variable. The null hypothesis is that the two variables are unrelated.

To explain the idea of the chi-squared test, let's refer back to the birthweight data in Table 10.1, and ask the question, 'Is there a relationship between the variables "birthing place" and "whether the mother smoked during pregnancy"?' The relevant data is summarised in the 2×2 table, Table 14.1. The columns of this table represent the two groups 'smokers' and 'non-smokers'. These two groups are *independent* - this is an essential requirement of the chi-squared test.⁴ The rows of the table represent the variable *birthing place* (either maternity unit, or home birth), again independent.

¹ The test is called the chi-squared test because it uses the *chi-squared*, or χ^2 , distribution. If a variable X is Normally distributed, then the variable X^2 has a χ^2 distribution. The χ^2 distribution is very skewed in small samples but becomes more similar in shape to the Normal distribution when samples are large.

² And pronounced as in *Kylie Minogue*

³ The method does not work for tables of percentages, proportions, or measurements.

⁴ If the two groups are matched, then *McNemar's test* is appropriate (see Table 12.1).

If we want to know whether the variables ‘birthing place’ and ‘smoking’ are related, the competing hypotheses will be:

H_0 : Birthing place and smoking status are *not* related in the population, i.e. the two variables are *independent*.

H_1 : Birthing place and smoking status *are* related in the population. The two variables are *dependent*.

Now, if the two variables are unrelated, then there is no reason why the proportion of smokers among mothers giving birth in a maternity unit, should be any different to the proportion of smokers among home-birth mothers.⁵ In other words, these two proportions should be the same. But we have already discussed a method for deciding whether two proportions are the same – by calculating a confidence interval for the difference in two population proportions – see p. 126 in Chapter 10. In fact, the two methods – asking if two variables are independent or if two proportions are the same – are equivalent whenever one of the variables *has only two* categories. However, although we can calculate a confidence interval in the two proportions approach as we saw in Chapter 10, we can’t with the chi-squared approach.

You can see that 10 out of the sample of 30 maternity-unit mothers smoked (a proportion of 0.333), and six out of 30 home-birth mothers smoked (a proportion of 0.2000). These sample proportions are definitely *not* the same, but this could be due to chance.

The crucial question is this, ‘What proportions would we *expect* to find if the null hypothesis of unrelated variables was true?’ The answer is, since we’ve got 16 smokers in a total of 60 women, a proportion of $16/60 = 0.2667$, we would expect to find 0.2667 or 26.67 per cent of the 30 in each category, which is $0.2667 \times 30 = 8$. So you’d expect about eight smokers in each group, rather than the observed values of 10 and six. An easier way to calculate *expected* frequencies is to use the expression:

$$\text{Expected cell frequency} = \frac{\text{Total of row cell is in} \times \text{total of column cell is in}}{\text{overall total frequency}}$$

For example, for the top left-hand cell, the row total is 30, the column total is 16 and the overall total is 60, so the expected value is $(30 \times 16)/60 = 8$. Since in this example the row totals are both 30, this means that the other two cells must each have an expected value of 22. In other words, the two-by-two table you would *expect* to see if the null hypothesis was true is that shown in Table 14.2.

Exercise 14.1 Calculate the expected values for the contingency table of ‘mother smoked’ and ‘Apgar score < 7’, shown in Table 2.11.

⁵ If there *was* a relationship, for example, maternity unit mothers tended to smoke more on average than home-birth mothers, then we would expect to find that the proportion of smokers among these mothers was consistently larger than among home-birth mothers. If there is *no* relationship, i.e. if the two variables are independent, then there is no reason to expect one proportion generally to be any larger or any smaller than the other.

Table 14.2 *Expected* cell values if the null hypothesis of unrelated variables (or equal proportions) is true

		Smoked?		
		Group 1: Yes	Group 2: No	Totals
Place of birth	Maternity unit	8	22	30
	Home	8	22	30
Totals		16	44	60

Are the observed and expected values close enough?

As you've seen, even if the null hypothesis is true, you wouldn't expect the difference between the observed and expected values to be *exactly* zero. But how far away from zero does this difference have to be, before you accept that the sample results are indicative of a true difference in the proportions in the population, rather than chance?

You can use the *chi-squared test* to answer this question: if the p -value associated with the chi-squared test is less than 0.05 (or 0.01), you can reject the null hypothesis and conclude that the two variables are not independent or, put another way, there is a statistically significant difference in the proportions.

The chi-squared test can be used with more than two categories in each variable, but with small sample sizes the maximum number of either is limited by the proviso that none of the *expected* values should be less than 1, and that 80 per cent of expected values should be greater than 5.⁶ There are two ways round the problem of low expected values. First, increase the sample size – usually impractical. Second, amalgamate two or more rows or columns, if this can be done and still make sense.

Calculation of a chi-squared test is not difficult to do by hand if the number of categories is small, but you would have to have available, and know how to use, a table of chi-squared values (I'm assuming here that calculation of the p -value is too difficult by hand, so this is a practical alternative). The procedure is as follows:

- Calculate the expected value E , for each cell in the table.
- For each cell calculate the value of $(O - E)$, where O is the observed value.
- Square each $(O - E)$ value.
- Divide each $(O - E)^2$ value by the E value for that cell.
- Sum all of the values in the previous step.
- Take the square root of the result from the previous step. This result is called the *test statistic*.⁷

⁶ There is some dispute among statisticians about the validity of this condition – some suggest that the chi-squared test still works well even with low expected frequencies.

⁷ For the mathematically minded, the test statistic = $\sqrt{\sum \left\{ \frac{(O - E)^2}{E} \right\}}$.

Table 14.3 Table of critical values for χ^2 test with statistical significance of 0.05. To reject the null hypothesis of unrelated, i.e. independent variables, or of equal proportions, the value of the test statistic must exceed the value in column two for the given table sizes in column one

(No. rows – 1) \times (No. cols – 1)	Value to be exceeded to reject null hypothesis of unrelated variables, or of equal proportions
1	3.84
2	5.99
4	9.49
6	12.59
9	16.92

To reject the null hypothesis of equal proportions, i.e. of independent variables, the value of the test statistic must exceed the critical chi-squared value obtained from a chi-squared table. Some of these values are shown in Table 14.3, for a level of significance of 0.05. For example, the test statistic must *exceed* 3.84 for a 2×2 table. In practice, you will, no doubt, use a computer program to supply the p -value for the chi-squared test, and thus to reject or not reject the null hypothesis that the two variables are independent, i.e. that the proportions are equal across categories.

Exercise 14.2 Calculate the value of the test statistic using the expected values you calculated in Exercise 14.1. With the help of Table 14.3, can you reject the null hypothesis that ‘smoking during pregnancy’ and ‘Apgar scores < 7 ’, are independent? Explain.

An example from practice

Table 14.4 is from the randomised controlled trial into ketorolac versus morphine for the treatment of limb pain (first referred to in Table 10.4) and shows the basic characteristics of the patients participating in the trial. The chi-squared test has been used four times to test whether the proportions (expressed here as percentages) in the ketorolac group and the morphine group are the same. First for ‘the proportion of men’ (categories ‘men’ and ‘not men’); then for ‘fracture site’; then for ‘referred for orthopaedic treatment’; and finally for ‘admitted to hospital’.

As you can see, the chi-squared test applied to the fracture sites data, for example, tests whether the proportions between the two groups is the same for all six sites, and gives rise to a 2×6 table. The p -value is 0.91, which is not less than 0.05, so you can conclude that the null hypothesis of equal proportions cannot be rejected. In fact, the p -value for the chi-squared test on each of the other three items are also all considerably greater than 0.05 indicating no difference between the two groups in any of them.

Table 14.4 Basic characteristics of the patients participating in a randomised controlled trial of ketorolac versus morphine for the treatment of blunt injury limb pain (see Table 10.5). The chi-squared test has been used four times to test whether the proportions in the ketorolac and morphine groups are the same for a number of items. Values are numbers (percentage*) unless stated otherwise. Reproduced from *BMJ*, **321**, 1247–51, by permission of BMJ

Variable	Ketorolac group (n = 75)	Morphine group (n = 73)	P value
Mean (SD) age (years)	53.9 (21.7)	53.2 (21.8)	0.85‡
No (%) of men	38 (51)	33 (45)	0.51§
Mean (SD) body mass index (kg/m ²)	22.8 (3.2)	23.0 (3.7)	0.77‡
Mean (interquartile range) time between injury and arrival at hospital (minutes)	95 (30–630)	82 (33–921)	0.75
Cause of injury:			
Motor vehicle crash	6 (8)	4 (5)	0.58¶
Falls	46 (61)	51 (70)	
Crush	20 (27)	14 (19)	
Other	3 (4)	4 (5)	
Fractures:	50 (67)	48 (66)	0.91§
Clavicle, humerus, elbow	5 (7)	8 (11)	
Radius, ulnar	8 (11)	11 (15)	
Hand	15 (20)	13 (18)	
Femur, patella	14 (19)	12 (16)	
Tibia, fibula	5 (7)	3 (4)	
Foot	2 (3)	1 (1)	
Non-fractures:			
Dislocation, upper limb	2 (3)	1 (1)	
Soft tissue injury, upper limb	10 (13)	10 (14)	
Soft tissue injury, lower limb	14 (19)	14 (19)	
Initial mean (SD) pain score:			
At rest	3.8 (1.1)	3.9 (1.1)	0.65‡
With activity	8.1 (1.2)	8.1 (1.2)	0.85‡
Referred for orthopaedic assessment	41 (55)	36 (49)	0.52§
Admitted to hospital†	38 (51)	29 (40)	0.18§
Admitted with adverse effects	0	3 (4)	

*Percentages may not sum to 100 because of rounding.

†Patient's admitted to hospital (to orthopaedic or emergency observation ward).

‡*t* test for unpaired means comparison.

§ χ^2 test.

¶Fisher's exact test.

Notice that the authors have used *Fisher's Exact* test (see Table 12.1 for a brief description) to compare the equality of the proportions between the two groups for 'cause of injury'. This is almost certainly because of low expected values in some cells.

The chi-squared test for trend

The *chi-squared trend test* is another useful application of the chi-squared distribution, and is appropriate if either variable has categories that *can be ordered*. I can best explain with a real example.

Table 14.5 Numbers of subjects by social class in cases and controls, in a study of stressful life events as a possible risk factor for breast cancer in women

Social class	Malignant diagnosis (cases) group	Benign diagnosis (control) group
I	10	20
II	38	82
III non-manual	28	72
III manual	13	24
IV	11	21
V	3	2
VI	3	4
Totals	106	226

An example from practice

Table 14.5 shows the social class categories (ordinal data) of the cases and controls in the unmatched case-control study of breast cancer in women (refer to Table 1.6). Recall that the subjects were women who attended with a breast lump. The cases were those women who received a malignant diagnosis, the controls those who received a benign diagnosis. These two groups are independent.

With two groups and seven ordered categories of 'social class', we have a 2×7 table. If you apply the chi-squared test here, you are testing whether the proportion of breast cancer cases is the same in each social class category, and simultaneously whether the two variables, diagnosis and social class, are independent. If the proportions are not the same you conclude that the variables are associated in some way.⁸



⁸ Note that to perform the chi-squared test for trend we have to number the categories.

The problem is that if social class *is* associated with diagnosis, then you would *expect* the proportion getting a benign diagnosis to vary systematically, either increasing or decreasing, as social class increased.⁹ In other words, the variability in the proportions may be due largely to this trend, rather than that the variables are associated.

In the chi-squared test for trend, the null hypothesis is that there is *no* trend, and the p -value is used in the usual way. Note that the test statistic for the trend test will always be less than that for the overall test described earlier. However, the trend test may produce a statistically significant result even when the overall test does not. This is because the test for trend is a *more powerful test*. The net result of all this is that if one or both of your variables has ordinal categories, you should use the chi-squared test for trend rather than the overall chi-squared test.

As a matter of interest, the overall chi-squared test for the data in Table 14.5 gives a p -value of 0.784, while the chi-squared trend test gives a p -value of 0.094. As it happens, neither of these is statistically significant, but is an illustration of how different the results from the two tests can be.

Exercise 14.3 Refer back to Table 1.6, the breast cancer and stress case-control study. The table footnote indicates four chi-squared trend tests. Comment on what each p -value reveals about the existence of a trend in the categories of each of the variable concerned.

The chi-squared test has a large number of other applications, one of which we'll meet in Chapter 19.

⁹The direction of change would depend on whether stressful life events were more, or less, common in higher social class groups.

VII

Getting up Close

15

Measuring the association between two variables

Learning objectives

When you have finished this chapter you should be able to:

- Explain the meaning of association.
- Draw and interpret a scatterplot, and from it assess the linearity, direction and strength of an association.
- Distinguish between negative and positive association.
- Explain what a correlation coefficient is.
- Describe Pearson's correlation coefficient r , its distributional requirements, and interpret a given value of r .
- Describe Spearman's correlation coefficient r_s and interpret a given value of r_s .
- Describe the circumstances under which Pearson's r or Spearman's r_s is appropriate.

Association

When we say that two ordinal or metric variables are *associated*, we mean that they behave in a way that makes them appear 'connected' - changes in either variable seem to coincide with

changes in the other variable. It's important to note (at this point anyway), that we are not suggesting that change in either variable is *causing* the change in the other variable, simply that they exhibit this commonality. As you will see, association, if it exists, may be *positive* (low values of one variable coincide with low values of the other variable, and high values with high values) or *negative* (low values with high values and vice versa).

In this chapter, I want to discuss two alternative methods of detecting an association. The first method relies on a plot of the sample data, called a *scatterplot*, in which values of one variable are plotted on the vertical axis and values of the other on the horizontal axis. The second approach is numeric, making both comparison and inference possible.

The scatterplot

A scatterplot will enable you to see if there is an association between the variables, and if there is, its strength and direction. But the scatterplot will only provide a *qualitative* assessment, and thus has obvious limitations. First, it's not always easy to say which of two sample scatterplots indicates the stronger association and second, it doesn't allow us to make *inferences* about possible associations in the population.

An example from practice

As part of a study of the possible association between Crohn's disease (CD) and ulcerative colitis (UC), researchers in Canada (Blanchard *et al.* 2001) produced the scatterplot shown in Figure 15.1. It doesn't matter which variable is plotted on which axis for the scatterplot itself, but in the study of causal *relationships* between variables (which I will discuss in Chapter 17), the choice of axis becomes more important.

Looking at the scatterplot it's not difficult to see that something is going on here. The scatter is not just a random cloud of points, but appears to display a pattern – low CD levels seem to be associated with low UC levels, and higher CD levels with high UC levels. You could justly claim that the two variables appear to be *positively associated*.

As a second example, Figure 15.2 shows a scatterplot taken from a study into the possible relationship between percentage mortality from aortic aneurysm, and the number of aortic aneurysm episodes dealt with per year, in each of 22 hospitals (McKee and Hunter 1995). This scatterplot displays a *negative association* between the two variables, low values for number of episodes seem to be associated with high values for percentage mortality, and vice versa.

As a final example from practice, Figure 15.3 shows a scatterplot taken from the cross-section study into the possible contribution of channel blockers (prescribed for depression), to the suicide rate in 284 Swedish municipalities (Lindberg *et al.* 1998), first referred to in Figure 3.10. The scatterplot here is very much more fuzzy than the two previous plots, and it would be hard to claim, merely from eyeballing it, that there is any notable association between the two variables (although admittedly there is some evidence of a rather weak positive association).

When you set out to investigate a possible association between two variables, a scatterplot is almost always worthwhile, and will often produce an insight into the way the two variables co-behave. In particular, it may reveal whether an association between them is *linear*. The

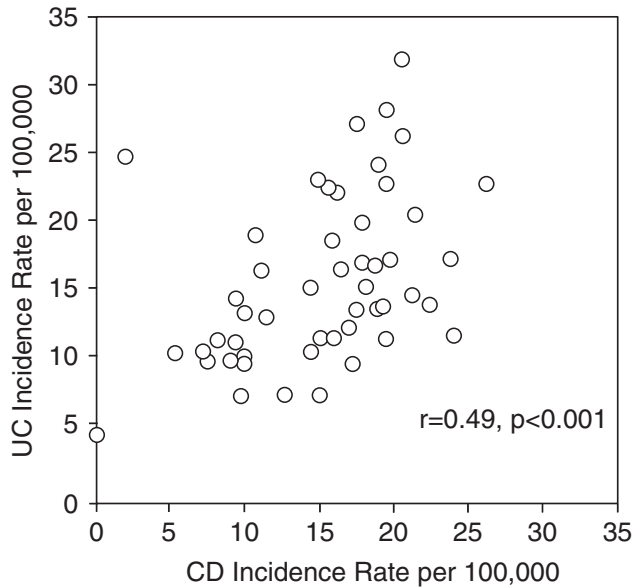


Figure 15.1 Scatterplot of the age-standardised incidence rates of Crohn's disease (CD) and ulcerative colitis (UC) by Manitoba postal area, Canada, 1987–1996. The scatterplot suggests a positive association between the two variables. Reproduced from *American Jnl of Epidemiology* 2001, **154**: 328–33, Fig. 3 p. 331, by permission of OUP

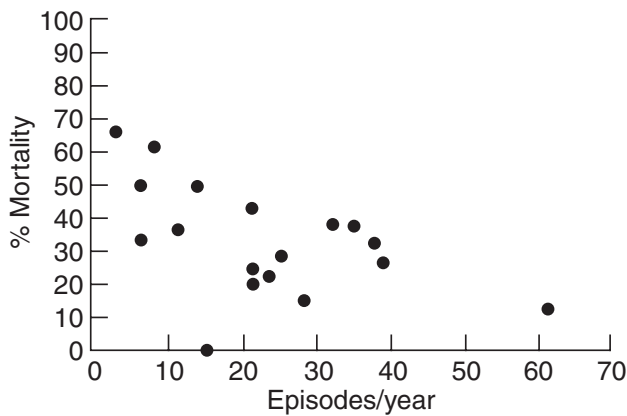


Figure 15.2 A scatterplot of percentage mortality from aortic aneurysm, and number of aortic aneurysm episodes dealt with per year, in 22 hospitals. The plot suggests a negative association between the two variables. Reproduced from *Quality in Health Care*, **4**, 5–12, courtesy of BMJ Publishing Group

property of linearity is important in some branches of statistics and we'll meet it again ourselves in Chapter 17. Put simply, a linear association is one in which the points in the scatterplot seem to cluster around a straight line. The two scatterplots in Figure 15.4 illustrate the difference between a linear and a non-linear association. The scatter in Figure 15.4a seems to be linear; but in Figure 15.4b it shows some curviness.

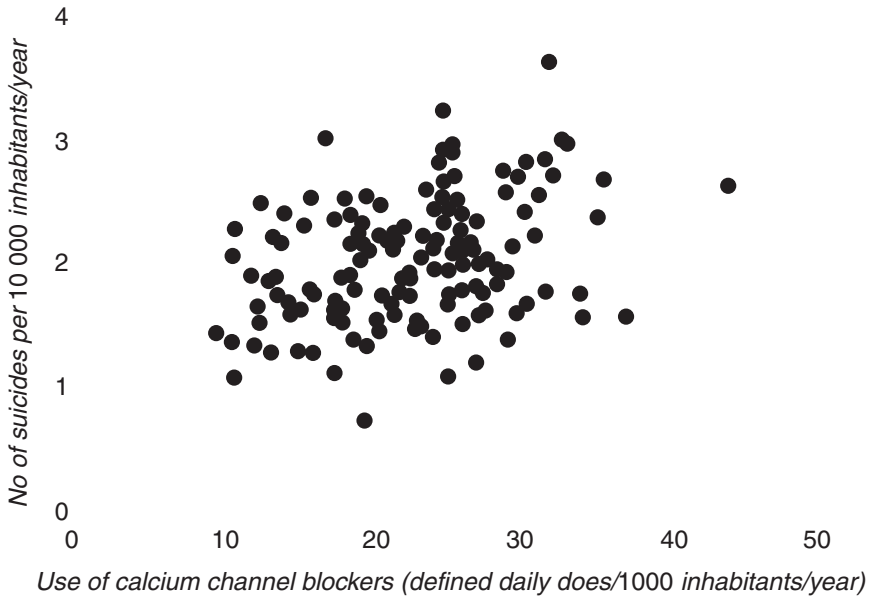


Figure 15.3 A scatterplot taken from a cross-section study into the possible contribution of channel blockers (prescribed for depression) to the suicide rate, in 284 Swedish municipalities. The plot suggests a weak, if any, relationship between the variables. Reproduced courtesy of BMJ Publishing Group

Exercise 15.1 Draw a scatterplot of Apgar score against birthweight for the 30 maternity-unit born infants using the data in Table 2.5, and comment on what it shows about any association between the two variables.

Exercise 15.2 The scatterplot in Figure 15.5 is from a study into the effect of passive smoking on respiratory symptoms (Janson *et al.* 2001). In addition, the ‘best’ straight line has been drawn through the points.¹ Comment on what the scatterplot suggests about the nature and strength of any association between the two variables.

Exercise 15.3 The scatterplot of percentage body fat against body mass index (bmi) in Figure 15.6 is from a cross-section study into the relationship between body mass index and body fat, in black populations in Nigeria, Jamaica and the USA (Luke *et al.* 1997). The aim of the study was to investigate whether per cent body fat rather than bmi could be used as a measure of obesity. What does the scatterplot tell you about the nature and strength of any association between these two variables?

¹I’ll have more to say about what constitutes the *best* straight line in Chapter 17, but loosely speaking, it’s the line which passes as close as possible to all the points.

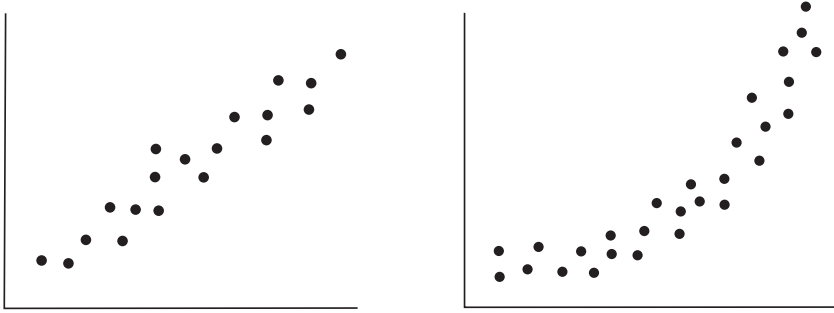


Figure 15.4 (a) A linear association (b) A non-linear association

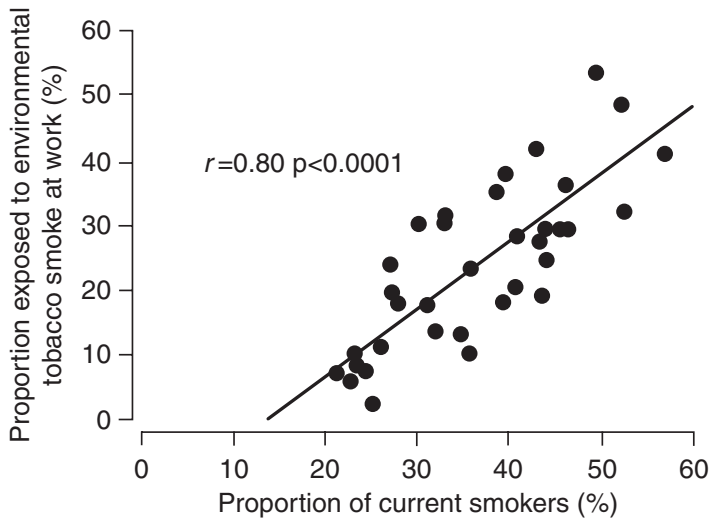


Figure 15.5 Scatterplot from a study into the effect of passive smoking on respiratory symptoms. Reprinted courtesy of Elsevier (*The Lancet* 2001, **358**, 2103–9, Fig. 1, p. 2105)

The correlation coefficient

The principal limitation of the scatterplot in assessing association is that it does not provide us with a *numeric* measure of the strength of the association; for this we have to turn to the *correlation coefficient*. Two correlation coefficients are widely used: *Pearson's* and *Spearman's*.

Pearson's correlation coefficient

Pearson's product-moment correlation coefficient, denoted ρ (Greek rho), in the population, and r in the sample, measures the strength of the *linear* association between two variables. Loosely speaking, the correlation coefficient is a measure of the average distance of all of the points from an imaginary straight line drawn through the scatter of points (analogous to the standard deviation measuring the average distance of each value from the mean).

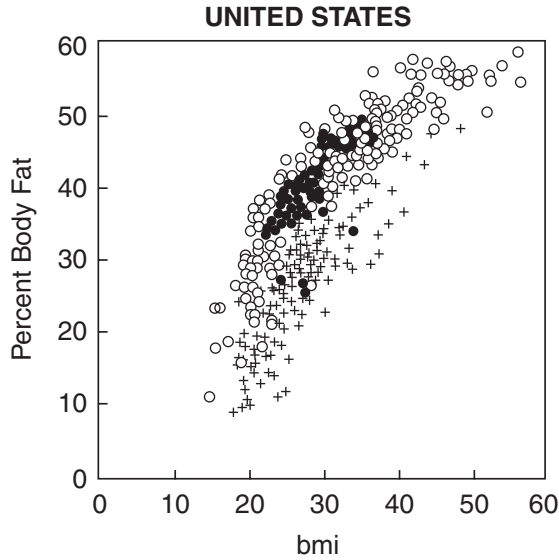


Figure 15.6 Scatterplot of per cent body fat against body mass index from a cross-section study into the relationship between bmi and body fat, in black population samples from Nigeria, Jamaica and the USA. Reproduced from *Amer. J. Epid.*, **145**, 620–8, courtesy of Oxford University Press

For Pearson's correlation coefficient to be appropriately used, both variables must be *metric continuous* and, if a confidence interval is to be determined, also approximately *Normally* distributed. The value of Pearson's correlation coefficient can vary as follows: from -1 , indicating a perfect negative association (all the points lie exactly on a straight line); through 0 , indicating no association; to $+1$, indicating perfect positive association (all points exactly on a line). In practice, with real sample data, you will never see values of -1 , 0 or $+1$. Calculation of r by hand is very tedious and prone to error, so we will avoid it here. But it can be done in a flash with a computer statistics program, such as SPSS or Minitab.

Is the correlation coefficient statistically significant in the population?

To assess the statistical significance of a *population* correlation coefficient and hence decide whether there is a statistically significant association between the two variables, you can either perform a hypothesis test (is the p -value less than 0.05 ?), or calculate a confidence interval (does it include zero?). For the hypothesis test, the hypotheses are:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

For example, for the data shown in the scatterplot in Figure 15.1, the sample $r = 0.49$, with a p -value < 0.001 . This indicates a statistically significant positive association in the population between incidence rate of Crohn's disease and ulcerative colitis.

A useful rule of thumb if you have a value for r but no confidence interval or p -value, is that to be statistically significant, r must be greater than $2/\sqrt{n}$, where n is the sample size. For example, if $n = 100$, then r has to be greater than $2/10 = 0.200$ to be statistically significant.

An example from practice

Table 15.1 is taken from the same cross-section study as Exercise 15.3, and shows the sample Pearson's correlation coefficient for the association between bmi and per cent body fat, with blood pressure, and waist and hip measurements, along with an indication of the statistical significance or otherwise of the p -value.

Unfortunately, the authors have not given the actual p -values, but only indicated whether they were less than 0.05 or less than 0.01. This is not good practice; the actual p -values should always be provided. As you can see, the population correlation coefficient between both bmi and per cent body fat, with waist and hip circumference, is positive and statistically significant in every case. However, bmi is more *closely* associated (higher r values) than body fat, except in Jamaican men. Apart from the association with systolic blood pressure in US males, there is no statistically significant association with either of the blood pressure measurements.

Exercise 15.4 Table 15.2 is from a case-control study of medical record validation (Olson *et al.* 1997), and shows the value of Pearson's r , and the 98 per cent confidence intervals, for the correlation between gestational age, as estimated by the mother, and as determined from medical records, for a number of demographic sub-groups (ignore the last column). The cases were the mothers of child leukaemia patients, the matched controls were randomly selected by random telephone calling. Identify: (a) any correlation coefficients not statistically significant; (b) the strongest correlation; (c) the weakest correlation.

Spearman's rank correlation coefficient

If *either* (or both) of the variables is ordinal, then *Spearman's rank correlation coefficient* (usually denoted ρ_s in the population and r_s in the sample) is appropriate. This is a non-parametric measure. As with Pearson's correlation coefficient, Spearman's correlation coefficient varies from -1 , through 0 , to $+1$, and its statistical significance can again be assessed with a p -value or a confidence interval. The null hypothesis is that the population correlation coefficient $\rho_s = 0$. Spearman's r_s is not quite as bad to calculate by hand as Pearson's r but bad enough, and once again you would want to do it with the help of a computer program.

An example from practice

Table 15.3 is from the same cross-section study first referred to in Figure 4.3, into the use of the Ontario mammography services. The authors wanted to know whether the variation in

Table 15.1 Correlation coefficients from a cross-section study into the relationship between body mass index (bmi) and body fat, in black population samples from Nigeria, Jamaica, and the USA. The aim of the study was to investigate whether body fat rather than bmi could be used as a measure of obesity.^{†,§} Reproduced from *Amer. J. Epid.*, **145**, 620–8, courtesy of Oxford University Press

Variable	Women						Men					
	Nigeria		Jamaica		United States		Nigeria		Jamaica		United States	
	BMI	% fat	BMI	% fat	BMI	% fat	BMI	% fat	BMI	% fat	BMI	% fat
Waist circumference	0.90 ^{**}	0.77 ^{**}	0.87 ^{**}	0.77 ^{**}	0.91 ^{**}	0.85 ^{**}	0.89 ^{**}	0.79 ^{**}	0.69 ^{**}	0.76 ^{**}	0.93 ^{**}	0.83 ^{**}
Hip circumference	0.93 ^{**}	0.81 ^{**}	0.91 ^{**}	0.82 ^{**}	0.93 ^{**}	0.87 ^{**}	0.89 ^{**}	0.76 ^{**}	0.64 ^{**}	0.72 ^{**}	0.93 ^{**}	0.82 ^{**}
Systolic blood pressure	0.24	0.24	0.16	0.15	0.21	0.21	0.09	0.09	0.24	0.24	0.24 [*]	0.23 [*]
Diastolic blood pressure	0.16	0.14	0.20	0.16	0.07	0.10	0.31	0.24	0.16	0.11	0.22	0.20

* $p < 0.05$; ** $p < 0.01$.

† Weight (kg)/height (m)².

#Data were adjusted for age.

§No significant difference was found between correlation coefficients for body mass index and percentage of body fat.

the ranked utilisation rates (number of visits per 1000 women) was similar across the age groups. They did this by measuring the strength of the association between the ranked rates for each pair of different age groups. When the association was strong and significant, they concluded that the variation in the usage rate was similar.

The results show that the r_s for the association between the ranked usage rates for 30–39 year-olds, and the 40–49 year-olds, across the 33 districts was 0.6496 (first row of table), with a p -value of 0.0005. So this association is positive and statistically significant in these two age group populations. Indeed, the correlation coefficients between all pairs of age groups are statistically significant, with all p -values < 0.05 . The authors thus concluded that variation in

Table 15.2 Pearson's r and 98 per cent confidence intervals for the association between gestational age, as estimated by the mother and from medical records, for a number of demographic sub-groups. Reproduced from *Amer. J. Epid.*, **145**, 58–67, courtesy of Oxford University Press

	Correlation of gestational age	98% CI*	Kappa statistic†
All gestational ages	0.839	0.817–0.859	0.62
Case/control status			
Cases	0.849	0.813–0.878	0.63
Controls	0.835	0.805–0.861	0.61
Education			
<High school	0.694	0.553–0.797	0.51
High school	0.833	0.790–0.868	0.63
>High school	0.835	0.804–0.861	0.62
Household income			
<\$22,000	0.791	0.734–0.837	0.59
\$22,000–\$ 34,999	0.882	0.849–0.908	0.62
≥\$35,000	0.843	0.800–0.877	0.65
Unknown	0.745	0.641–0.823	0.60
Time (years) from delivery to interview			
<2	0.896	0.862–0.921	0.64
2–3.9	0.821	0.784–0.852	0.63
4–5.9	0.828	0.775–0.869	0.61
6–8	0.852	0.734–0.920	0.42
Maternal age (years)			
<25	0.822	0.773–0.861	0.64
25–29	0.889	0.862–0.912	0.63
30–34	0.760	0.694–0.813	0.57
≥35	0.888	0.824–0.930	0.64
Birth order			
First born	0.880	0.853–0.903	0.67
Second born	0.815	0.778–0.846	0.57
≥Third born	0.632	0.416–0.781	0.52
Maternal race			
White	0.846	0.822–0.866	0.64
Other	0.782	0.680–0.855	0.42

* CI, confidence interval.

† Three categories, <38, 38–41, ≥42 weeks.

Table 15.3 Spearman correlation coefficients from a cross-section study of the use of the Ontario mammography services in relation to age. Each correlation coefficient measures the strength of the association in the variation between the ranked usage rate across the 33 health districts for each pair of age groups. Reproduced from *J. Epid. Comm. Health*, **51**, 378–82, courtesy of BMJ Publishing Group

Age group (y)	30–39y	40–49y	50–69y	70+y
30–39	1.0000	0.6496 (p < 0.0001)	0.5949 (p = 0.0005)	0.5488 (p = 0.0014)
40–49		1.0000	0.9021 (p < 0.0001)	0.8985 (p < 0.0001)
50–69			1.0000	0.9513 (p < 0.0001)
70+				1.0000

usage rate was similar for the four age groups across the 33 health districts. However, whether association is the correct way to measure similarity in two sets of values is a question I will return to in the next chapter.

Two other correlation coefficients can only be mentioned briefly. Kendall's rank-order correlation coefficient, denoted τ (tau), is appropriate in the same circumstances as Spearman's r_s , i.e. with ranked data (which may be ordinal or continuous). Tau is available in SPSS, but not in Minitab. The *point-biserial* correlation coefficient is appropriate if one variable is metric continuous and the other is truly dichotomous (which means that the variable can take only two values; alive or dead, male or female, etc.). Unfortunately, this latter measure of association is not available in either SPSS or Minitab.

If you plan to use a correlation coefficient you should ensure that the assumptions referred to above are satisfied, in particular that the association is linear - which can be checked by a scatterplot. Moreover, with Pearson's correlation coefficient you should interpret any results with suspicion if there are outliers present in either data set, since these can distort the results.

Finally it is worth noting again that just because two variables are significantly associated, does *not* mean that there is a cause-effect *relationship* between them.

16

Measuring agreement

Learning objectives

When you have finished this chapter you should be able to:

- Explain the difference between association and agreement.
- Describe Cohen's kappa, calculate its value and assess the level of agreement.
- Interpret published values for kappa.
- Describe the idea behind ordinal kappa.
- Outline the Bland–Altman approach to measuring agreement between metric variables.

To agree or not agree: that is the question

Association is a measure of the inter-connectedness of two variables; the degree to which they tend to change together, either positively or negatively. *Agreement* is the degree to which the values in two sets of data actually *agree*. To illustrate this idea look at the hypothetical data in Table 16.1, which shows the decision by a psychiatrist and by a psychiatric social worker (PSW) whether to section (Y), or not section (N), each of 10 individuals with mental ill-health. We would say that the two variables were in perfect *agreement* if every pair of values were the same.

Table 16.1 Decision by a psychiatrist and a psychiatric social worker whether or not to section 10 individuals suffering mental ill-health

Patient	1	2	3	4	5	6	7	8	9	10
Psychiatrist	Y	Y	N	Y	N	N	N	Y	Y	Y
PSW	Y	N	N	Y	N	N	Y	Y	Y	N

In practical situations this won't happen, and here you can see that only seven out of the 10 decisions are the same, so the *observed* level of *proportional agreement* is 0.70 (70 per cent).

Cohen's kappa

However, if you had asked each clinician simply to toss a coin to make the decision (heads – section, tails – don't section), some of their decisions would probably still have been the same – by *chance* alone. You need to adjust the observed level of agreement for the proportion you would have *expected* to occur by chance alone. This adjustment gives us the *chance-corrected proportional agreement statistic*, known as *Cohen's kappa*, κ :

$$\kappa = \frac{(\text{proportion of observed agreement} - \text{proportion of expected agreement})}{(1 - \text{proportion of expected agreement})}$$

We can calculate the *expected* values using a contingency table in exactly the same way as we did for chi-squared (row total \times column total \div overall total – see Chapter 14). Table 16.2 shows the data in Table 16.1 expressed in the form of a contingency table, with the psychiatrist's scores in the rows, the PSW's scores in the columns, and with row and column totals added. The expected values are shown in brackets in each cell.

Table 16.2 Contingency table showing observed (and *expected*) decisions by a psychiatrist and a psychiatric social worker on whether to section 10 patients (data from Figure 16.1)

		Psychiatric Social Worker		Totals
		Yes	No	
Psychiatrist	Yes	4 (3)	2 (3)	6
	No	1 (2)	3 (2)	4
Totals		5	5	10

Expected value:
(5 \times 6)/10 = 3

We have seen that the *observed* agreement is 0.70, and we can calculate the expected agreement to be 5 out of 10 or 0.50.¹ Therefore:

$$\kappa = (0.70 - 0.50)/(1 - 0.50) = 0.20/0.50 = 0.40$$

¹ We can expect the two clinicians to agree on 'Yes' three times, and 'No' two times, making five agreements in total.

Table 16.3 How good is the agreement – assessing kappa

Kappa	Strength of agreement
≤0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good

So after allowing for chance agreements, agreement is reduced from 70 per cent to 40 per cent. Kappa can vary between zero (agreement no better than chance), and one (perfect agreement), and you can use Table 16.3 to assess the quality of agreement. It's possible to calculate a confidence interval for kappa, but these will usually be too narrow (except for quite small samples) to add much insight to your result.

An example from practice

Table 16.4 is from a study into the development of a new quality of life scale for patients with advanced cancer and their families – the Palliative Care Outcome scale (POS) (Hearn *et al.* 1998). It shows agreement between the patient and staff (who also completed the scale questionnaires) for a number of items on the POS scale. The table also contains values of Spearman's r_s , and the proportion of agreements within one point on the POS scale. The level of agreement between staff and patient is either fair or moderate for all items, and agreement within one point is either good or very good.

Table 16.4 From a palliative care outcome scale (POS) study showing levels of agreement between the patient and staff assessment for a number of items on the POS scale. Reproduced from *Quality in Health Care*, **8**, 219–27, courtesy of BMJ Publishing Group

Item	No of patients	Patient score (% severe)	Staff score (% severe)	K	Spearman correlation	Proportion agreement within 1 score
At first assessment: 145 matched assessments						
Pain	140	24.3	20.0	0.56	0.67	0.87
Other symptoms	140	27.2	26.4	0.43	0.60	0.86
Patient anxiety	140	23.6	30.0	0.37	0.56	0.83
Family anxiety	137	49.6	46.0	0.28	0.37	0.72
Information	135	12.6	13.4	0.39	0.36	0.79
Support	135	10.4	14.1	0.22	0.32	0.79
Life worthwhile	133	13.6	16.5	0.43	0.54	0.82
Self worth	132	15.9	23.5	0.37	0.53	0.82
Wasted time	135	5.9	6.7	0.33	0.32	0.95
Personal affairs	129	7.8	13.2	0.42	0.49	0.96

Table 16.5 Injury Severity Scale (ISS) scores given from case notes by two experienced trauma clinicians to 16 patients in a major trauma unit. Reproduced from *BMJ*, **307**, 906–9. by permission of BMJ Publishing Group

Observer no.	Case no.															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	9	14	29	17	34	17	38	13	29	4	29	25	4	16	25	45
2	9	13	29	17	22	14	45	10	29	4	25	34	9	25	8	50

Exercise 16.1 Do the highest and the lowest levels of agreement in Table 16.4 coincide with the highest and lowest levels of correlation? Will this always be the case?

Exercise 16.2 Table 16.5 is from a study in a major trauma unit into the variation between two experienced trauma clinicians in assessing the degree of injury of 16 patients from their case notes (Zoltie *et al.* 1993). The table shows the Injury Severity Scale (ISS) score awarded to each patient.² Categorise the scores into two groups: ISS scores of less than 16, and of 16 or more. Express the results in a contingency table, and calculate: (a) the observed and expected proportional agreement; (b) kappa. Comment on the level of agreement.

A limitation of kappa is that it is sensitive to the proportion of subjects in each category (i.e. to prevalence), so caution is needed when comparing kappa values from different studies – these are only helpful if prevalences are similar. Moreover, Cohen’s kappa as described above is only appropriate for *nominal* data, as in the sectioning example above, although most data can be ‘nominalised’, like the ISS values above. In the next paragraph, however, I describe, briefly, a version of kappa which can handle ordinal data.

Measuring agreement with ordinal data – weighted kappa

The idea behind weighted kappa is best illustrated by referring back to the data in Table 16.5. The two clinician’s ISS scores agree for only five patients. So the proportional observed agreement is only $5/16 = 0.3125$ (31.25 per cent). However, in several cases the scores have a ‘near miss’; patient 2, for example, with scores of 14 and 13. Other pairs of scores are further apart, patient 15 is given scores of 25 and 8! Weighted kappa gives credit for near misses, but its calculation is too complex for this book.

Measuring the agreement between two metric continuous variables

When it comes to measuring agreement between two metric continuous variables the obvious problem is the large number of possible values – it’s quite possible that *none* of them will be

² The ISS is used for the assessment of severity of injury, with a range from 0 to 75. ISS scores of 16 or above indicate potentially life-threatening injury, and survival with ISS scores above 51 is considered unlikely.



the same. One solution is to use a Bland-Altman chart (Bland and Altman 1986). This involves plotting, for each pair of measurements, the *differences* between the two scores (on the vertical axis) against the *mean* of the two scores (on the horizontal axis).

A pair of tramlines, called the 95 per cent *limits of agreement*, are drawn a distance of two s_d above and below the zero difference line (where s_d = standard deviations of the differences). If

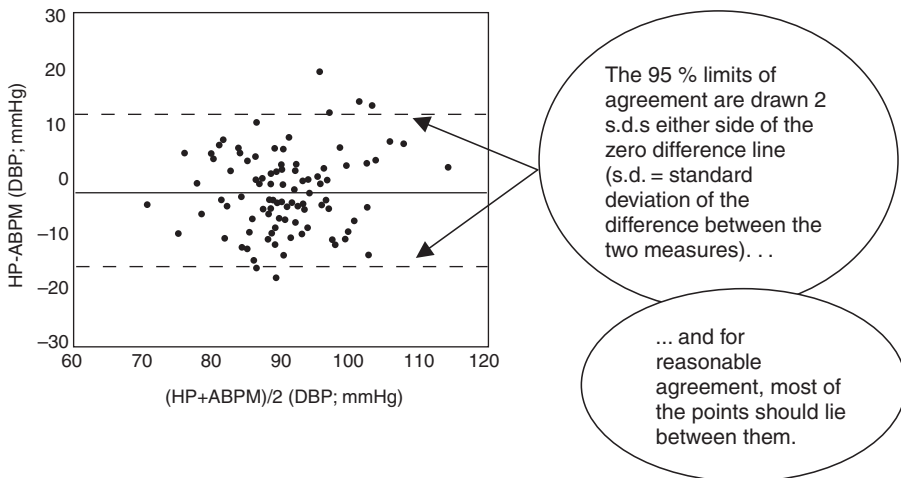


Figure 16.1 A Bland-Altman chart to measure agreement between two metric continuous variables; diastolic blood pressure as measured by patients at home with a cuff-measuring device (HP), and as measured by the same patients using an ambulatory device (ABPM). Reproduced from *Brit. J. General Practice*, **48**, 1585–9, courtesy of the Royal College of General Practitioners

all of the points on the graph fall between the tramlines, then agreement is 'acceptable', but the more points there are outside the tramlines, the less good the agreement. Moreover the spread of the points should be reasonably horizontal, indicating that differences are not increasing (or decreasing) as the values of the two variables increase.

An example from practice

The idea is illustrated in Figure 16.1, for agreement between two methods of measuring diastolic blood pressure (Brueren *et al.* 1998). In this example, there are only a few points outside the ± 2 standard deviation tramlines and the spread of points is broadly horizontal. We would assess this chart as suggesting reasonably good agreement between the two methods of blood pressure measurement.

The continuous horizontal line across the middle of the chart represents the mean of the differences between the two measures. Note that this is below the zero mark indicating some bias in the measures. It looks as if the ABPM values are greater on the whole than the HP values.

To sum up, two variables that are in reasonable agreement will be strongly associated, but the opposite is not necessarily true. The two measures are not equivalent; association does not measure agreement.

VIII

Getting into a Relationship

17

Straight line models: linear regression

Learning objectives

When you have finished this chapter you should be able to:

- Describe the difference between an association and a cause-and-effect relationship.
- Estimate the equation of a straight line from a graph, and draw a straight line knowing its equation.
- Describe what is meant by a linear relationship and how the linear regression equation can be used to model it.
- Identify the constant and slope parameters, and the dependent and independent variables.
- Explain the role of the residual term.
- Summarise the model building process.
- Provide a brief explanation of the idea behind the method of ordinary least squares estimation.
- List the basic assumptions of the simple linear regression model.

- Interpret computer-generated linear regression results.
- Explain what goodness-of-fit is and how it is measured in the simple linear regression model.
- Explain the role of \bar{R}^2 in the context of multiple linear regression.
- Interpret published multiple linear regression results.
- Explain the adjustment properties of the regression model.
- Outline how the basic assumptions can be checked graphically.

Health warning!

Although the maths underlying the idea of linear regression is a little complicated, some explanation of the idea is necessary if you are to gain any understanding of the procedure and be able to interpret regression computer outputs sensibly. I have tried to keep the discussion as brief and as non-technical as possible, but if you have an aversion to maths you might want to skim the material in the next few pages.

Relationship and association

In Chapter 15, I emphasised the fact that an association between two variables does *not* mean that there is a cause-and-effect relationship between them. For example, body mass index and systolic blood pressure may appear to be closely associated, but this does *necessarily* mean that an increase in body mass index will *cause* a corresponding increase in systolic blood pressure (or indeed the other way round). In this chapter and the next, I am going to deal with the idea of a *causal* relationship between variables, such that changes in the value of one variable bring about or *cause* changes in the value of another variable. Or to put it another way, variation among a group of individuals in say their blood pressure is caused, or explained, by the variation among those same individuals in their body mass index.

In the clinical world demonstrating a cause–effect relationship is difficult, and requires a number of conditions to be satisfied; the relationship should be plausible, repeatable, predictable, with a proved mechanism, and so on. I will assume in the remainder of this chapter that a cause-effect relationship between the variables has been satisfactorily demonstrated, and that this relationship is *linear* (see pp. 172/3 for an explanation of linearity).

A causal relationship – explaining variation

Let's begin with a simple example. Suppose that systolic blood pressure (SBP), in mmHg, is effected by body mass index (bmi) in kg/m^2 , and the two variables are related by the following expression:

$$\text{SBP equals } 110 \text{ plus } \frac{3}{4} \text{ of bmi}$$

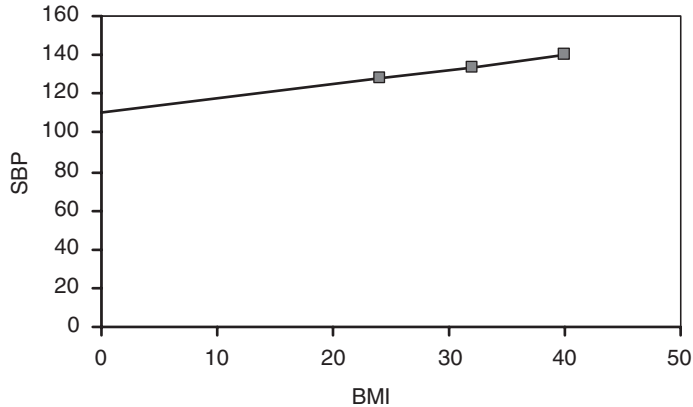


Figure 17.1 A plot of systolic blood pressure (SBP) against body mass index (bmi) produces a straight line, and shows that the relationship between the two variables is linear

So for example, when $\text{bmi} = 40$, SBP equals 110 plus $\frac{3}{4}$ of 40, or 110 plus 30, which equals 140. This equation is a *linear* equation. If you plot it with pairs of values of bmi and SBP, you will see a straight line. For instance, when $\text{bmi} = 24$, $\text{SBP} = 128$, and when $\text{bmi} = 32$, $\text{SBP} = 134$. We already know that when $\text{bmi} = 40$, $\text{SBP} = 140$, and if we plot these three pairs of values, and draw a line through them, we get Figure 17.1. This is clearly a straight line.

We can write the above expression more mathematically as an equation:

$$\text{SBP} = 110 + 0.75 \times \text{bmi}$$

This equation explains the *variation* in systolic blood pressure from person to person, in terms of corresponding *variation* from person to person in body mass index. I have referred to this relationship as an equation, but I could also have described it as a *model*. We are *modelling* the variation in systolic blood pressure in terms of corresponding variation in body mass index. We can write this equation in a more general form in terms of two variables Y and X , thus:¹

$$Y = b_0 + b_1 X$$

The term b_0 is known as the *constant coefficient*, or the coefficient of intersection – it's where the line cuts the Y axis (110 in our Figure 17.1). The term b_1 is known as the *slope coefficient*, (0.75 in our equation), and will be positive if the line slopes upwards from left to right (as in Figure 17.1), and negative if the line slopes down from left to right (as in Figure 15.2). Higher values of b_1 means more steeply sloped lines. One important point: the value of b_1 (+ 0.75 in the example) is the amount by which SBP would change if the value of bmi *increased* by 1 unit. I'll come back to this later.

¹ You may remember this from school as: $y = mx + c$, or some other variation.

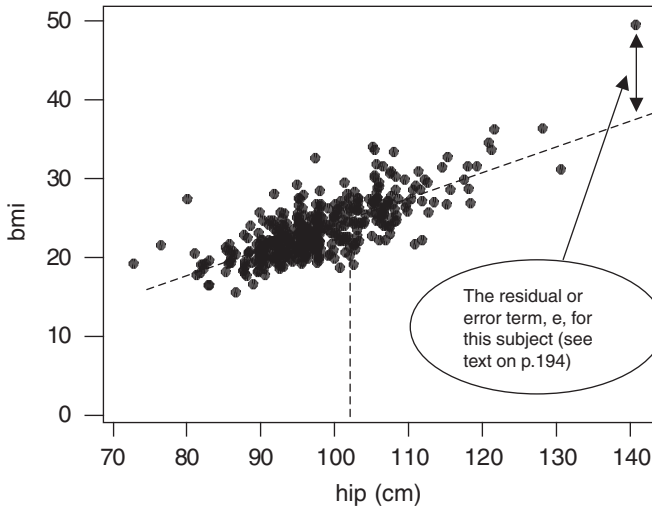


Figure 17.2 A scatterplot of body mass index against hip circumference, for a sample of 412 women in a diet and health cohort study. The scatter of values appears to be distributed around a straight line. That is, the relationship between these two variables appears to be broadly *linear*

Exercise 17.1 Plot the following values for the variables X and Y on a scatter plot and draw the straight line through the points. What is the equation of this line?

Y	5	4	2	1
X	2	4	8	10

The linear regression model

In Figure 17.1 all of the points lie *exactly* on the straight line. In practice this won't happen, and the scatterplot in Figure 17.2 is more typical of what you might see. Here we have body mass index, bmi , (in kg/m^2), and hip circumference, HIP (cm), for a sample of 412 British women from a study into the relationship between diet and health. Suppose we believe that there is a causal relationship between bmi and HIP – changes in hip measurement lead to changes in bmi . If we want to investigate the nature of this relationship then we need to do three things, which I'll deal with in turn:

- Make sure that the relationship is linear.²
- Find a way to determine the equation linking the variables, i.e. get the values of b_0 and b_1 .
- See if the relationship is statistically significant, i.e. that it is present in the population.

² Because we are only dealing with linear relationships in this chapter.

Is the relationship linear?

One way of investigating the linearity of the relationship is to examine the scatterplot, such as that in Figure 17.2.

The points in the scatterplot do seem to cluster along a straight line (shown dotted), which I have drawn, ‘by eye’, through the scatter. This suggests a *linear* relationship between bmi and HIP. So far, so good. We can write the equation of this straight line as:

$$\text{bmi} = b_0 + b_1 \times \text{HIP}$$

This equation is known as *the sample regression equation*. The variable on the left-hand side of the equation, bmi, is known variously as the *outcome*, *response* or *dependent* variable. I’m going to refer to it as the dependent variable in this chapter. It must be *metric continuous*. It gives us the *mean* value of bmi for any specified HIP measurement. In other words, it would tell us (if we knew b_0 and b_1) what the mean body mass index would be for all those women with some particular hip measurement.

The variable on the right-hand side of the equation, HIP, is known variously as the *predictor*, *explanatory* or *independent* variable, or the covariate. I will use the term *independent variable* here. The independent variable can be of any type: nominal, ordinal or metric. This is the variable that’s doing the ‘causing’. It is changes in hip circumference that cause body mass index to change in response, but not the other way round.

Incidentally, my ‘by eye’ line has the equation:

$$\text{bmi} = -8.4 + 0.33 \times \text{HIP}$$

This means that the *mean* body mass index of *all* the women with, say, HIP = 100 cm in this sample is equal to 24.6 kg/m².

Clearly drawing a line by eye through a scatter is not satisfactory – ten people would get ten different lines. So the obvious question arises, ‘What is the ‘best’ straight line that can be “drawn” through a scatter of sample values, and how do I find out what it is?’

Exercise 17.2 (a) Draw by eye the best straight line you can through the scatterplot in Figure 15.1, and write down the regression equation. By how much would the mean incidence rate of ulcerative colitis (UC) change if the rate of Crohn’s disease (CD) changed by one unit? (b) Draw, by eye, the best straight line you can through the scatterplot in Figure 15.2, and write down the regression equation. What change in mean percentage mortality would you expect if the mean number of episodes per year increased by 1? (c) What is the equation of the regression line shown in Figure 15.5? What value of mean per cent exposed at work would you expect if per cent of current smokers in a workplace was 35 per cent?

Estimating b_0 and b_1 – the method of ordinary least squares (OLS)

The second problem is to find a method of getting the values of the sample coefficients b_0 and b_1 , which will give us a line that fits the scatter of points better than any other line, and which

will then enable us to write down the equation linking the variables. The most popular method used for this calculation is called *ordinary least squares*, or OLS. This gives us the values of b_0 and b_1 , and the straight line that *best* fits the sample data. Roughly speaking, ‘best’ means the line that is, on average, closer to all of the points than any other line. How does it do this? Look back at Figure 17.2. The distance of each point in the scatter from the regression line is known as the *residual* or error, denoted e . I have shown the e for just one of the points. If all of these residuals are squared and then added together, to give the term Σe^2 ,³ then the ‘best’ straight line is the one for which the sum, Σe^2 , is smallest. Hence the name ordinary ‘least squares’.

Now: the calculations involved with OLS are too tedious to do by hand, but you can use a suitable computer program to derive their values quite easily (both SPSS and Minitab will do this). It is important to note that the sample regression coefficients b_0 and b_1 are *estimates* of the population regression coefficients β_0 and β_1 . In other words, we are using the sample regression equation:

$$Y = b_0 + b_1 X$$

to estimate the *population regression equation*:

$$Y = \beta_0 + \beta_1 X$$

Basic assumptions of the ordinary least squares procedure

The ordinary least squares procedure is only guaranteed to produce the line that best fits the data if the following assumptions are satisfied:

- The relationship between Y and X is linear.
- The dependent variable Y is metric continuous.
- The residual term, e , is Normally distributed, with a mean of zero, for each value of the independent variable, X .
- The spread of the residual terms should be the same, whatever the value of X . In other words, e shouldn’t spread out more (or less) when X increases.

Let me explain the last two assumptions. Suppose you had, say, 50 women with a hip circumference of 100 cm. As the scatterplot in Figure 17.2 indicates, most of these women have a different body mass index. As you have seen, the difference between each individual woman’s bmi and the regression line is the residual e . If you arranged these 50 residual values into a frequency distribution then the third assumption stipulates that this distribution should be Normal.

The fourth assumption demands that if you repeated the above exercise for each separate value of hip circumference, then the spreads (the standard deviations) of each distribution of

³ Known as the sum of squares. Σ is the Greek ‘sigma’, which means sum all the values.

residual values should be the same, for all hip sizes. If the residual terms have this latter property then they are said to be *homoskedastic*.

These assumptions may seem complicated, but the consequences for the accuracy of the ordinary least squares estimators may be serious if they are violated. Needless to say, these assumptions need to be checked. I'll return to this later.

Back to the example – is the relationship statistically significant?

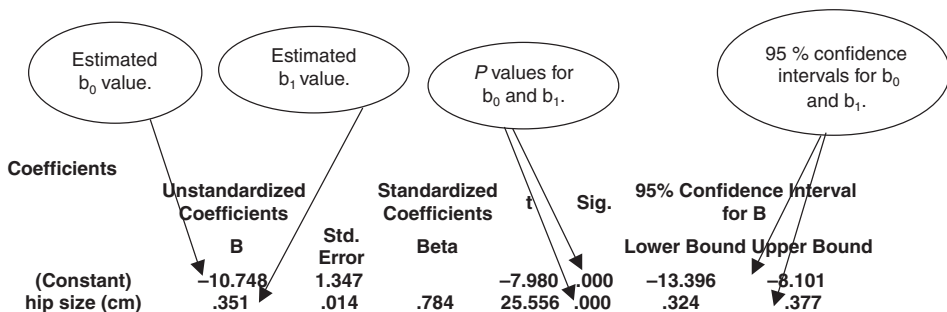
Having calculated b_1 and b_2 , we now need to address the third question; is the relationship statistically significant in the population? We can use either confidence intervals for β_0 and β_1 or hypothesis tests, to judge statistical significance. We then ask: ‘Does the confidence interval for β_1 include zero (or is its p value > 0.05)?’ If the answer in either case is yes, then you *can't* reject the null hypothesis that β_1 is equal to zero; which means that the relationship is not statistically significant. Whatever the value of HIP, once multiplied by a b_1 equal to zero, it disappears from the regression equation and can have no effect on bmi.

SPSS and Minitab for example, will give you confidence intervals and/or p values. In practice we have very little interest in the constant coefficient β_0 ; it's only there to keep a mathematical equality between the left- and right-hand sides of the equation. Besides, in reality it often has no sensible interpretation. For example, in the current example, β_0 would equal the body mass index of individuals with a hip circumference equal to zero!

Thus the focus in linear regression analysis is to use b_1 to estimate β_1 , and then examine its statistical significance. If β_1 is statistically significant, then the relationship is established (well at least with a confidence level of 95 per cent).

Using SPSS

If you use the SPSS linear regression program with the data on the 412 women in Figure 17.2, you will get the output shown in Figure 17.3. SPSS provides both a p value and a 95 per cent confidence interval.



a Dependent Variable: Body Mass Index (weight/height^2) (kg/m2)
 $R^2 = 0.614$; $R^{-2} = 0.613^1$.

¹Values for R^2 and R^{-2} appear in a separate table in the SPSS output. For convenience I have copied them to this table. See below for comment on R^2 .

Figure 17.3 Output from SPSS for ordinary least squares regression applied to the body mass index/hip circumference example

Using Minitab

With Minitab you get the output shown in Figure 17.4. Minitab calculates only the p value, otherwise the results are the same as for SPSS.

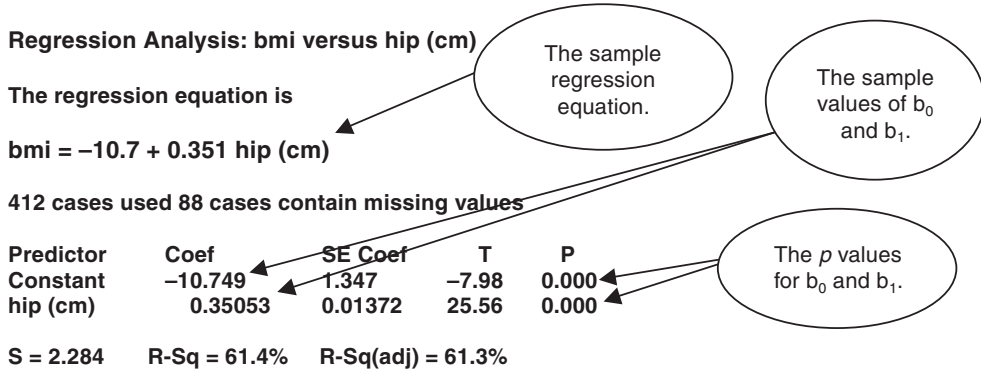


Figure 17.4 Output from Minitab for ordinary least squares regression applied to the body mass index/hip circumference example

Between them, Figure 17.3 and Figure 17.4 provide us with the estimates of b_0 and b_1 , their 95 per cent confidence intervals and their p values, along with the value of R^2 (see below). Regression results are often summarised in a table such as that in Table 17.1.

Table 17.1 Summary of results from the regression of BMI on HIP

Dependent variable	Coefficient	Estimated value	(95 % CI)	(p -value)	R^2	\bar{R}^2
BMI	b_0	-10.748	(-13.396 to -8.101)	0.000	61.4%	61.3%
	b_1	0.351	(0.324 to 0.377)	0.000		

The 95 per cent confidence interval and the p value is shown alongside each sample coefficient. Both parameters β_0 and β_1 are statistically significant, since neither confidence interval includes zero, and both p values are less than 0.05. Thus the result of this analysis is that bmi and HIP are statistically significantly related through the linear regression equation:

$$\text{bmi} = -10.7 + 0.351 \times \text{HIP}^4$$

The value of +0.351 for b_1 means that for every unit (1 cm) increase in hip circumference, the mean bmi will increase by 0.351 kg/m². Knowing the equation, you can, if you wish, draw this best OLS estimated regression line onto the scatterplot.

The regression equation also enables us to *predict* the value of the mean bmi, for any value of hip circumference, *within* the range of the sample hip circumference values (71 cm to 140 cm). For example, for individuals with a hip circumference of 100 cm, you can substitute HIP = 100

⁴ Compare with the by-eye line of: $\text{bmi} = -8.4 + 0.33\text{HIP}$.

into the sample regression equation and thus calculate a value for mean bmi of 24.4 kg/m². Prediction of bmi for hip circumference values *outside* the original sample data range requires a more complex procedure, and will not be discussed here.

Exercise 17.3 What does the model predict for mean bmi for women with a hip circumference of 130 cm?

Goodness-of-fit – R^2

Figure 17.3, Figure 17.4, and Table 17.1, contain values for something called R^2 , and \bar{R}^2 (SPSS calls the latter ‘R-Sq(adj)’). What are these? Suppose you think that waist circumference, WST, might be used as a measure of obesity, so you repeat the above procedure, but use WST as your independent variable instead of HIP. Your results indicate that b_1 is again statistically significant. Now you have two models, in both of which the independent variable has a statistically significant linear relationship with bmi. But which model is best? The one with HIP or the one with WST?

In fact, the best model is the one that ‘explains’ the greatest proportion of the observed variation in bmi from subject to subject, that is, has the best *goodness-of-fit*. One such measure of this explanatory power is known as the *coefficient of determination*, and is denoted R^2 .

As a matter of interest, $R^2 = 0.614$, or 61.4 per cent, for the hip circumference model, and $R^2 = 0.501$, or 50.1 per cent, for the waist circumference model. So variation in hip circumference explains 61 per cent of the observed variation in bmi, while variation in waist circumference explains only 50 per cent of the variation. So using hip circumference as your independent variable gives you a better fitting model.

Here’s a thought. If only 61 per cent of the variation in bmi is explained by variation in hip circumference, what is the remaining 39 per cent explained by? One possibility is that the rest is due to chance, to random effects. A more likely possibility is that, as well as hip circumference, there are other variables that contribute something to the variation in bmi from subject to subject. It would be naïve to believe that variation in bmi, or any clinical variable, can be totally explained by only one variable. Which brings us neatly to the *multiple* linear regression model.

Multiple linear regression

A *simple* linear regression model is one with only one independent variable on the right-hand side. When you have *more* than one independent variable the regression model is called a *multiple linear regression model*. For example, having noticed that both hip and waist circumference are each significantly related to bmi, you might include them *both* as independent variables. This gives the following model, which now gives us *mean* bmi for the various possible combinations of sample values of both HIP and WST:

$$\text{bmi} = b_0 + b_1 \times \text{HIP} + b_2 \times \text{WST}$$

	Variable	Estimated coefficient	(95 % CI)	(<i>p</i> -value)	R^2	\bar{R}^2
Model (& dependent variable)	constant	$b_0 = -10.748$	(-13.396 to -8.101)	0.000		
1. BMI	HIP	$b_1 = 0.351$	(0.324 to 0.377)	0.000	61.4%	61.3%
2. BMI	constant	$b_0 = -9.645$	(-12.250 to -7.041)	0.000		
	HIP	$b_1 = 0.261$	(0.219 to 0.303)	0.000		
	WST	$b_2 = 0.105$	(0.065 to 0.144)	0.000	63.7%	63.5%

Figure 17.5 Multiple linear regression output (last three rows) from SPSS for model with body mass index as the dependent variable and both hip and waist circumferences as independent variables

Note that when we move from the simple to the multiple linear regression model, we need to add a further basic assumption to the list on p. 194. That is, that there is no perfect association or *collinearity* between any of the independent variables. When this assumption is not met, we refer to the model as having *multicollinearity*. The consequence of this condition is wide and thus imprecise confidence intervals.

If you use SPSS to derive the OLS estimators of the above model containing both HIP and WST you get the output shown in Figure 17.5 (last three rows).

Using these results, we can write the estimated multiple linear regression model as:

$$\text{bmi} = -9.645 + 0.261 \times \text{HIP} + 0.105 \times \text{WST}$$

So for example, for all of those women in the sample for whom $\text{HIP} = 100$ and $\text{WST} = 75$, then the above equation estimates their *mean* bmi to be:

$$\text{bmi} = -9.645 + 0.261 \times 100 + 0.105 \times 75 = 24.330$$

The other information in Figure 17.5 tells us that parameters β_1 and β_2 are both statistically significant as neither confidence interval includes zero. Compared to the simple regression model containing only HIP as an independent variable, goodness of fit has improved marginally, with R^2 increasing from 61.4 per cent to 63.7 per cent. Note that in the multiple linear regression model, R^2 measures the explanatory power with *all* of the variables currently in the model acting together.

Exercise 17.4 If we add ‘age’ as a third independent variable to the bmi model, then Minitab produces the results shown in Figure 17.6. (a) Comment on the statistical significance of the three independent variables. (b) How does an increase in age effect mean body mass index values? (c) Has goodness of fit improved compared to the model with only HIP and WST included? (d) What is the mean body mass index of all of those women in the sample with a hip circumference of 100 cm, and a waist circumference of 75 cm, who are aged: (i) 30; (ii) 60?

Regression Analysis : BMI versus hip(cm), waist(cm), Age

The regression equation is

$$\text{BMI} = -12.4 + 0.289 \text{ hip(cm)} + 0.125 \text{ waist(cm)} - 0.0249 \text{ Age}$$

Predictor	Coef	SE Coef	T	P
Constant	-12.425	1.353	-9.18	0.000
Hip (cm)	0.28876	0.02041	14.15	0.000
Waist (cm)	0.12549	0.01762	7.12	0.000
Age	-0.02492	0.01104	-2.26	0.024

S = 2.24817 R-Sq = 64.0% R-Sq(adj) = 63.8%

Figure 17.6 Output from Minitab for regression of bmi on HIP, WST and AGE

Dealing with nominal independent variables: design variables and coding

In linear regression, most of the independent variables are likely to be metric, or at least ordinal. However any independent variable that is *nominal* must be coded into a so-called *design* (or *dummy*) variable, before being entered into a model. There is only space for a brief description of the process here.

As an example, suppose in a study of hypertension, you have systolic blood pressure (SBP) as your dependent variable, and age (AGE) and smoking status (SMK), as your independent variables. SMK, is a nominal variable, having the categories: non-smoker, ex-smoker, and current smoker. This gives the model:

$$\text{SBP} = b_0 + b_1\text{AGE} + b_2\text{SMK} \quad (1)$$

To enter SMK into your computer, you would have to score the three smoking categories in some way – but how? As 1, 2, 3, or as 0, 1, 2, etc. As you can imagine, the scores you attribute to each category will effect your results. The answer is to *code* these three categories into *two* design variables. Note that the number of design variables is always one *less* than the number of categories in the variable being coded. In this example, we set out the coding design as in Table 17.2.

So you replace smoking status (with its dodgy numbering), with two new design variables, D_1 and D_2 , which take the values in Table 17.2, according to smoking status. The model now

Table 17.2 Coding design for a nominal variable with three categories

Smoking status	Design variable values	
	D_1	D_2
Non-smoker	0	0
Ex-smoker	0	1
Current smoker	1	0

becomes: $Y = b_0 + b_1\text{Age} + b_2D_1 + b_3D_2$. For example, if the subject is a current smoker, $D_1 = 1$ and $D_2 = 0$; if an ex-smoker, $D_1 = 0$ and $D_2 = 1$; if a non-smoker, $D_1 = 0$ and $D_2 = 0$. Notice in the last situation that the smoking status variable effectively disappears from the model.

This coding scheme can be extended to deal with nominal variables with any reasonable number of categories, depending on the sample size.⁵ The simplest situation is a nominal variable with only *two* categories, such as sex, which can be represented by one design variable with values 0 (if male) or 1 (if female).

Exercise 17.5 The first three subjects in the study of systolic blood pressure and its relationship with age and smoking status are, a 50-year-old smoker, a 55-year-old non-smoker and a 35-year-old ex-smoker, respectively. Fill in the first three rows of the data sheet shown in Table 17.3, as appropriate.

Table 17.3 Data sheet for systolic blood pressure relationship

Subject	Age	D_1	D_2
1			
2			
3			

Model building and variable selection

At the beginning of this chapter we chose body mass index as the variable to explain or model systolic blood pressure. In practice, researchers may or may not have an idea about which variables they think are relevant in explaining the variation in their dependent variable. Whether they do or they don't will influence their decision as to which variables to include in their model, i.e. their *variable selection procedure*.

There are two main approaches to the model-building process:

- First, *automated* variable selection – the computer does it for you. This approach is perhaps more appropriate if you have little idea about which variables are likely to be relevant in the relationship.
- Second, *manual* selection – *you* do it! This approach is more appropriate if you have a particular hypothesis to test, in which case you will have a pretty good idea which independent

⁵ As a rule of thumb, you need at *the very least* 15 subjects for each independent variable in your model. If you've got, say, five ordinal and/or metric independent variables in your model, you would need a minimum of 75 subjects. If you want also to include a single nominal variable with five categories (i.e. four design variables), you would need another 60 subjects. In these circumstances, it might help to amalgamate some categories.

variable is likely to be the most relevant in explaining your dependent variable. However, you will almost certainly want to include other variables to control for confounding (see p. 81 for an account of confounding).

Both of these methods have a common starting procedure, as follows:⁶

- Identify a list of independent variables that you think might possibly have some role in explaining the variation in your dependent variable. Be as broad-minded as possible here.
- Draw a scatterplot of each of these candidate variables (if it is not a nominal variable), against the dependent variable. Examine for linearity. If any of the scatterplots show a strong, but not a linear relationship with the dependent variable, you will need to code them first before entering them into the computer data sheet. For example, you might find that the relationship between the dependent variable and 'age' is strong but not linear. One approach is to group the *age* values into four groups, using its three quartile values to define the group boundaries, and then code the groups with three design variables.
- Perform a series of univariate regressions, i.e. regress each candidate independent variable in turn against the dependent variable. Note the *p*-value in each case.
- At this stage, all variables that have a *p*-value of at least 0.2 should be considered for inclusion in the model. Using a *p*-value less than this may fail to identify variables that could subsequently turn out to be important in the final model.

With this common starting procedure out of the way, we can briefly describe the two variable selection approaches, starting with automated methods.

Automated variable selection methods

- *Forwards selection*: The program starts with the variable that has the lowest *p*-value from the univariate regressions. It then adds the other variables one at a time, in lowest *p*-value order, regressing each time, retaining all variables with *p*-values < 0.05 in the model.
- *Backwards selection*: The reverse of forwards selection. The program starts with *all* of the candidate variables in the model, then the variable that has highest *p*-value > 0.05 , is removed. Then the next highest *p*-value variable, and so on, until only those variables with a *p*-value < 0.05 are left in the model, and all other variables have been discarded.
- *Forwards or backwards stepwise selection*: After each variable is added (or removed), the variables which were already (or are left) in the model are re-checked for statistical significance; if no longer significant they are removed. The end result is a model where all variables have a *p*-value < 0.05 .

⁶Note that the criteria used by the different computer regression programs to select and de-select variables differ.

These automated procedures have a number of disadvantages, although they may be useful when researchers have little idea about which variables are likely to be relevant. As an example of the automated approach, the authors of a study into the role of arginase in sickle cell disease, in which the outcome variable was \log_{10} arginase activity (Morris *et al.* 2005), comment:

This modelling used a stepwise procedure to add independent variables, beginning with the variables most strongly associated with \log_{10} arginase with $P \leq 0.15$. Deletion of variables after initial inclusion in the model was allowed. The procedure continued until all independent variables in the final model had $P \leq 0.05$, adjusted for other independent variables, and no additional variables had $P \leq 0.05$.

Manual variable selection methods

Manual, DIY methods, are often more appropriate if the investigators know in advance which is likely to be their principal independent variable. They will include this variable in the model, together with any other variables that they think may be potential confounders. The identity of potential confounders will have been established by experience, a literature search, discussions with colleagues and patients and so on. There are two alternative manual selection procedures:

- *Backward elimination:* The main variable plus all of the potentially confounding variables are entered into the model at the start. The results will then reveal which variables are statistically significant (p -value < 0.05). Non-significant variables can then be dropped, one at a time in decreasing p -value order, from the model, regressing each time. However, if the coefficient of any of the remaining variables changes markedly⁷ when a variable is dropped, the variable should be retained, since this may indicate that it is a confounder.
- *Forward elimination:* The main explanatory variable of interest is put in the model, and the other (confounding) variables are added one at a time in order of (lowest) p -value (from the univariate regressions). The regression repeated each time a variable is added. If the added variable is statistically significant it is retained, if not it is dropped, unless any of the coefficients of the existing variables change noticeably, suggesting that the new variable may be a confounder.

The end result of either of these manual approaches should be a model containing the same variables (although this model may differ from a model derived using one of the automated procedures). In any case, the overall objective is *parsimony*, i.e. having as few explanatory variables in the model as possible, while at the same time explaining the maximum amount of variation in the dependent variable. Parsimony is particularly important when sample size is on the small side. As a rule of thumb, researchers will need at least 15 observations for each independent variable to ensure mathematical stability, and at least 20 observations to obtain reasonable statistical reliability (e.g. narrow-ish confidence intervals).

As an example of the manual backwards selection approach, the authors of a study of birthweight and cord serum EPA concentration (Grandjean *et al.* 2000), knew that cord serum

⁷ There is no rule about how big a change in a coefficient should be considered noteworthy. A value of 10 per cent has been suggested, but this seems on the small side.

EPA was their principal independent variable, and but wanted to include possible confounders in their model. They commented:

Multiple regression analysis was used to determine the relevant importance of predictors of the outcome (variable). Potential confounders were identified on the basis of previous studies, and included maternal height and weight, smoking during pregnancy, diabetes, parity, gestational length, and sex of the child. Covariates⁸ were kept in the final regression equation if statistically significant ($p < 0.01$) after backwards elimination.

Incidentally, the main independent variable, cord serum concentration, was found to be statistically significant (p -value = 0.037), as were all of the confounding variables.

Goodness-of-fit again: \bar{R}^2

When you add an *extra* variable to an existing model, and want to compare goodness-of-fit with the old model, you need to compare not R^2 , but *adjusted* R^2 , denoted \bar{R}^2 . The reasons don't need to concern us here, but R^2 will increase when an extra independent variable is added to the model, without there necessarily being any increase in explanatory power. However, if \bar{R}^2 increases, then you know that the explanatory power (its ability to explain more of the variation in the dependent variable) of the model *has* increased. From Figure 17.3 or Figure 17.4, $\bar{R}^2 = 0.613$ in the *simple* regression model with only hip circumference as an independent variable. From Figure 17.5, with both hip and waist circumferences included, \bar{R}^2 increases to 0.635, so this multiple regression model does show a small but real improvement in goodness-of-fit, and would be preferred to either of the simple regression models. Of course, you might decide to explore the possibility that other independent variables might also have a significant role to play in explaining variation in body mass index; age is one obvious contender, as is sex, and should be included in the model.

Exercise 17.6 Table 17.4 contains the results of a multiple linear regression model from a cross-section study of disability, among 1971 adults aged 65 and over in 1986 (Kavanagh and Knapp 1998). The objective of the study was to examine the utilisation rates of general practitioners' time by elderly people resident in communal establishments. The dependent variable was the *natural log* of weekly utilisation (minutes) per resident.⁹ There were 10 independent variables, as shown in the figure.

(a) Identify those independent variables whose relationship with the dependent variable is statistically significant. (b) What is the effect on the natural log of utilisation time, and what is this in actual minutes, if there is an increase of: (i) one person in the number in a

⁸ i.e. independent variables.

⁹ Probably because the researchers believed the utilisation rate to be skewed. See Figure 5.6 for an example of transformed data.

private residential home; (ii) one unit in the severity of disability score? (c) How much of the variation in general practitioners' utilisation time is explained by the variation in the independent variables?

Table 17.4 Sample regression coefficients from a linear regression model, where the dependent variable is the natural log of the utilisation time (minutes) of GPs, by elderly patients in residential care. The independent variables are as shown. Reproduced from *BMJ*, **317**, 322–7, courtesy of BMJ Publishing Group

Explanatory variable	β coefficient (SE)	P value
Constant	0.073 (0.353)	0.837
Age	<0.0005 (0.004)	0.923
Male sex	0.024 (0.060)	0.685
Severity of disability	0.043 (0.005)	<0.0001
Mental disorders	0.120 (0.061)	0.047
Nervous system disorders	0.116 (0.062)	0.063
Circulatory system disorders	0.122 (0.066)	0.063
Respiratory system disorders	0.336 (0.115)	0.003
Digestive system disorders	0.057 (0.070)	0.415
Type of accommodation:		
Local authority	—	—
Voluntary residential home	−0.084 (0.183)	0.649
Voluntary nursing home	0.562 (0.320)	0.079
Private residential home	−0.173 (0.157)	0.272
Private nursing home	0.443 (0.228)	0.053
Size of establishment (No of residents)		
Local authority	−0.004 (0.003)	0.170
Voluntary residential home	−0.004 (0.002)	0.069
Voluntary nursing home	−0.002 (0.002)	0.245
Private residential home	0.006 (0.002)	0.017
Private nursing home	−0.007 (0.007)	0.362

$R^2 = 0.1098$, $F_{(17,415)} = 9.71$, $P = <0.0001$. Sample size = 1971 in 433 sampling units.

Adjustment and confounding

One of the most attractive features of the multiple regression model is its ability to *adjust* for the effects of possible association between the independent variables. It's quite possible that two or more of the independent variables will be associated. For example, hip (HIP) and waist (WST) circumference are significantly positively associated with $r = +0.783$ and p -value = 0.000. The consequence of such association is that increases in HIP are likely to be accompanied by increases in WST. The increase in HIP will cause bmi to increase both directly, but also indirectly via WST. In these circumstances it's difficult to tell how much of the increase in bmi is due *directly* to an increase in HIP, and how much to the *indirect* effect of an associated increase in WST.

The beauty of the multiple regression model is that each regression coefficient measures only the *direct* effect of its independent variable on the dependent variable, and controls or adjusts for any possible interaction from any of the other variables in the model. In terms of the results in Figure 17.5, an increase in HIP of 1 cm will cause mean bmi to increase by 0.261 kg/m² (the value

of b_1), and *all* of this increase is caused by the change in hip circumference (plus the inevitable random error). Any effect that a concomitant change in waist circumference might have is discounted. The same applies to the value of -0.0249 for b_3 on the 'age' variable in Figure 17.6.

We can use the adjustment property to deal with confounders in just the same way. You will recall that a confounding variable has to be associated with *both* one of the independent variables *and* the dependent variable (see the discussion in Chapter 6). Notice that the coefficient b_1 , which was 0.351 in the simple regression model with HIP the only independent variable, decreases to 0.261 with two independent variables. A marked change like this in the coefficient of a variable already in the model when a new variable is added, is an indication that one of the variables is possibly a confounder. As you have already seen in the model-building section above, in these circumstances both variables should be retained in the model.

An example from practice

Table 17.5 is from a cross-section study into the relationship between bone lead and blood lead levels, and the development of hypertension in 512 individuals selected from a cohort study (Cheng *et al.* 2001). The table shows the outcome from three multiple linear regression models with systolic blood pressure as the dependent variable. The first model includes blood lead as an independent variable, along with six possible *confounding* variables.¹⁰ The second and third models were the same as the first model, except tibia and patella lead, respectively, were substituted for blood lead. The results include 95 per cent confidence intervals and the R^2 for each model.

As the table shows, the tibia lead model has the best goodness-of-fit ($R^2 = 0.1015$), but even this model only explains 10 per cent of the observed variation in systolic blood pressure. However, this is the only model that supports the relationship between hypertension and lead levels; the 95 per cent confidence interval for tibia lead (0.02 to 2.73) does not include zero. The only confounders statistically significant in all three models are age, family history of hypertension and calcium intake.

Exercise 17.7 From the results in Table 17.5: (a) which independent variables are statistically significant in all three models? (b) Explain the 95 per cent confidence interval of (0.28 to 0.64) for *age* in the blood lead model. (c) In which model does a unit (1 year) increase in age change systolic blood pressure the most?

Diagnostics – checking the basic assumptions of the multiple linear regression model

The ordinary least squares method of coefficient estimation will only produce the best estimators if the basic assumptions of the model are satisfied. That is: a metric continuous

¹⁰ The inclusion of Age^2 in the model is probably an attempt to establish the linearity of the relationship between systolic blood pressure and age. If the coefficient for Age^2 is not statistically significant then the relationship is probably linear.

Table 17.5 Multiple regression results from a cross-section study into the relationship between bone lead and blood lead levels and the development of hypertension in 512 individuals selected from a cohort study. The figure show the outcome from three multiple linear regression models, with systolic blood pressure as the dependent variable. Reproduced from *Amer. J. Epidemiol.*, 153, 164–71, courtesy of Oxford University Press

Variable	Baseline model + blood lead		Baseline model + tibia lead		Baseline model + patella lead	
	Parameter estimate	95% CI	Parameter estimate	95% CI	Parameter estimate	95% CI
Intercept	128.34		125.90		127.23	
Age (years)	0.46*	0.28, 0.64	0.39*	0.20, 0.58	0.44*	0.26, 0.63
Age squared (years ²)	-0.02*	-0.04, -0.00	-0.02*	-0.04, -0.00	-0.02*	-0.04, -0.00
Body mass index (kg/m ²)	0.36*	0.01, 0.72	0.33	-0.02, 0.69	0.35	-0.00, 0.71
Family history of hypertension (yes/no)	4.36*	1.42, 7.30	4.36*	1.47, 7.25	4.32*	1.42, 7.22
Alcohol intake (g/day)	0.08*	0.00, 0.149	0.07	-0.00, 0.14	0.07	-0.00, 0.14
Calcium intake (10 mg/day)	-0.04*	-0.08, -0.00	-0.04*	0.07, -0.00	-0.04*	-0.07, -0.00
Blood lead (SD) [†]	-0.13	-1.35, 1.09				
Tibia lead (SD) [†]			1.37*	0.02, 2.73		
Patella lead (SD) [†]					0.57	-0.71, 1.84
Model R ²	0.0956		0.1015		0.0950	

* $p < 0.05$

[†] Parameter estimates are based on 1 standard deviation (SD) in blood lead level (4.03 µg/dl), tibia lead level (13.65 µg/g), and patella lead level (19.55 µg/g).

dependent variable; a linear relationship between the dependent and each independent variable; error terms with constant spread and Normally distributed; and the independent variables not perfectly correlated with each other. How can we check that these assumptions are satisfied?

- *A metric continuous dependent variable.* Refer to Chapter 1 if you are unsure how to identify a metric continuous variable.
- *A linear relationship between the dependent variable and each independent variable.* Easiest to investigate by plotting the dependent variable against each of the independent variables; the scatter should lie approximately around a straight line.¹¹ The other possibility is to plot the residual values against the *fitted* values of the independent variable (bmi in our example). These are the values the estimated regression equation would give for mean bmi, for every combination of values of HIP and WST. The scatter should be evenly spread around zero, with no discernible pattern, such as in Figure 17.7(a).
- *The residuals have constant spread across the range of values of the independent variable.* Check with a plot of the residual values against the *fitted* values of bmi. The spread of the residuals should be fairly constant around the zero value, across the range of fitted values of the independent variable. Figure 17.7(b) is an example of non-constant variance. The spread appears to increase as the value of the independent variable increases. Figure 17.7(c) is an example of both non-linearity and non-constant variance.
- *The residuals are Normally distributed for each fitted value of the independent variable.* This assumption can be checked with a histogram of the residuals. For our bmi example, the histogram in Figure 17.10 indicates that, apart from a rather worrying outlier, the distribution is Normal. You might want to identify which woman this outlier represents and check her data for anomalies.
- *The independent variables are not perfectly correlated with each other.* Unfortunately, this is not an easy assumption to check. Some degree of correlation is almost certain to exist among some of the independent variables.

Exercise 17.8 (a) Explain briefly each of the basic assumptions of the multiple linear regression model. (b) With the aid of sketches where appropriate, explain how we can test that these assumptions are satisfied.

¹¹Notice that we only have to establish this property of linearity for the metric independent variables in the model. Any binary variables are linear by default – they only have two points, which can be joined with a straight line. Any ordinal independent variables will have to be expressed as binary dummies – again linear by default for the same reason.

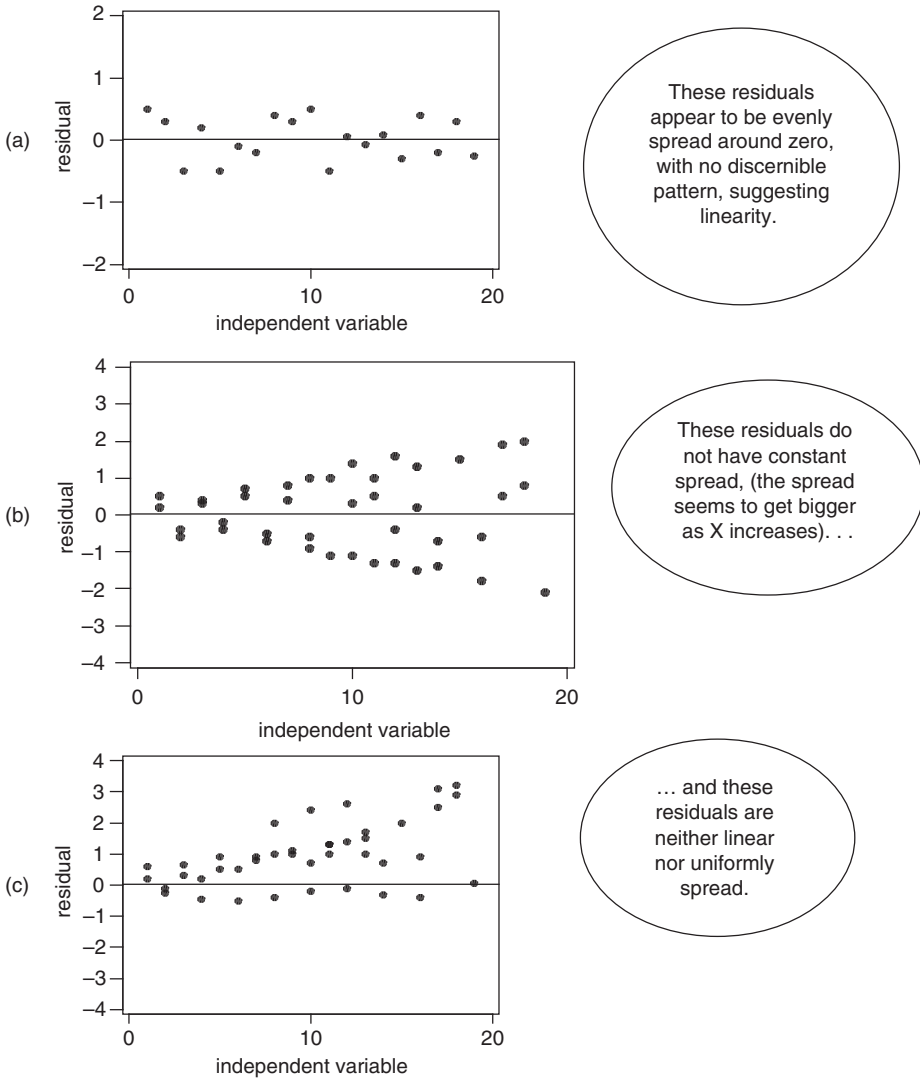


Figure 17.7 Testing the basic assumptions of the linear regression model by plotting the residuals against the fitted values of the regression equation

An example from practice

Let's apply the above ideas to check the basic assumptions of the ordinary least squares method as applied to the multiple linear regression of body mass index (bmi) on hip circumference (HIP) and waist circumference (WST), which we considered earlier. Recall that the model was:

$$\text{bmi} = b_0 + b_1 \times \text{HIP} + b_2 \times \text{WST}$$

and that both HIP and WST were found to be statistically significant explainers of the variation in bmi.

The first assumption is that bmi is a metric continuous dependent variable, which it is. The second assumption is that the relationship between bmi and HIP and bmi and WST should be linear. If we draw a scatterplot of bmi against each of these variables, we get the scatterplots shown in Figure 17.8. These indicate a reasonable degree of linearity in each case. Notice though that the spread in bmi appear to get larger as WST increases.

The third assumption is that the residuals have constant spread over the range of fitted values of the model. Figure 17.9 is a plot of these residuals against the fitted values of bmi. This third assumption appears not to be completely satisfied. The spread of residual values appear to increase as the fitted bmi value increases. This may be an indication that an important independent variable is missing from the model. However, the distribution of points above and below the zero line seems reasonably symmetric, supporting the linearity assumption demonstrated in the scatterplots.

The fourth assumption of the Normality of the residuals is checked with the histogram of the residuals, see Figure 17.10. These do appear to be reasonably Normal, although there is some suggestion of positive skew.

Thus all of the basic assumptions appear to be reasonably well satisfied (apart from the multicollinearity assumption which we have not tested), and the ordinary least squares regression estimates b_1 and b_2 of the population parameters β_1 and β_2 , are the 'best' we can get, i.e. they fit the data at least as well as any other estimates.¹²

Multiple linear regression is popular in clinical research. Much more popular though, for reasons which will become clear in the next chapter, is logistic regression.

Analysis of variance

Analysis of variance (ANOVA) is a procedure that aims to deal with the same problems as linear regression analysis, and many medical statistics books contain at least one chapter describing ANOVA. It has a history in the social sciences, particularly psychology. However, regression and ANOVA are simply two sides of the same coin – the *generalised linear model*. As Andy Field (2000) says:

Anova is fine for simple designs, but becomes impossibly cumbersome in more complex situations. The regression model extends very logically to these more complex designs, without getting bogged down in mathematics. Finally, the method (Anova) becomes extremely unmanageable in some circumstances, such as unequal sample sizes. The regression method makes these situations considerably more simple.

In view of the fact that anything ANOVA can do, regression can also do, and, for me anyway, do it in a way that's conceptually easier, I am not going to discuss ANOVA in this book. If you are interested in exploring ANOVA in more detail, you could do worse than read Andy Field's book, or that of Altman (1991).

¹²There are other methods of estimating the values of the regression parameters, which I don't have the space to consider. However, provided the basic assumptions are satisfied, none will be better than the ordinary least squares estimators.

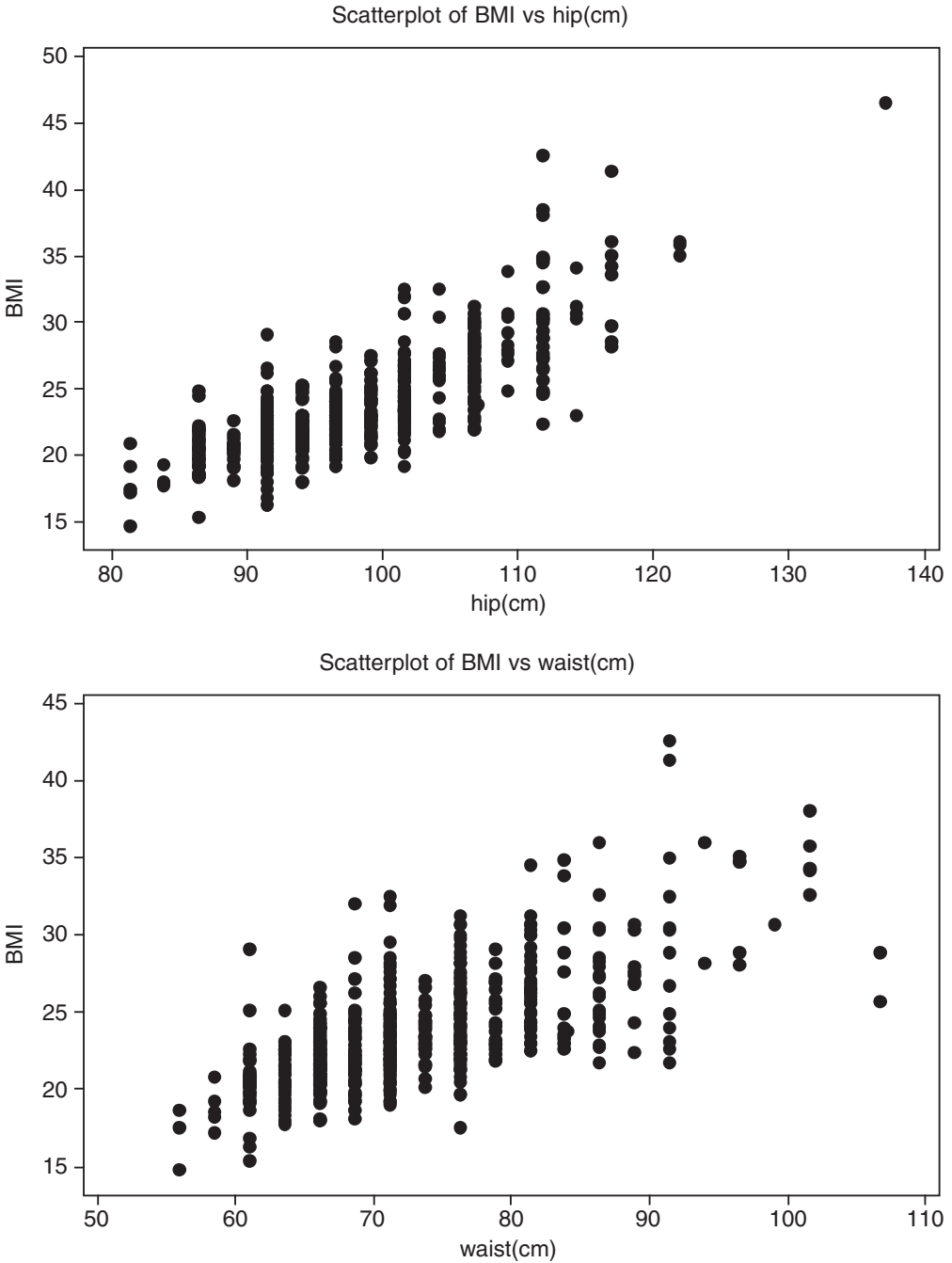


Figure 17.8 Scatterplots of the dependent variable body mass index (bmi) against hip circumference (HIP) – top plot – and waist circumference (WST) – bottom plot. As you can see, both plots indicate a more-or-less linear relationship

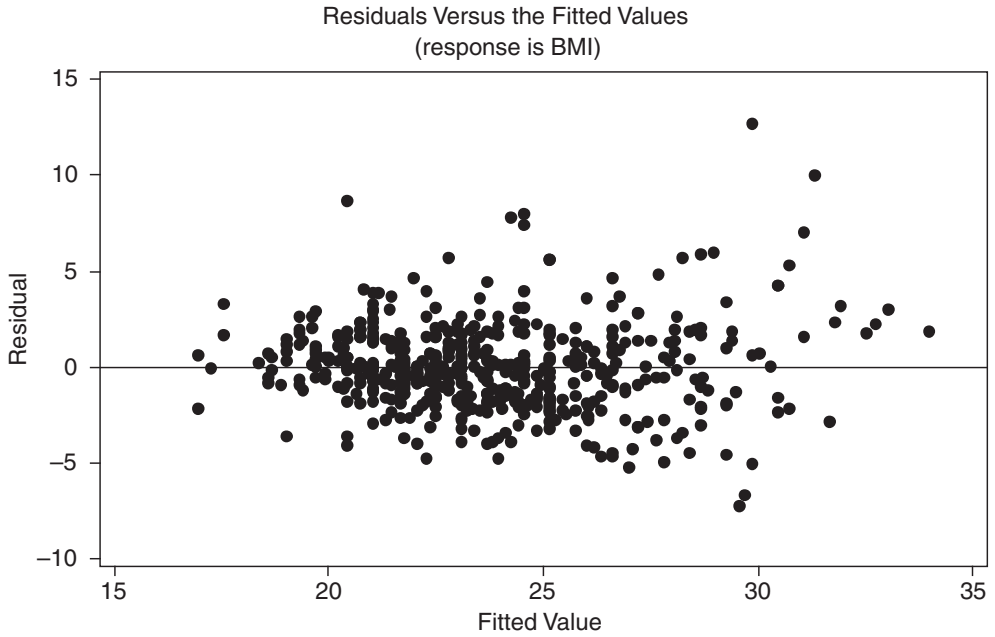


Figure 17.9 A plot of the residuals versus the fitted bmi values, as a check of the basic assumptions of the linear regression model

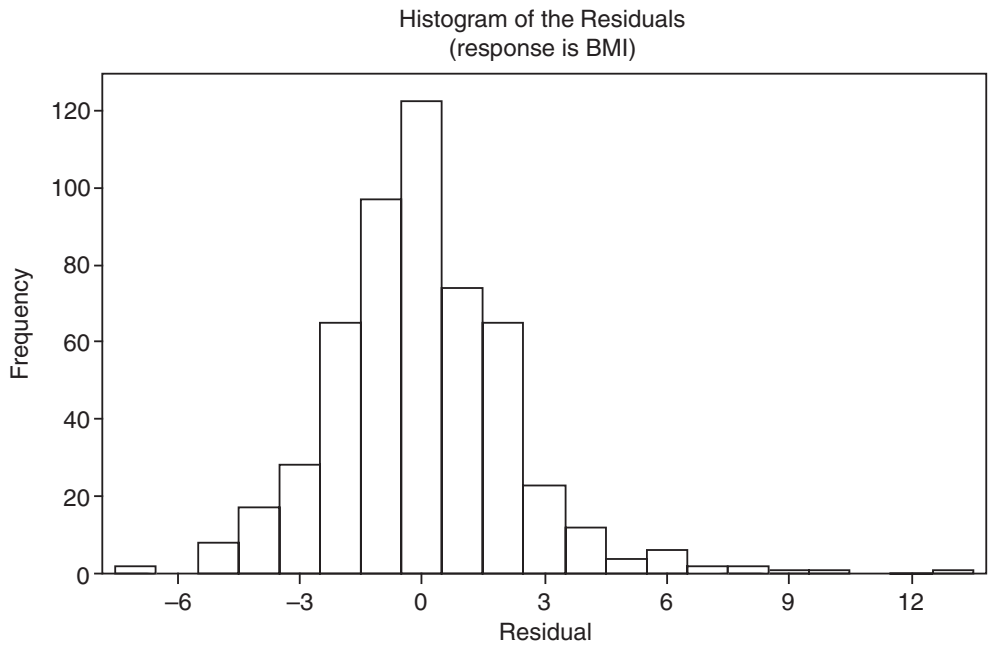


Figure 17.10 A plot of the residuals in the body mass index example, showing reasonable Normality, and thus satisfying the fourth assumption governing the use of the ordinary least squares estimation method

18

Curvy models: logistic regression

Learning objectives

When you have finished this chapter you should be able to:

- Explain why a linear regression model is not appropriate if the dependent variable is binary.
- Explain what the logit transformation is and what it achieves.
- Write down the logic regression equation.
- Explain the concept of linearity and outline how this can be tested for and dealt with.
- Explain how estimates of the odds ratios can be derived directly from the regression parameters.
- Describe how the statistical significance of the population odds ratio is determined.
- Interpret output from SPSS and Minitab logistic regression programs.

A second health warning!

Although the maths underlying the logistic regression model is perhaps more complicated than that in linear regression, once more a brief description of the underlying idea is necessary if you are to gain some understanding of the procedure and be able to interpret logistic computer outputs sensibly.

Binary dependent variables

In linear regression the dependent or outcome variable must be metric continuous. In clinical research, however, the outcome variable in a relationship will very often be dichotomous or *binary*, i.e. will take only *two* different values: alive or dead; malignant or benign; male or female and so on. In addition, variables that are not naturally binary can often be made so. For example, birthweight might be coded ‘less than 2500 g’, and ‘2500 g or more’, Apgar scores coded ‘less than 7’, ‘7 or more’, etc. In this chapter I want to show how a binary dependent variable makes the linear regression model inappropriate.

Finding an appropriate model when the outcome variable is binary

If you are trying to find an appropriate model to describe the relationship between two variables Y and X , when Y , the dependent variable, is continuous, you can draw a scatterplot of Y against X (Figure 17.2 is a good example) and if this has a linear shape you can model the relationship with the linear regression model. However, when the outcome variable is binary, this graphical approach is not particularly helpful.

For example, suppose you are interested in using the breast cancer/stress data from the study referred to in Table 1.6, to investigate the relationship between the outcome variable ‘diagnosis’, and the independent variable ‘age’. Diagnosis is, of course, a binary variable with two values: $Y = 1$ (malignant) or $Y = 0$ (benign). If we plot *diagnosis* against *age*, we get the scatterplot shown in Figure 18.1, from which it’s pretty well impossible to draw any definite conclusions about the nature of the relationship.

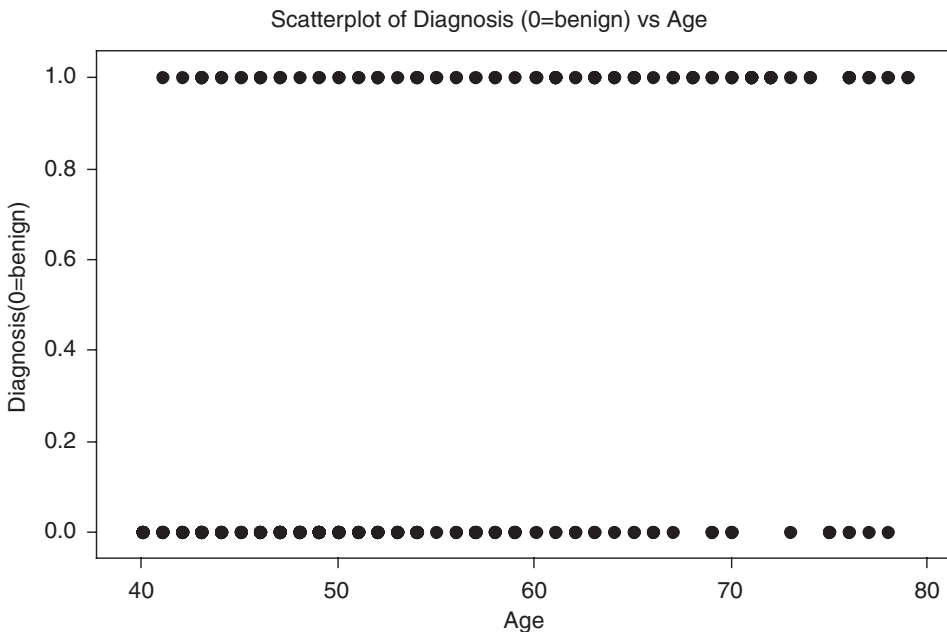


Figure 18.1 Scatter plot of *diagnosis* against *age* for the 332 women in the breast cancer and stress study referred to in Table 1.6

Table 18.1 Proportion of women with malignant lump in each age group

Proportion with malignant lump, i.e. $Y = 1$. Or the probability that $Y = 1$, i.e. $P(Y = 1)$	Midpoint of age group
0.140	45
0.226	55
0.635	65
0.727	75

The problem is that the large variability in age, in both the malignant and benign groups, obscures the difference in age (if any) *between* them. However, if you *group* the age data: 40-49, 50-59, etc., and then calculate the *proportion* of women with a malignant diagnosis (i.e. with $Y = 1$) in each group, this will reduce the variability, but preserve the underlying relationship between the two variables. The results of doing this are shown in Table 18.1.

Notice that I've labelled the first column as the 'Proportion with $Y = 1$, or the Probability that $Y = 1$, written as $P(Y = 1)$ '. Here's why. In linear regression, you will recall that the dependent variable is the *mean* of Y for a given X . But what about a binary dependent variable? Can we find something analogous to the mean? As it happens, the mean of a set of binary, zero or one, values is the same as the *proportion* of ones,¹ so an appropriate equivalent version of the binary dependent variable would seem to be 'Proportion of ($Y = 1$)s'.

But proportions can be interpreted as probabilities (see Chapter 8). So the dependent variable becomes the 'Probability that $Y = 1$ ', or $P(Y = 1)$, for a given value of X . For example the probability of a malignant diagnosis ($Y = 1$) for all of those women aged 40, which we can write as, $P(Y = 1)$ given $X = 40$.

You can see in Table 18.1, the proportion with malignant breast lumps (the probability that $Y = 1$) increases with age, but does it increase linearly? A scatterplot of the proportion with malignant lumps, $Y = 1$, against group age midpoints is shown in Figure 18.2, which does suggest *some* sort of relationship between the two variables. But it's definitely *not* linear, so a *linear* regression model won't work. In fact, the curve has more of an elongated S shape, so what we need is a mathematical equation that will give such an S-shaped curve.

There are several possibilities, but the *logistic* model is the model of choice. Not only because it produces an S-shaped curve, which we want, but, critically, it has a meaningful clinical interpretation. Moreover, the value of $P(Y = 1)$ is restricted by the maths of the logistic model to lie between zero and one, which is what we want, since it's a probability.

The logistic regression model

The simple² *population* logistic regression equation is:

$$P(Y = 1) = (e^{\beta_0 + \beta_1 X}) / (1 + e^{\beta_0 + \beta_1 X}) \quad (1)$$

¹ For example, the mean of the five values: 0, 1, 1, 0, 0 is $2/5 = 0.4$, which is the same as the proportion of 1s, i.e. 2 in 5 or 0.4.

² 'Simple' because there is only one independent variable – so far.

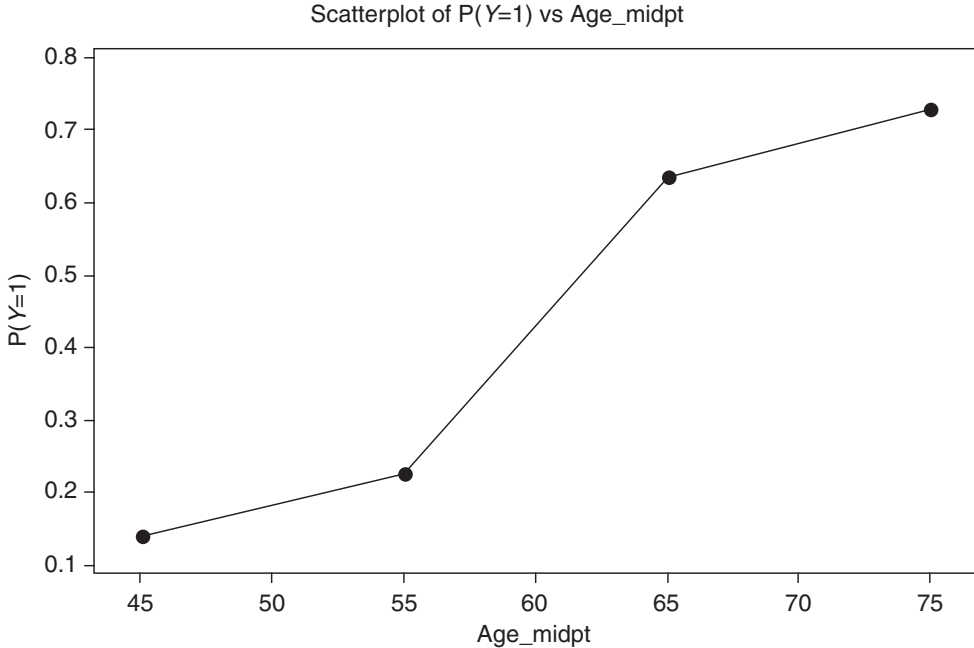


Figure 18.2 Scatterplot of the proportion of women with a malignant diagnosis ($Y = 1$) against midpoints of age group

which we estimate with the *sample* logistic regression equation:

$$P(Y = 1) = (e^{b_0 + b_1 X}) / (1 + e^{b_0 + b_1 X}) \quad (2)$$

by determining the values of the estimators b_0 and b_1 . We'll come back to this problem in a moment. Note that e is the exponential operator, equal to 2.7183, and has nothing to do with the residual term in linear regression. As you can see, the logistic regression model is mathematically a bit more complicated than the linear regression model.

The outcome variable, $P(Y = 1)$, is the probability that $Y = 1$ (the lump is malignant), for some given value of the independent variable X . There is no restriction on the type of *independent* variable, which can be nominal, ordinal or metric.

As an example, let's return to our breast cancer study (Figure 1.6). Our outcome variable is *diagnosis*, where $Y = 1$ (malignant) or $Y = 0$ (benign). We'll start with one independent variable – *ever used an oral contraceptive pill* (OCP), Yes = 1, or No = 0. We are going to treat OCP use as a possible risk factor for receiving a malignant diagnosis. This gives us the sample regression model:

$$P(Y = 1) = (e^{b_0 + b_1 \times \text{OCP}}) / (1 + e^{b_0 + b_1 \times \text{OCP}}) \quad (3)$$

So all we've got to do to determine the probability that a woman picked at random from the sample will get a malignant diagnosis ($Y = 1$), with or without OCP use, is to calculate

the values of b_0 and b_1 somehow, and then put them in the logistic regression equation, with $OCP = 0$, or $OCP = 1$.

Estimating the parameter values

Whereas the linear regression models use the method of ordinary least squares to estimate the regression parameters β_0 and β_1 , logistic regression models use what is called *maximum likelihood estimation*. Essentially this means choosing the population which is *most likely* to have generated the sample results observed. Figure 18.3 and Figure 18.4, respectively, show the output from SPSS's and Minitab's logistic regression program for the above OCP model.

SPSS's and Minitab's logistic regression program both give $b_0 = -0.2877$ and $b_1 = -0.9507$. If we substitute these values into the logistic regression model of Equation (3), we get:³

$$\text{if } OCP = 0 \text{ (has never used OCP), } P(Y = 1) = 0.4286$$

$$\text{if } OCP = 1 \text{ (has used OCP), then } P(Y = 1) = 0.2247$$

So a woman who has never used an oral contraceptive pill has a probability of getting a malignant diagnosis nearly twice that of a woman who *has* used an oral contraceptive. Rather than being a risk factor for a malignant diagnosis, in this sample the use of oral contraceptives seems to confer some protection against a breast lump being malignant.

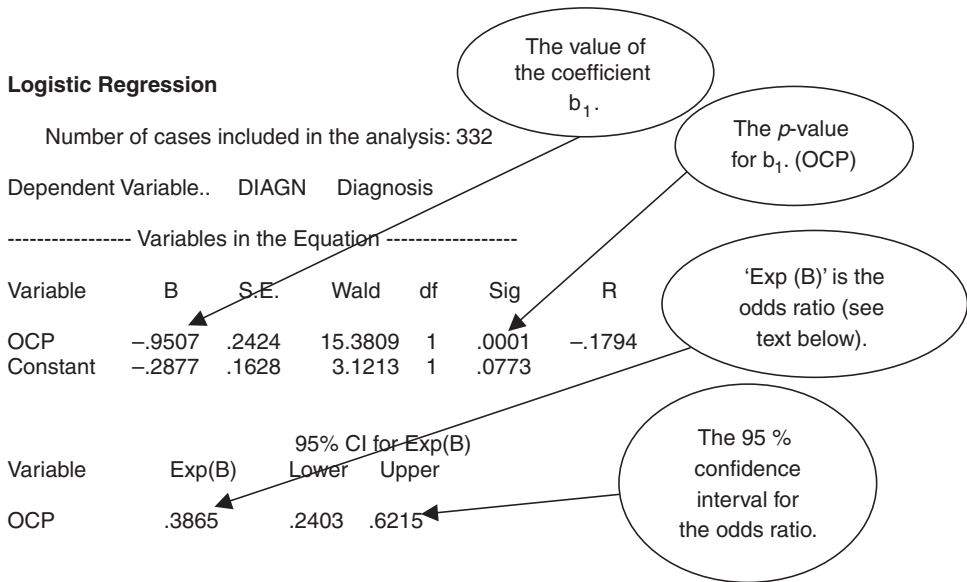


Figure 18.3 Abbreviated output from SPSS for a logistic regression with *diagnosis* as the dependent variable, and *use of oral contraceptive pill (OCP)* as the independent variable or risk factor

³ You'll first need to work out the values of $(b_0 + b_1 \times OCP)$, then $(1 + b_0 + b_1 \times OCP)$, then raise e to each of these powers. Then divide the former by the latter.

Binary Logistic Regression: Diagnosis versus OCP?

Response Information

Variable	Value	Count
Diagnosis	1	106 (Event)
	0	226
Total		332

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-0.2877	0.1628	-1.77	0.077			
OCP? 1	-0.9507	0.2424	-3.92	0.000	0.39	0.24	0.62

Log-Likelihood = -200.009

Test that all slopes are zero: G = 15.860, DF = 1, P-Value = 0.000

* NOTE * No goodness of fit tests performed.
 * The model uses all degrees of freedom.

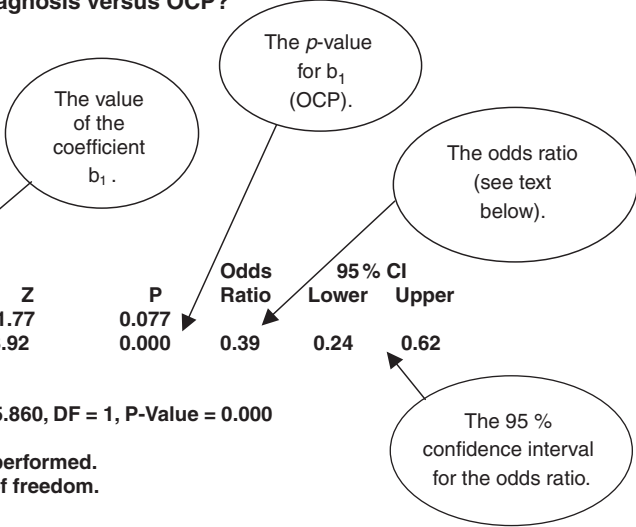


Figure 18.4 Output from Minitab for a logistic regression with *Diagnosis* as the dependent variable and *Use of Oral Contraceptive Pill (OCP)* as the independent variable or risk factor

The odds ratio

The great attraction of the logistic regression model is that it readily produces odds ratios. But how? There's quite a lot of maths involved, but eventually we can get to the following result:

$$\text{Odds ratio} = e^{b_0+b_1} / e^{b_0} = e^{b_1}$$

It is this ability to produce odds ratios that has made the logistic regression model so popular in clinical studies. Thus to find the odds ratio all you need to do is raise e to the power b₁, easily done on a decent calculator.

For example, in our Diagnosis/OCP model, b₀ = -0.2877 and b₁ = -0.9507, so the odds ratio for a malignant diagnosis for woman using OCP compared to women not using OCP is:

$$\text{Odds ratio} = e^{-0.9507} = 0.386$$

In other words, a woman who has used OCP has only about a third of the odds of getting a malignant diagnosis as a woman who has not used OCP. This result seems to confirm our earlier result that use of OCP provides some protection against a malignancy. Of course we don't know whether this result is due to chance or whether this represents a real statistically

⁴ Making use of the rule: X^a/X^b = X^{a-b}.

significant result in the population. To answer this question we will need either a confidence interval for β_1 or a p -value. I'll deal with this question shortly.

Exercise 18.1 Explain why, in terms of the risk of using OCP and the probability of getting a malignant diagnosis, that the values $P(Y = 1) = 0.4286$ when $OCP = 0$, and $P(Y = 1) = 0.2247$, when $OCP = 1$, are compatible with an odds ratio = 0.386 for a malignant diagnosis, among women using OCP compared to women not using OCP.

Interpreting the regression coefficient

In linear regression, the coefficient b_1 represents the increase in Y for a unit increase in X . We are not so much interested in the meaning of b_1 in the logistic regression model, except to note that if the independent variable is ordinal or metric, then you might be more interested in the effect on the odds ratio of changes of *greater* than one unit. For example, if the independent variable is age, then the effect on the odds ratio of an increase in age of one year may not be as useful as say a change of 10 years. In these circumstances, if the change in age is c years, then the change in the odds ratio is e^{cb_1} .

Exercise 18.2 (a) In linear regression we can plot Y against X to determine whether the relationship between the two variables is linear. Explain why this approach is not particularly helpful when Y is a binary variable. What approach might be more useful? (b) Is age significant? (c) Figure 18.5 shows the output from Minitab for the regression of *diagnosis* on *age* for the breast cancer example. Use the Minitab values to write down the estimated logistic regression model. (d) Calculate the probability that the diagnosis will be malignant, $P(Y = 1)$, for women aged: (i) 45; (ii) 50. (e) Calculate $[1 - P(Y = 1)]$ in each case, and hence calculate the odds ratio for a malignant diagnosis in women aged 45 compared to women aged 50. Explain your result. (f) Confirm that the antilog_e of the coefficient on *age* is equal to the odds ratio. (g) What effect does an increase in *age* of 10 years have on the odds ratio?

Logistic Regression Table. Dependent variable is Diagnosis.						95% CI	
Predictor	Coef	SE Coef	Z	P	Odds Ratio	Lower	Upper
Constant	-6.4672	0.7632	-8.47	0.000			
Age	0.10231	0.01326	7.72	0.000	1.11	1.08	1.14

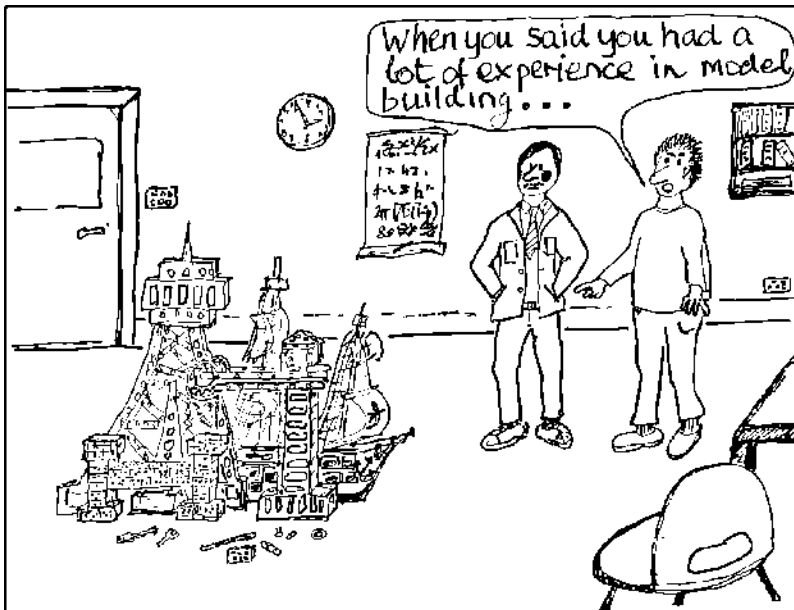
Figure 18.5 Output from Minitab for the logistic regression of *diagnosis* on *age*

Statistical inference in the logistic regression model

As you saw in Chapter 11, if the population odds ratio is equal to 1, then the risk factor in question has no effect on the odds for any particular outcome; that is, the variable concerned is *not* a statistically significant risk (or benefit). We can use either the p -value or the confidence interval to decide whether any departures from a value of 1 for the odds ratio is due merely to chance or is an indication of statistical significance.

In fact, in Figure 18.4, the 95 per cent confidence interval for the odds ratio for OCP use is (0.24 to 0.62), and since this does not include 1, the odds ratio is statistically significant in terms of receiving a malignant diagnosis. In addition the p -value = 0.000, so a lot less than 0.05. However, we still need to be cautious about this result because it represents only a crude odds ratio, which, in reality, would need to be adjusted for other possible confounding variables, such as age. We can make this adjustment in logistic regression just as easily as in the linear regression model, simply by including the variables we want to adjust for on the right-hand side of the model.

Notice that Minitab, Figure 18.4, uses the z distribution to provide a p -value, whereas SPSS, Figure 18.3, uses the *Wald statistic*, which can be shown to have a z distribution in the appropriate circumstances.



Exercise 18.3 Figure 18.6 shows the output from SPSS for the regression of *diagnosis* on *body mass index* (BMI). Comment on the statistical significance of body mass index as a risk factor for receiving a malignant diagnosis.

		B	Wald	Sig.	Exp(B)	95.0% CI for EXP(B)	
						Lower	Upper
Step 1(a)	BMI	.082	10.943	.001	1.085	1.034	1.139
	Constant	-2.859	19.313	.000	.057		

Figure 18.6 The output from SPSS for the regression of *diagnosis* on *body mass index* (some columns are missing)

The multiple logistic regression model

In my explanation of the odds ratio above I used a simple logistic regression model, i.e. one with a single independent variable (OCP), because this offers the simplest treatment. However, the result we got, that the odds ratio is equal to e^{b_1} , applies to *each* coefficient if there is more than one independent variable, i.e. e^{b_2} , e^{b_3} , etc. The usual situation is to have a risk factor variable plus a number of confounder variables (the usual suspects – age, sex, etc.). Suppose, for example, that you decided to include *age* and *body mass index* (BMI) along with OCP as independent variables. Equation (1) would then become:

$$P(Y = 1) = (e^{\beta_0 + \beta_1 \times \text{OCP} + \beta_2 \times \text{age} + \beta_3 \times \text{BMI}}) / (1 + e^{\beta_0 + \beta_1 \times \text{OCP} + \beta_2 \times \text{age} + \beta_3 \times \text{BMI}})$$

$P(Y = 1)$ is still of course the probability that the woman will receive a malignant diagnosis, $Y = 1$. The odds ratio for *age* is e^{b_2} ; the odds ratio for BMI is e^{b_3} . Moreover, as with linear regression, each of these odds ratios is *adjusted* for any possible interaction between the independent variables.

As an example, output from Minitab for the above multiple regression model of *diagnosis* against *use of oral contraceptives* (OCP), *age* and *body mass index* (BMI), is shown in Figure 18.7.

Exercise 18.4 Comment on what is revealed in the output in Figure 18.7 about the relationship between diagnosis and the three independent variables shown.

Building the model

The strategy for model building in the logistic regression model is similar to that for linear regression:

- Make a list of candidate independent variables.
- For any nominal or ordinal variables in the list construct a contingency table and perform a chi-squared test.⁵ Make a note of the *p*-value.

⁵ Provided the number of categories isn't too big for the size of your sample – you don't want any empty cells or low expected values (see Chapter 14)

Binary Logistic Regression: Diagnosis versus OCP, Age, BMI

Variable	Value	Count	
Diagnosis (0=benign)	1	106	(Event)
	0	224	
	Total	330	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds	95% CI	
					Ratio	Lower	Upper
Constant	-9.24814	1.30391	-7.09	0.000			
OCP	0.356767	0.329147	1.08	0.278	1.43	0.75	2.72
Age	0.111670	0.0164348	6.79	0.000	1.12	1.08	1.15
BMI	0.0812739	0.0275908	2.95	0.003	1.08	1.03	1.14

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	329.603	321	0.358
Deviance	328.516	321	0.374
Hosmer-Lemeshow	2.581	8	0.958

Figure 18.7 Minitab output for the model *diagnosis* against OCP, *age* and BMI

- For any metric variables, perform either a two-sample t test, or a univariate logistic regression; note the *p*-value in either case.
- Pick out all those variables in the list whose *p*-value is 0.25 or less. Select the variable with the smallest *p*-value (if there is more than one with the smallest *p*-value pick one arbitrarily) to be your first independent variable. This is your starting model.
- Finally, add variables to your model one at a time, each time examining the *p*-values for statistical significance. If a variable, when added to the model, is not statistically significant, drop it, unless there are noticeable changes in coefficient values, which is indicative of confounding.

Goodness-of-fit

In the linear regression model you used R^2 to measure goodness-of-fit. In the logistic regression model measuring goodness-of-fit is much more complicated, and can involve graphical as well as numeric measures. Two numeric measures that can be used are the *deviance coefficient* and the *Hosmer-Lemeshow statistic*. Very briefly, both of these have a chi-squared distribution, and we can use the resulting *p*-value to reject, or not, the null hypothesis that the model *provides a good fit*. The graphical methods are quite complex and you should consult more specialist

sources for further information on this and other aspects of this complex procedure. Hosmer and Lemeshow (1989) is an excellent source.

Exercise 18.5 Use the Hosmer-Lemeshow goodness-of-fit statistic in the output of Figure 18.7 to comment on the goodness-of-fit of the model shown.

Linear and logistic regression modelling are two methods from a more general class of methods known collectively as *multivariable* statistics. *Multivariate* statistics on the other hand, is a set of procedures applicable where there is more than one dependent variable, and includes methods such as principal components analysis, multidimensional scaling, cluster and discriminant analysis, and more. Of these, principal components analysis appears most often in the clinical literature, but even so is not very common. Unfortunately, there is no space to discuss any of these methods.

IX

Two More Chapters

19

Measuring survival

Learning objectives

When you have finished this chapter you should be able to:

- Explain what censoring means.
- Calculate Kaplan-Meier survival probabilities.
- Draw a Kaplan-Meier survival curve.
- Use the Kaplan-Meier curve to estimate median survival time (if possible).
- Explain the use of the log-rank test to determine if the survival experience of two or more groups is significantly different.
- Explain the role of the hazard ratio in comparing the relative survival experience of two groups.
- Outline the general idea behind Cox proportional hazards regression and interpret the results from such a regression.

Introduction

Imagine that you have a patient who has overdosed on paracetamol. A spouse asks you what their chances of 'coming through it' are. Or suppose a patient with breast cancer wants to

know which of two possible treatments offers the best chance of survival. You can answer questions like these with the help of a procedure known as *survival analysis*. The basis of this method is the measurement of the time from some *intervention* or *procedure* to some *event of interest*.

For example, if you were studying survival after mastectomy for breast cancer (the procedure), you would want to know how long each woman survived following surgery. Here, the event of interest would be death. For practical reasons you usually have to limit the duration of the study, for example, to one year, or five years, etc. Very often you will want to compare the survival experiences of two groups of patients; for example women having a mastectomy, with women having less radical surgery.

Censored data

One particular problem, which makes this type of analysis tricky, is that you often don't observe the event of interest in *all* of the subjects. For example, after five years, by no means all of the women will have died following the mastectomy. We don't know how long these particular patients will live after the end of the study period, only that they are still alive when the study period ends. In addition, some patients may withdraw from the study during the study period; they may move away, or simply refuse further participation, or die from a cause unrelated to the study. These types of incomplete data are said to be *censored*.

A final problem is that not all patients may enter the study at the same time. Fortunately, methods have been developed to deal with these difficulties. One of which, known as the *Kaplan–Meier method*, gives us a table of survival probabilities which can be charted as the Kaplan–Meier chart. The two questions that are often of the greatest interest are:

- What's the probability of a patient surviving for some given period of time?
- What's the *comparative* survival experience of two groups of patients?

A simple example of survival in a single group

Look at the data in Table 19.1, which shows survival data (in months) for a group of 12 patients diagnosed with a brain tumour, who were followed up for 12 months.

Table 19.1 shows that seven patients died, two left the study prematurely and three survived. This means that you have seven definite and five censored survival times. We can represent the survival times in the last column of Table 19.1 graphically, as in Figure 19.1, where the survival times are arranged in *ascending* order.

Calculating survival probabilities and the proportion surviving: the Kaplan–Meier table

The Kaplan–Meier method requires a Kaplan–Meier table like Table 19.2, with, strictly speaking, rows *only* for time periods when a death occurs (shown in bold in the table). However, I have

Table 19.1 Survival times (months) over a 12-month study period, of 12 patients diagnosed with brain tumour. *Indicates censored data (patient survived, S, or left study prematurely, P). The *actual* survival time for these patients is not known

Patient	Month of entry to study (0 indicates present at beginning of study)	Time after study start date to death or censoring (months)	Outcomes: Died (D), Survived (S) or left study prematurely (P)	Survival times
1	0	12	S*	12
2	0	12	S*	12
3	0	11	D	11
4	0	8	D	8
5	1	6	P*	5
6	2	12	S*	10
7	2	4	D	2
8	2	5	D	3
9	2	9	D	7
10	3	9	P*	6
11	3	8	D	5
12	3	7	D	4

included all 12 rows in the table to help illustrate the method more clearly. The second column tells us how many people were still alive, n , at the beginning of each month, t . This will equal the total initial number of patients in the study, minus both the total number of deaths and the total number of premature withdrawals up to the beginning of the month. Column 4 records the number of deaths d in each month. Column 5 records the total number at risk during the month, r . By dividing column 4 by column 5, we get d/r , the probability that a patient still alive at the beginning of the month will die during it (which is equivalent to the *proportion* of patients dying in that month). The value of d/r is shown in column 6.

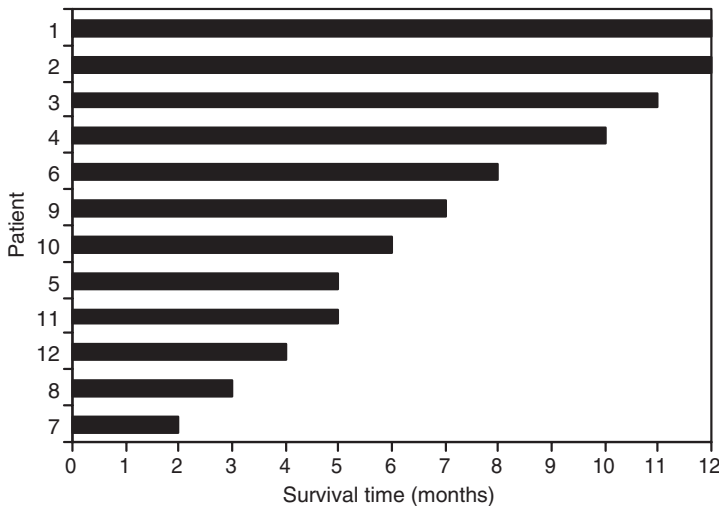


Figure 19.1 Chart of survival times (in ascending order) from Table 19.1

Table 19.2 Calculation of Kaplan-Meier survival probabilities

1	2	3	4	5	6	7	8
Month	Number still in study at start of month t	Withdrawn prematurely during month t	Deaths in month t	Number at risk in month t	Probability of death in month t	Probability of surviving month t	Cumulative probability of surviving to month t
t	n	w	d	r	d/r	$p = 1 - d/r$	S
1	12	0	0	12	0	1	1
2	12	0	0	12	0	1	1
3	12	0	0	12	0	0	1
4	12	0	1	11	$1/11 = 0.091$	0.909	0.909
5	11	0	1	10	$1/10 = 0.100$	0.900	0.818
6	10	1	0	9	0	1	1
7	9	0	1	8	$1/8 = 0.125$	0.875	0.716
8	8	0	2	6	$2/6 = 0.333$	0.667	0.478
9	6	1	1	4	$1/4 = 0.250$	0.750	0.358
10	4	0	0	4	0	1	1
11	4	0	1	3	$1/3 = 0.333$	0.667	0.239
12	3	0	0	3	0	1	1

Since d/r is the probability of dying during a time period, then $(1 - d/r)$ must be the probability of surviving to the end of the time period. This survival probability is shown in column 7. To calculate the probability of surviving *all* of the preceding time periods *and* the current time period, you must successively multiply the probabilities in column 7 together (ignoring any 0's). The resultant cumulative probabilities, labelled S , are shown in column 8. For example, the value for S of 0.818 in row 5 is $1 \times 1 \times 1 \times 0.909 \times 0.900$. These column 8 values are the *Kaplan-Meier survival probabilities*.

Table 19.2 indicates that the probability of a patient surviving to the end of the third month is 1, to the end of the fourth month is 0.909, and so on, and for the full 12 months after the diagnosis is 0.239.

We can also interpret these values as *proportions*. For example, 0.909 of the patients (or 90.9 per cent, will survive to the end of the fourth month. About a quarter (23.9 per cent) will survive the full 12 months. We can generalise these results to the *population* of patients of whom this sample is representative, and who have the same type of brain tumour, at the same stage of development, and receive the same level of care. In addition, we may want to adjust for possible confounding variables such as age, sex, etc. We'll deal with this question later.

The Kaplan-Meier chart

If you plot the cumulative survival probabilities in the last column of Table 19.2 against time, you get the *Kaplan-Meier curve*, shown in Figure 19.2. Notice that the survival 'curve' looks like a staircase, albeit with uneven steps. Every time there is a death, the curve steps down. Since there are seven deaths, there are seven steps down.¹

¹ Notice there is a double step down at period 8 because of the two deaths.

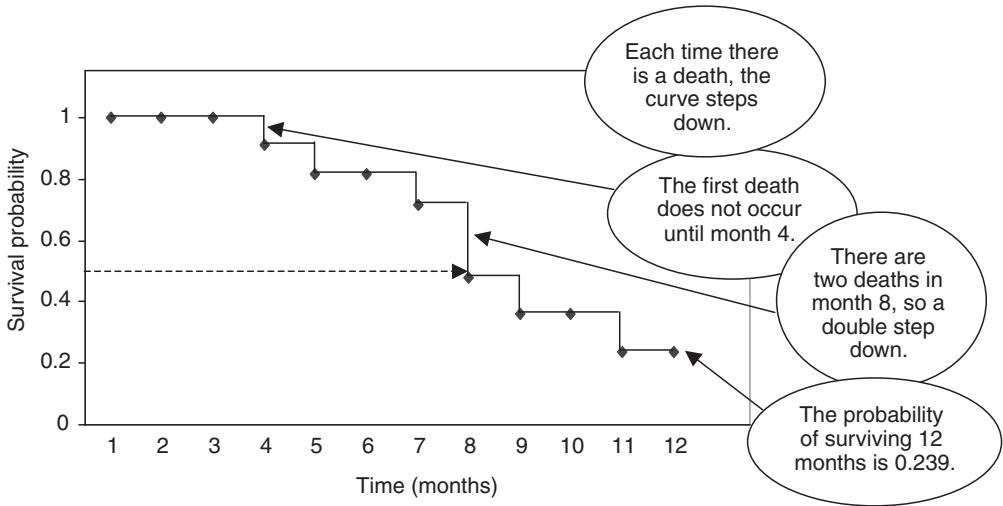


Figure 19.2 The Kaplan-Meier survival curve drawn from the data in Table 19.2 (the dotted line indicates median proportion surviving – see text below)

Exercise 19.1 The data in Table 19.3 shows the survival times (in days) of eight patients with acute myocardial infarction, treated with a new reperfusion drug Explase, as part of a fibrinolytic regimen. Patients were followed up for 14 days. Calculate survival probabilities and plot Kaplan-Meier survival curves. Comment on your results.

Table 19.3 The survival times (in days) of eight patients with acute myocardial infarction. Patients were followed up for 14 days

Patient	Day of entry to study (0 indicates present at beginning of study)	Time after study start date to death or censoring (days)	Outcomes: Died (D), Survived (S) or Left study prematurely (P)
1	0	3	D
2	0	14	S
3	0	8	D
4	0	12	P
5	1	14	S
6	2	13	D
7	2	14	S
8	2	14	S

Determining median survival time

One of the consequences of not knowing the actual survival times of all of those subjects who survive beyond the end of the study period is that we cannot calculate the mean survival time of the whole group. However, if you interpret the probabilities on the vertical axis of

a Kaplan-Meier chart as proportions or percentages, you can often easily determine *median* survival times. It is that value which corresponds to a probability of 0.5 (i.e. 50 per cent). In Figure 19.2, the median survival time is 8 months. At this time, half of the patients still survived. Obviously the survival time of any proportion of the sample can be determined in this same way, including the interquartile range values, provided that the Kaplan-Meier curve goes down far enough (unfortunately it often doesn't).

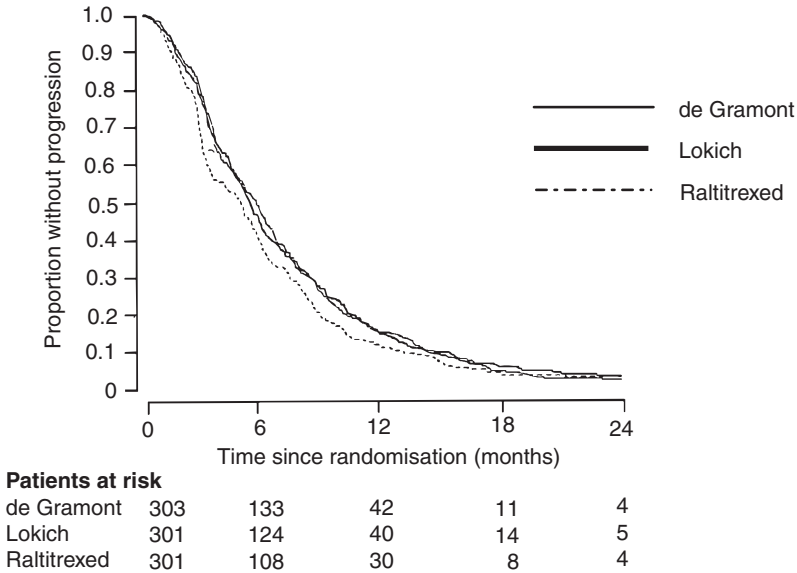
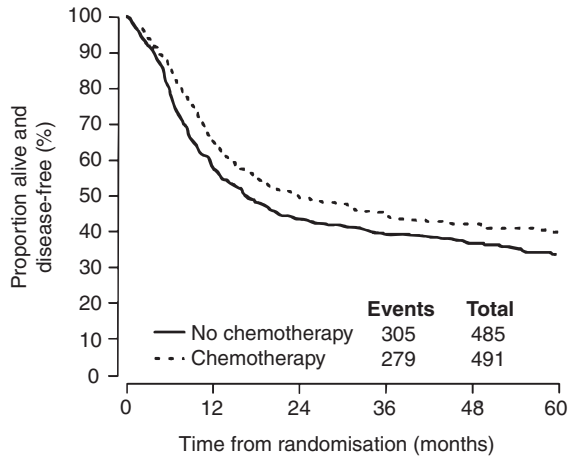


Figure 19.3 Kaplan-Meier curves for overall survival for three groups of patients in a comparison of three chemotherapy regimes in the treatment of colorectal cancer. Reprinted from *The Lancet*, 2002, 359, 1559 with permission from Elsevier

Exercise 19.2 Figure 19.3 shows Kaplan-Meier curves for progression-free survival, for three groups of patients in a comparison of three chemotherapy regimes used for the treatment of colorectal cancer (Maughan *et al.* 2002). The three regimes were: the de Gramont regimen; the Lokich regimen; and Raltitrexed. What were the approximate median survival times for progression-free survival with each of the three regimes?

Comparing survival with two groups

Although the survival curve for a single group may sometimes be of interest, much more usual is the desire to compare the survival experience of two or more groups. For example, Figure 19.4 is taken from a study of chemotherapy for the treatment of bladder cancer (Medical Research Council Advanced Bladder Working Group 1999). One group of patients ($n = 485$) was randomly assigned to receive conventional radical surgery (cystectomy) or radiotherapy, while a second group ($n = 491$) received the conventional treatment *plus* chemotherapy. The



Patients at risk

No chemotherapy	485	271	192	145	102	63
Chemotherapy	491	308	222	163	116	75

Figure 19.4 Survival curves for two groups of patients with bladder cancer, one group given conventional surgery or radiotherapy, the other group given the conventional treatment *plus* chemotherapy. Reprinted courtesy of Elsevier (*The Lancet*, 1999, Vol No. 354, p. 533–9)

question asked, ‘Was the survival experience of the chemotherapy group any better over the five year follow up?’

The two Kaplan-Meier curves seem to show that the proportions surviving in the chemotherapy group was larger than those in the conventional group throughout the duration of the study, since the survival curve for the former was higher than that of the latter. In fact, the authors of this study report *median* values for disease-free survival of 20 months for the chemotherapy group and 16.5 months for the no-chemotherapy group. The 95 per cent confidence interval for the difference in medians was (0.5 to 7.0) months, so the difference in medians was statistically significant.

Notice that the authors have provided a table showing the numbers at risk at each time interval. This is to remind us that the smaller numbers of survivors towards the end of a trial produce less reliable results. As a direct consequence of this effect, you should not assume that just because the gap between two survival curves gets progressively larger (as it is often seen to do), that this is *necessarily* due to an actual divergence in the survival experiences in the two groups. It might well be caused simply by the low numbers of subjects still at risk. This can make the ends of the curves unreliable.

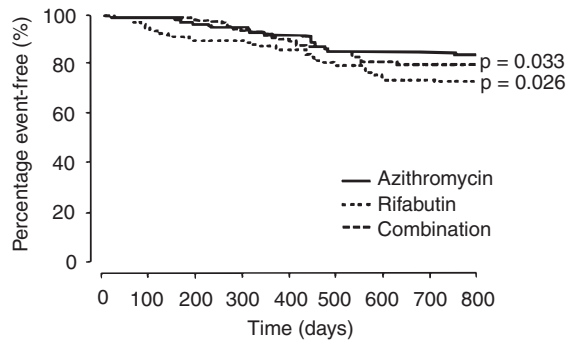
The log-rank test

If you want to compare the *overall* survival experience of two (or more) groups of patients (rather than say comparing just the median survival times as we did above), then one possible approach is to use the non-parametric *log-rank test*. Essentially, the null hypothesis to be tested is that the two samples (the two groups) are from the same population as far as their survival experience is concerned. In other words there is *no difference* in the survival experiences.

The log-rank test of this hypothesis uses a comparison of observed with expected events (deaths, say), given that the null hypothesis is true.² If the p value is less than 0.05 you can reject the null hypothesis and conclude that there is a statistically significant difference between the survival experience of the groups. You can then use the Kaplan-Meier curves to decide which group had the significantly better survival. A limitation of the log-rank test is that it cannot be used to explore the influence on survival of more than one variable, i.e. the possibility of confounders – for this you need Cox’s proportional regression, which we’ll come to shortly.

The authors in the bladder cancer study reported a log-rank test p value of 0.019 for the difference in survival times at *three years*, but unfortunately don’t give the results of the test over the whole five year duration of the study.

Exercise 19.3 What do you conclude about the statistical significance of the difference in three year survival times of the chemotherapy and non-chemotherapy groups from the results given in the previous paragraph?



Number of patients at risk

Azithromycin	233	185	141	117	78	50	29	11	2
Rifabutin	236	172	133	106	72	44	26	10	0
Combination	224	178	150	121	89	52	31	16	2

Figure 19.5 Kaplan-Meier curves from a study to assess the clinical efficacy of azithromycin for prophylaxis of *Pneumocystitis carinii* pneumonia in HIV-1 infected patients. Reprinted courtesy of Elsevier (*The Lancet*, 1999, Vol No. 354, p. 1891–5)

An example of the log-rank test in practice

Figure 19.5 shows the Kaplan-Meier curves from a study to assess the clinical efficacy of azithromycin for prophylaxis of *Pneumocystitis carinii* pneumonia in HIV-1 infected patients (Dunne *et al.* 1999). Patients were randomly assigned to one of three treatment groups: the first group given azithromycin, the second rifabutin and the third a combination of both drugs. The figure shows the event-free (no *Pneumocystitis carinii* pneumonia) survival experiences over an 800 day period for the three treatment groups.

² You may have spotted the similarity with the chi-squared test considered earlier in the book. In fact the calculations are exactly the same.

The log-rank test was used to test the hypothesis that there is no difference in the percentage event-free between the azithromycin and rifabutin groups (p value = 0.033), and between the azithromycin and the combination groups (p -value = 0.026). The authors concluded that azithromycin as prophylaxis for *Pneumocystis carinii* pneumonia, provides additional protection over and above standard *Pneumocystis carinii* pneumonia prophylaxis. However, these results should be treated with caution because of the very small size of the survivor group towards the end of the study.

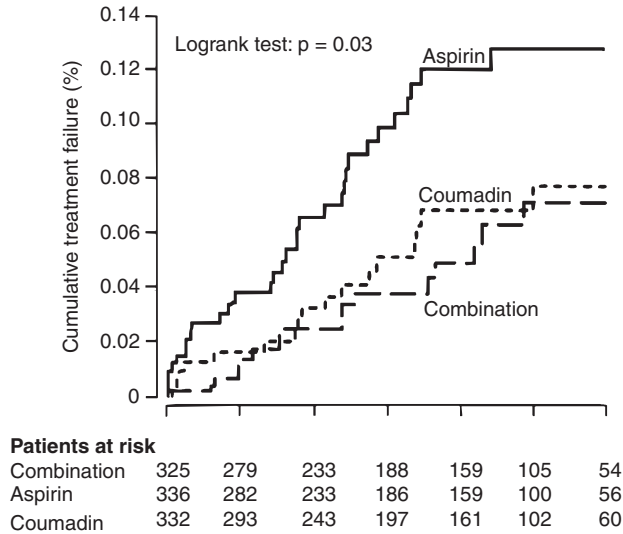


Figure 19.6 Kaplan-Meier curves of percentage number of subsequent ischaemic events from a randomised controlled trial into the relative effectiveness of aspirin and oral anticoagulants (coumadin), used for antiplatelet treatment, following myocardial infarction. Reprinted courtesy of Elsevier (*The Lancet* 2002, **360**, 109–14, Fig. 3, p. 111)

Exercise 19.4 Figure 19.6 shows the Kaplan-Meier curves for the percentage number of ischaemic events from a randomised controlled trial into the relative effectiveness of aspirin and oral anticoagulants (coumadin) for antiplatelet treatment following myocardial infarction (van Es *et al.* 2002). The object was to investigate which of these two drugs is more effective for the long-term reduction of subsequent ischaemic events, and whether the combination of the two drugs offers greater benefit than either drug alone. Is there a statistically significant difference in mortality between the three possible treatments? Which treatment seems to offer the best survival?

The hazard ratio

The log-rank test is limited by the fact that it is just that – a test. It will tell you if there is a significant difference between the survival experience of two (or more) groups, but does not quantify that difference. For this we need what is called the *hazard ratio* (based on the ratio

of observed and expected events for the two groups), along with which we can calculate a confidence interval. As a matter of interest, the authors of the bladder cancer study reported, for those alive and disease free, a hazard ratio of 0.82, with a 95 per cent confidence interval of (0.70 to 0.97). We can interpret this result to mean that the group who had chemotherapy had a risk of dying in the study period of only 82 per cent compared to the risk for the non-chemotherapy group, and this difference was statistically significant (confidence interval does not include 1).

Exercise 19.5 The survival curves shown in Figure 19.4 from the bladder cancer study are for subjects who are alive *and* disease free. For subjects who were alive but not necessarily disease free, the authors report the following results. What do these results tell you?

Comparison of the survival time in the two groups gave a hazard ratio of 0.85 [95 per cent CI of (0.71 to 1.02)]. The absolute difference in 3-year survival was 5.5 per cent, 50.0 per cent in the chemotherapy group, 55.5 per cent in the non-chemotherapy group [95 per cent CI of (−0.5 to 11.0)]. The median survival time for the chemotherapy group was 44 months and for the no-chemotherapy group was 37.5 months [95 per cent CI of (−0.5 to 15)].

The proportional hazards (or Cox's) regression model

Although researchers can use the log-rank test to distinguish survival between two groups, the test only provides a *p* value; it would be more useful to have an estimate of any difference in survival, along with the corresponding confidence interval. The hazard ratio mentioned above provides this, but neither the log-rank test nor the simple hazard ratio allow for adjustment for possible confounding variables, which may significantly affect survival. For this we can use an approach known as *proportional hazards (or Cox's) regression*. This procedure will provide both estimates and confidence intervals for variables that affect survival, and enable researchers to adjust for confounders. We will discuss briefly the principle underlying the method, and the meaning of some of the terms used.

The focus of *proportional hazards regression* is the *hazard*. The hazard is akin to a failure rate. If the end-point is death, for example, then the hazard is the rate at which individuals die at some point during the course of a study. The hazard can go up or down over time, and the distribution of hazards over the length of a study is known as the *hazard function*. You won't see authors quote the hazard regression function or equation, but for those interested it looks like this:

$$\text{Hazard} = h_0 + e^{(\beta_1 X_1 + \beta_2 X_2 + \dots)}$$

h_0 is the baseline hazard and is of little importance. The explanatory or independent variables can be any mixture of nominal, ordinal or metric, and nominal variables can be 'dummied', as described in Chapter 17 and Chapter 18. The same variable selection procedures as in linear or logistic regression models can also be used, i.e. either automated or by hand.

The most interesting property of this model is that e^{β_1} , e^{β_2} , etc. give us the *hazard ratios*, or HRs, for the variables X_1 , X_2 , and so on (notice the obvious similarity with the odds ratios in

logistic regression). The hazard ratios are essentially *risk ratios*, but called hazard ratios in the context of survival studies. For example, in a study of the survival of women with breast cancer, the variable X_1 might be ‘micrometastases present (Y/N)’. In which case, the hazard ratio HR_1 (the risk of death for a patient when micrometastases are present compared to that for a patient where they are absent, is equal to e^{b_1} . All of this is only true if the relative effect (essentially the *ratio*) of the hazard on the two groups (for example, the relative effect of micrometastases on the survival of each group) remains constant over the whole course of the study.

An application from practice

As an example of proportional hazards regression, Table 19.4 is taken from a study into the relative survival of two groups of patients with non-metastatic colon cancer; one group having open colectomy (OC), the other laparoscopy-assisted colectomy (LAC) (Lacy *et al.* 2002). The figure shows hazard ratios and their confidence intervals: for the probability of being free of recurrence; for overall survival; and for cancer-related survival, after the patients were stratified according to tumour stage.

So, for example, patients with lymph-node metastasis do only about a third as well in terms of being recurrence-free over the course of the study compared to patients without lymph-node metastasis (hazard ratio = 0.31), and this difference is statistically significant since the confidence interval does not include 1 (and the *p* value of 0.0006 is < 0.05). Patients with lymph-node metastasis also compare badly with non-metastasis patients in terms of both

Table 19.4 Results of a Cox proportional hazards regression analysis comparing the survival of patients with laparoscopy-assisted colectomy versus open colectomy, for the treatment of non-metastatic colon cancer. Reproduced courtesy of Elsevier (*The Lancet*, 2002, Vol No. 359, page 2224–30

	Hazard ratio (95% CI)	p
Probability of being free of recurrence		
Lymph-node metastasis (presence vs absence)	0.31 (0.16–0.60)	0.0006
Surgical procedure (OC vs LAC)	0.39 (0.19–0.82)	0.012
Preoperative serum CEA concentrations (≥ 4 ng/mL vs < 4 ng/mL)	0.43 (0.22–0.87)	0.018
Overall survival		
Surgical procedure (OC vs LAC)	0.48 (0.23–1.01)	0.052
Lymph-node metastasis (presence vs absence)	0.49 (0.25–0.98)	0.044
Cancer-related survival		
Lymph-node metastasis (presence vs absence)	0.29 (0.12–0.67)	0.004
Surgical procedure (OC vs LAC)	0.38 (0.16–0.91)	0.029

Type of surgical procedure, laparoscopy-assisted vs open colectomy, is significantly beneficial in terms of recurrence-free and cancer-related survival, but not in terms of overall survival.

OC = open colectomy; LAC = laparoscopy-assisted colectomy; CEA = carcinoembryonic antigen.

overall survival (only about half as well, HR = 0.49), and cancer-related survival (just over a quarter as well, HR = 0.29). Both of these results are statistically significant. Note that type of surgery; laparoscopy-assisted versus open colectomy, is not statistically significant in terms of overall survival as the confidence interval of (0.23 to 1.01) includes 1.

Table 19.5 Hazard ratios due to a number of risk factors in a univariate (unadjusted), and multivariate (adjusted) cohort analysis of the risk to HIV+ women of vulvovaginal and perianal condylomata acuminata and intraepithelial neoplasia. Reproduced courtesy of Elsevier (*The Lancet*, 2002, Vol No. 359, page 108–14

Risk factor	Number of women	Univariate analysis*		Multivariate analysis†	
		Hazard ratio (95% CI)	p	Adjusted hazard ratio (95% CI)	p
HIV-1 infection	726	17.0 (4.07–70.9)	0.0007	6.96 (1.51–32.2)	0.01
CD4 T lymphocyte count‡	707	3.38 (2.24–5.10)	<0.0001	1.66 (1.03–2.69)	0.04
Human papillomavirus infection	699	4.86 (2.21–10.7)	0.0006	3.76 (1.67–8.43)	0.0013
History of injecting two or more drugs three or more times per week	726	3.09 (1.57–6.07)	0.003	2.32 (1.14–4.71)	0.02
Less than a highschool education	725	2.15 (1.09–4.22)	0.03	1.99 (1.00–3.98)	0.05
Cigarette smoking at enrolment	726	0.84 (0.43–1.64)	0.61	0.71 (0.35–1.44)	0.34
Age <35 years at enrolment	726	1.85 (0.93–3.68)	0.08		
Currently unmarried	726	2.48 (0.96–6.38)	0.06		
Annual income <US\$ 10 000	711	1.15 (0.56–2.34)	0.71		
First sex at <16 years of age	723	1.33 (0.69–2.59)	0.40		
>7 lifetime sex partners	722	1.40 (0.71–2.79)	0.33		
History of prostitution	722	1.83 (0.90–3.74)	0.10		
History of ever injecting drugs	726	1.74 (0.90–3.39)	0.10		
History of sexually transmitted disease§	654	1.58 (0.72–3.45)	0.25		

*In univariate analysis, vulvovaginal lesion was the outcome variable.

†355 HIV-1-positive and 325 HIV-1-negative women were included in the multivariate analysis, with vulvovaginal or perianal lesion as the outcome variable and HIV-1 infection, CD4 T lymphocyte count, human papillomavirus infection, less than a highschool education, cigarette smoking, and history of injection of two or more drugs three or more times per week as covariates.

‡ HIV-1 negative women were presumed to have a CD4 count >500 cell/, μL .

§ Does not include a history of genital warts.

Exercise 19.6 Table 19.5 shows the hazard ratios (unadjusted and adjusted) due to a number of risk factors in a cohort analysis of the risk to HIV+ women of vulvovaginal and perianal condylomata acuminata and intraepithelial neoplasia (Conley *et al.* 2002). Interpret the multivariate results. How do these differ from the univariate results?

Checking the proportional hazards assumption

The proportional hazards assumption can be checked graphically using what is known as the *log-log* plot. Unfortunately, this procedure is beyond the scope of this book.

20

Systematic review and meta-analysis

Learning objectives

When you have finished this chapter you should be able to:

- Provide a broad outline of the idea of systematic review.
- Outline a typical search procedure.
- Describe what is meant by publication bias and its implications.
- Describe how we can use the funnel plot to examine for the presence of publication bias.
- Explain the importance of heterogeneity across studies and how the L'Abbé plot can be used in this context.
- Explain the meaning of meta-analysis.
- Outline the role of the Mantel-Haenszel procedure in combining studies.
- Describe what a forest plot is and how it is used.

Introduction

If you have a patient with a particular condition and you want to know the current consensus on the most effective treatment, then you could perhaps ask the opinions of colleagues (although they may know no more than you) or maybe look through some pharmaceutical publicity material. Or read all the relevant journals lying around your clinic or office. Better still, if you have access to one of the clinical databases, such as Medline, then the job will be that much easier; in fact, anything like an adequate search is almost impossible otherwise. If you want your search to capture everything written on your topic then you will need a systematic approach. This process of searching for all relevant studies (or trials) is known as a *systematic review*.

However you do your systematic review, you are likely to encounter some difficulties:

- Many of the studies you turn up will be based on smallish samples. As you know, small samples may well produce unreliable results.
- Partly as a consequence of the above problem, many of the studies come to different and conflicting conclusions.
- There will be some studies that you simply do not find. Perhaps because they are published in obscure and/or non-English-language journals, or are not published at all (for example, internal pharmaceutical company reports, or research dissertations). This shortfall may lead to what is known as publication bias.

To some extent you can address the first two of these problems by combining all of these individual studies into one large study, as you will see later (a process called *meta-analysis*), and you will also want to deal with the potential for publication bias. But let's start with a brief description of systematic review.

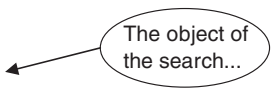
Systematic review

The basis of a systematic review is a comprehensive search that aims to identify all similar and relevant studies that satisfy a pre-defined set of *inclusion and exclusion criteria*. As an example, the following extract from a systematic review and meta-analysis of studies of dietary intervention to lower blood cholesterol, shows the inclusion and exclusion criteria, together with a brief description of the search procedure (Tang *et al.* 1998).

Methods

Identification of trials and extraction of data

We aimed to identify all unconfounded randomised trials of dietary advice to lower cholesterol concentration in free-living subjects published before 1996. Trials were eligible for



The object of the search...

inclusion if there were at least two groups, of which one could be considered a control group; treatment assignment was by random allocation; the intervention was a global dietary modification (changes to various food components of the diet to achieve the desired targets); and lipid concentration were measured before and after the intervention.

...the
inclusion
criteria...

Trials of diets to reduce fat intake in women considered to be at risk of breast cancer were included because the diets were similar to those aimed at lowering cholesterol concentration. We excluded trials of specific supplementation diets (such as those with particular oils or margarine, garlic, plant sterol, or fibre supplements, etc.), multifactorial intervention trials, trials aimed primarily at lowering body weight or blood pressure, and trials whose interventions lasted less than four weeks. Trials based on randomisation of workplace or general practice were also excluded.

...the
exclusion
criteria...

To identify these trials we identified four electronic databases (Medline, Human Nutrition, EMBASE, and Allied and Alternative Medicine). These databases included trials published after 1966. We also identified trials by hand searching the American Journal of Human Nutrition by scrutinising the references of review articles and of each relevant randomised trial, and by consulting experts on the subject.

...and the
search
strategy.

Reports that appeared only in non-English language journals were examined with the help of translators. Trials were categorised according to their approximate target diet into four groups.

The end result of a systematic review then, is a list of studies, each one of which provides a value for the specified outcome measure. In the above example, this outcome measure was the percentage difference in mean total blood cholesterol between the intervention (dietary advice) group and the control group. Examination of this list of outcome values may provide the required insights into treatment effectiveness.

Exercise 20.1 Briefly outline the systematic review procedure and some of the problems that may arise.

The forest plot

The list of studies produced by the systematic review is often accompanied by what is known as a *forest plot*. This plot has study outcome on the vertical axis, usually arranged by size of study (i.e. by sample size), and the outcome measure on the horizontal axis. The outcome measure

might be odds or risk ratios, means or proportions, or their differences, and so on. There are a number of ways of displaying the data. For example, by using a box with a horizontal line through it, whose length represents the width of the 95 per cent confidence interval for whatever outcome measure is being used. Or with a diamond, whose width represents the 95 per cent confidence interval. The area of each box or diamond should be proportional to its sample size. As an example, the forest plot for the cholesterol study referred to above is shown in Figure 20.1.

Here the 22 individual studies, each represented by a black square whose size is proportional to sample size, are divided into four groups according to their approximate target diet (we don't need to go into the details). The aggregated mean percentage reduction in cholesterol (with a 95 per cent confidence interval) for each of these groups is represented by a white square, whose size is proportional to the sample size of the aggregated individual studies. The large white square at the bottom of the plot is the aggregated value for all the studies combined. I'll come back to this shortly.

The horizontal axis represents mean percentage change in blood cholesterol. As you can see, 21 of the 22 studies show a reduction in percentage cholesterol (the study fourth from the top lies exactly on the zero, or no difference, line). However, in seven of the studies the confidence interval crosses the zero line, indicating that the reduction in cholesterol is not statistically significant. The remaining 15 studies show a statistically significant reduction (95 per cent confidence interval does not cross the zero line), as do all four group summary values. Thus there appears to be plenty of evidence that dietary interventions of the type included here do manage to achieve statistically significant reductions in total blood cholesterol.

Exercise 20.2 The results in Table 20.1 show the outcomes (relative risk for proportion of subjects with side effects), from each of six randomised trials comparing antibiotic with placebo for treating acute cough in adults (Fahey *et al.* 1998). Draw a forest plot of this data and comment briefly on what it shows. Note: relative risks greater than 1 favour the placebo (i.e. fewer side effects).

Table 20.1 The outcomes (relative risk for proportion of subjects with side effects), from each of six randomised trials comparing antibiotic with placebo for treating acute cough in adults. Reproduced from *BMJ* 1998, **316**: 906–10. Figure 4, p. 909. Figures 2 and 3, p. 908, courtesy of BMJ Publishing Group

Study	Sample size	Relative risk (95 % CI)
Briskfield <i>et al.</i>	50	0.51 (0.20 to 1.32)
Dunlay <i>et al.</i>	57	7.59 (0.43 to 134.81)
Franks and Gleiner	54	3.48 (0.39 to 31.38)
King <i>et al.</i>	71	2.30 (0.93 to 5.70)
Stott and West	207	1.49 (0.63 to 3.48)
Verheij <i>et al.</i>	158	1.71 (0.80 to 3.67)
Total	597	1.51 (0.86 to 2.64)

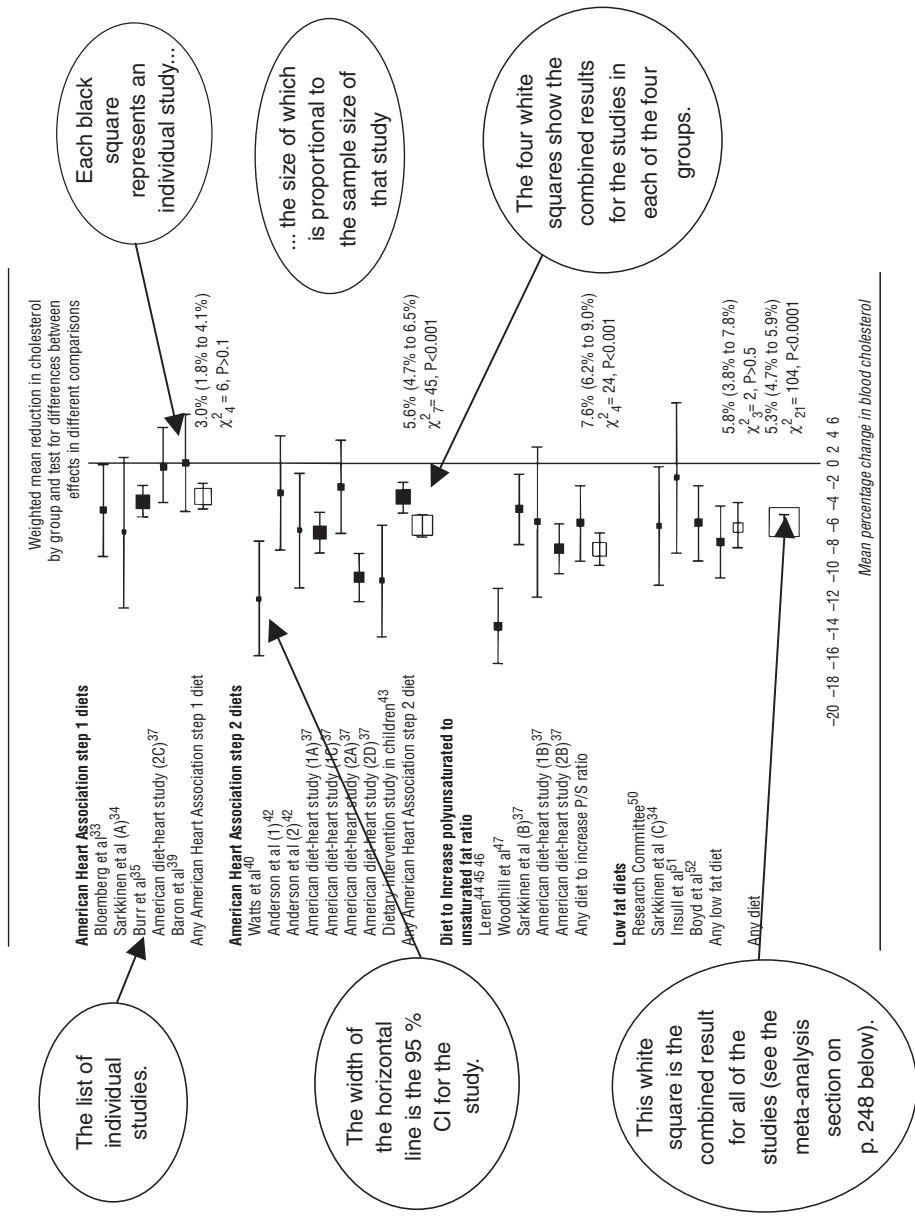


Figure 20.1 Forest plot for the dietary intervention and blood cholesterol study. Mean percentage changes (with 95 per cent confidence intervals) in blood total cholesterol concentration. Reproduced from *BMJ* 1998, **316**: 1213-20, courtesy of BMJ Publishing Group

Publication and other biases

The success of any systematic review depends critically on how thorough and wide-ranging the search for relevant studies is. One frequently quoted difficulty is that of *publication bias*, which can arise from a number of sources:

- The tendency for journals to favour the acceptance of studies showing *positive* outcomes at the expense of those with negative outcomes.
- The tendency for authors to favour the submission to journals of studies showing *positive* outcomes at the expense of those with negative outcomes.
- Studies with positive results are more likely to be published in English language journals giving them a better chance of capture in the search process.
- Studies with positive results are more likely to be cited, giving them a better chance of capture in the search process.
- Studies with positive results are more likely to be published in more than one journal, giving them a better chance of capture in the search process.
- Some studies are never submitted for publication. For example, those that fail to show a positive result, those by pharmaceutical companies (particularly if the results are unfavourable), graduate dissertations and so on.

In the light of all this it is important that possible presence of publication bias should be addressed. One possibility is to use what is known as a *funnel plot*.

The funnel plot

In a funnel plot the *size* of the study is shown on the vertical axis and the size of the treatment's effect (for example, as measured by an odds or risk ratio, or a difference in means, etc.) is shown on the horizontal axis. In the absence of bias the funnel plot should have the shape of a *symmetric* upturned cone or funnel. Larger studies shown at the top of the funnel will be more precise (their results will not be so spread out), smaller studies, shown towards the bottom less precise, and therefore more spread out. These differences produce the funnel shape. However, if the funnel is asymmetrical, for example, if parts of the funnel are missing or poorly represented – and this will usually be near the bottom of the funnel where the smaller studies are located – then this is suggestive of bias of one form or another.¹

¹ There are a number of other possible causes of bias in systematic reviews. Those interested should look, for example, at Egger and Davey Smith (1998), where other possible biases are discussed.

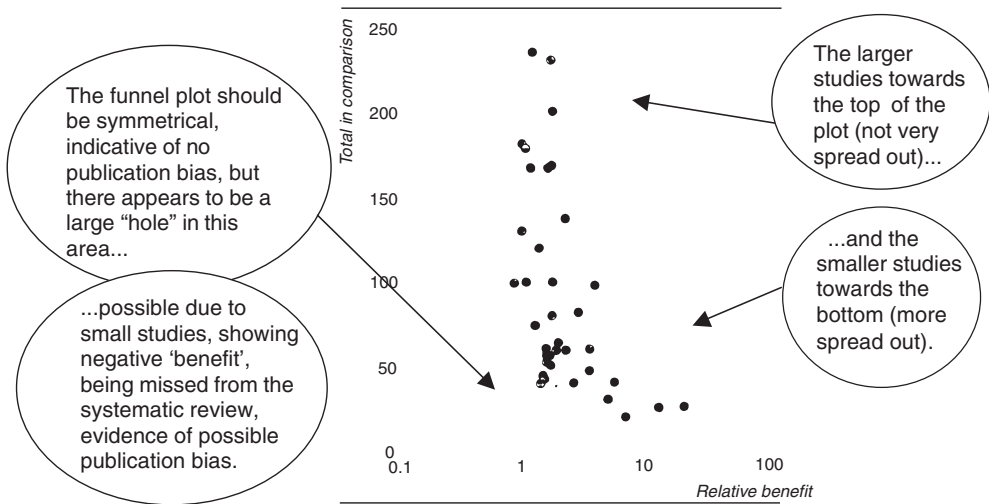


Figure 20.2 Funnel plot used to check for publication bias in a systematic review of the effectiveness of topically applied non-steroidal anti-inflammatory drugs. The asymmetry of the funnel is an indication of publication bias (see text). Reproduced from *BMJ*, Jan 1998; **316**: 333–338, courtesy of BMJ Publishing Group

As an example, Figure 20.2 is a funnel plot from a systematic review of the effectiveness of topically applied non-steroidal anti-inflammatory drugs in acute and chronic pain conditions (Moore *et al.* 1998). Relative benefit (risk ratio) is shown on the horizontal axis. Each point in the figure represents one of the studies. Values to the left of the value of 1 on the horizontal axis show negative ‘benefit’, values to the right, positive benefit.

The asymmetry in the funnel is quite marked, with a noticeable absence of small studies showing negative ‘benefit’ (risk ratio less than 1). The authors comment:

The funnel plot might be interpreted as showing publication bias. The tendency for smaller trials to produce a larger analgesic effect might be construed as supporting the absence of trials showing no difference between topical non-steroidal and placebo. We made strenuous efforts to unearth unpublished data and contacted all pharmaceutical companies in the United Kingdom that we identified as producing non-steroidal products. One company made unpublished data available to us, but the others did not feel able to do so.

Exercise 20.3 a) Outline the major sources of publication bias. (b) Figure 20.3 shows a funnel plot from a systematic review of trials of beta blockers in secondary prevention after myocardial infarction (Egger and Davey Smith 1998). The plot has odds ratio (horizontal axis) against sample size. Comment on the evidence for publication bias.

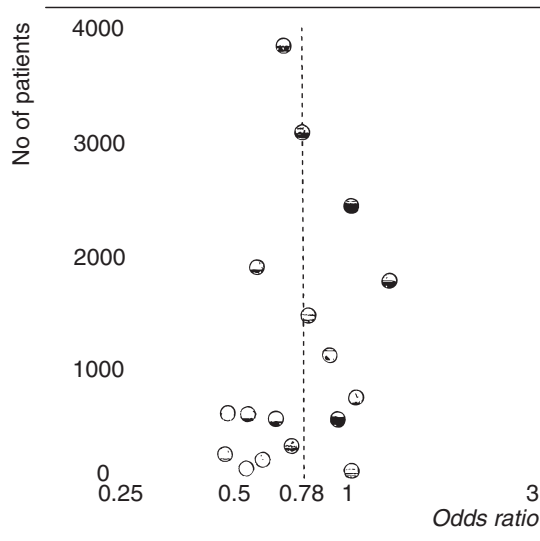


Figure 20.3 Funnel plot from a systematic review of trials of beta blockers in secondary prevention after myocardial infarction. Reproduced from *BMJ* 1998, **316**: 61–6. Figure 2, p. 64, courtesy of BMJ Publishing Group

Combining the studies

Meta-analysis is the process of combining a number of separate studies to produce one ‘super-study’. So, for example, we might have three separate studies, with sample sizes of 40, 80 and 150. When combined, we get a super-study with a sample size of 270. The assumption of the meta-analysis is that this super-study will provide a more reliable and precise overall result for the output variable in question, than do any of the smaller individual studies. We can use the *Mantel-Haenszel* procedure to combine the studies.² Before studies can be combined, however, they must satisfy the *homogeneity* criterion. A few words about that first, before we look at an example of meta-analysis.

Homogeneity among studies

Even when a set of potentially similar studies has been identified, authors have to make sure they are similar, or *homogeneous*, enough to be combined. For example, they should have similar subjects, have the same type and level of intervention, the same output measure, the same treatment effect and so on. Only if studies are *homogeneous* in this way can they be properly combined. Studies which don’t have this quality are said to suffer from *heterogeneity*. The underlying assumption (i.e. the null hypothesis) of meta-analysis is that all of the studies measure the same effect in the same population, and that any differences between them is due to chance alone. When the results are combined the chance element cancels out.

² Note that this is not to be confused with the Mantel-Haenszel test for heterogeneity.

You might find the comments on heterogeneity by the authors of the diet and cholesterol study quoted earlier illuminating (Tang, *et al.* 1998):

Heterogeneity between study effects

The design and results of these dietary studies differed greatly. They were conducted over 30 years and varied in their aims, in the intensity and type of intervention, and in the different baseline characteristics of the subjects included. Completeness and duration of follow up also differed. Unsurprisingly, the heterogeneity between their effects on blood cholesterol concentration was also significant. Among the longer trials some, but not all, of the heterogeneity between the effects on blood cholesterol concentration seemed to be due to the type of diet recommended. Deciding which trials should be included in which groups is open to different interpretation and, although we tried to be consistent, for some trials the target diets either were not clearly stated or did not fit neatly into recognised categories such as the step 1 and 2 diets. It is important to be cautious in interpreting meta-analysis when there is evidence of significant heterogeneity; although there was no evidence that the overall results were influenced by trials with outlying values.

The homogeneity assumption should be tested. One possibility is for the authors to provide readers with a *L'Abbé plot*. The *L'Abbé plot* displays *outcomes* from a number of studies, with the percentage of successes (or reduction in risk, etc.) with the treatment group on the vertical axis, and same measure for the control/placebo group on the horizontal axis. The 45° line is thus the boundary between effective and non-effective treatment. Values above the line show beneficial results. If possible, varying sized plotting points proportional to sample size should be shown. The more compact the plot, the more homogeneous the studies.

As an example, Figure 20.4 is a *L'Abbé plot* showing outcomes from 37 placebo-controlled trials of topical non-steroidal anti-inflammatory drugs in acute (●), and chronic (■), pain

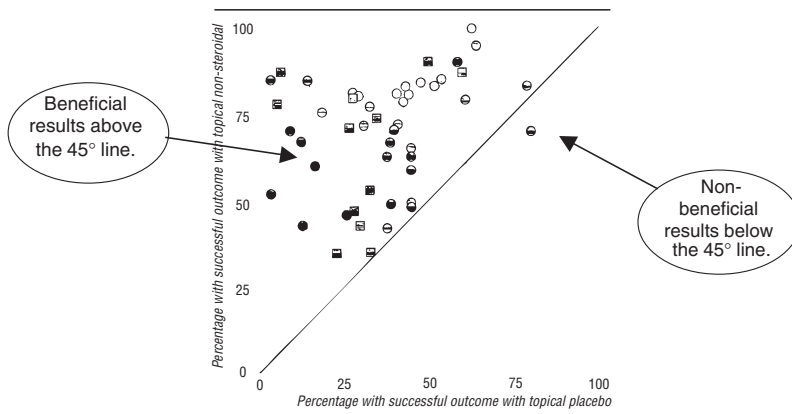


Figure 20.4 *L'Abbé plot* showing outcomes from 37 placebo-controlled trials of topical non-steroidal anti-inflammatory drugs in acute (●) and chronic (■) pain conditions. The compactness of the plotted points is a measure of homogeneity across the studies. Reproduced from *BMJ*, Jan 1998; **316**: 333–338, courtesy of BMJ Publishing Group

conditions (Moore *et al.* 1998). In this plot, the authors have not plotted the points in proportion to sample size. Whether the degree of spread of the points in Figure 20.4 is indicative of homogeneity among the studies is a matter of judgement, which can only be made by those experienced in the interpretation of these charts. Note that the *overall* meta-analytic result can also be plotted on this same plot (but is not shown in Figure 20.4).

Mantel-Haenszel test for heterogeneity

A more commonly used alternative is the *Mantel-Haenszel test for heterogeneity*, which uses the chi-squared distribution (see Chapter 14). The null hypothesis is that the studies are homogeneous. An example of its use is given in Table 20.2, which is taken from a study that ‘aimed to identify and evaluate all published randomised trials of hospital versus general practice care for people with diabetes’ (Griffin 1998). The author’s Table 2 presents a summary of the weighted (by sample size) mean differences, for a number of different outcomes. The author’s Table 3 presents similar information for different outcomes in terms of the odds ratio. The *p*-values for the Mantel-Haenszel test (using chi-squared) are given in the last column. Only one set of studies (Referral to chiropody, *p*-value < 0.005) displays evidence of heterogeneity, but since this comprised only two studies, the result is somewhat meaningless.

Meta-analysis and the Mantel-Haenszel procedure

If the studies pass the homogeneity test then we can combine them using the Mantel-Haenszel procedure, to produce the meta-analysis; this will give us an overall value for the outcome in question. The procedure is often accompanied by a forest plot, showing the individual studies, together with the combined result, as in the next example.

This is a report of a meta-analysis of randomised controlled trials to compare antibiotic with placebo, for acute cough in adults, referred to above (Fahey *et al.* 1998). The focus was on placebo-controlled trials, which reported two specific outcomes: the proportion of subjects reporting productive cough; and the proportion of subjects reporting no improvement at follow-up.³ Figure 20.5 shows the forest plots for these two acute cough outcomes, in terms of the risk ratios (called by the authors ‘relative risks’) in favour of the specific outcome.

The overall net outcome effect is shown with a diamond shape here (one for each of the two outcomes). The area of the diamond is proportional to the total number of studies represented, and the width the 95 per cent confidence interval. Values to the left of an odds ratio of 1 (bottom axis) show reductions in fatalities among cases, those to the right an increase in fatality (compared to control groups).

The Mantel-Haenszel procedure was used to produce the final result shown at the bottom of the forest plot in Figure 20.5. The aggregate relative risks are 0.85 for productive cough and 0.62 for no improvement at follow-up. These appear to show reductions in the risk for both conditions and favour the antibiotic over the placebo. However, since both have 95 per cent confidence intervals which include 1, neither is in fact significant, confirmed by the fact that

³ There was a third outcome concerned with side-effects which is not considered here. See Exercise 20.2 above.

Table 20.2 The Mantel-Haenszel test for heterogeneity across studies, with a number of different outcomes in the diabetes care study. The null hypothesis is that the studies are homogeneous. Only one outcome (chiroprody) has significant heterogeneity. Reproduced from *BMJ* 1998, **317**: 390–6. Table 2, p. 392, courtesy of BMJ Publishing Group

Outcome	Weighted difference in mean values (95% CI)			χ^2 test of between trial heterogeneity	
	Favours prompted GP care	Favours hospital care			P value
Glycated haemoglobin (%) (3 trials, n = 535)	-0.28 (-0.59 to 0.03)			3.90	>0.10
Systolic blood pressure (mm Hg) (2 trials, n = 369)		1.62 (-3.30 to 6.53)		2.56	>0.10
Diastolic blood pressure (mm Hg) (2 trials, n = 369)		0.56 (-1.69 to 2.80)		0.10	>0.75
Frequency of review (per patient per year) (2 trials, n = 402)	0.27 (0.07 to 0.46)			0.59	>0.30
Frequency of glycated haemoglobin test (per patient per year) (2 trials, n = 402)	1.60 (1.45 to 1.75)			0.05	>0.80

Outcome	Odds ratios (95% CI)			χ^2 test of between trial heterogeneity	
	Favours prompted GP care	Favours hospital care			P value
Mortality (2 trials, n = 455)		1.06 (0.53 to 2.11)		0.0	1.0
Losses to follow up (3 trials, n = 589)	0.37 (0.22 to 0.61)			1.63	>0.30
Referral to chiroprody (2 trials, n = 399)	2.51 (1.59 to 3.97)			9.77	<0.005
Referral to dietitian (2 trials, n = 399)		0.61 (0.40 to 0.92)		0.56	>0.30

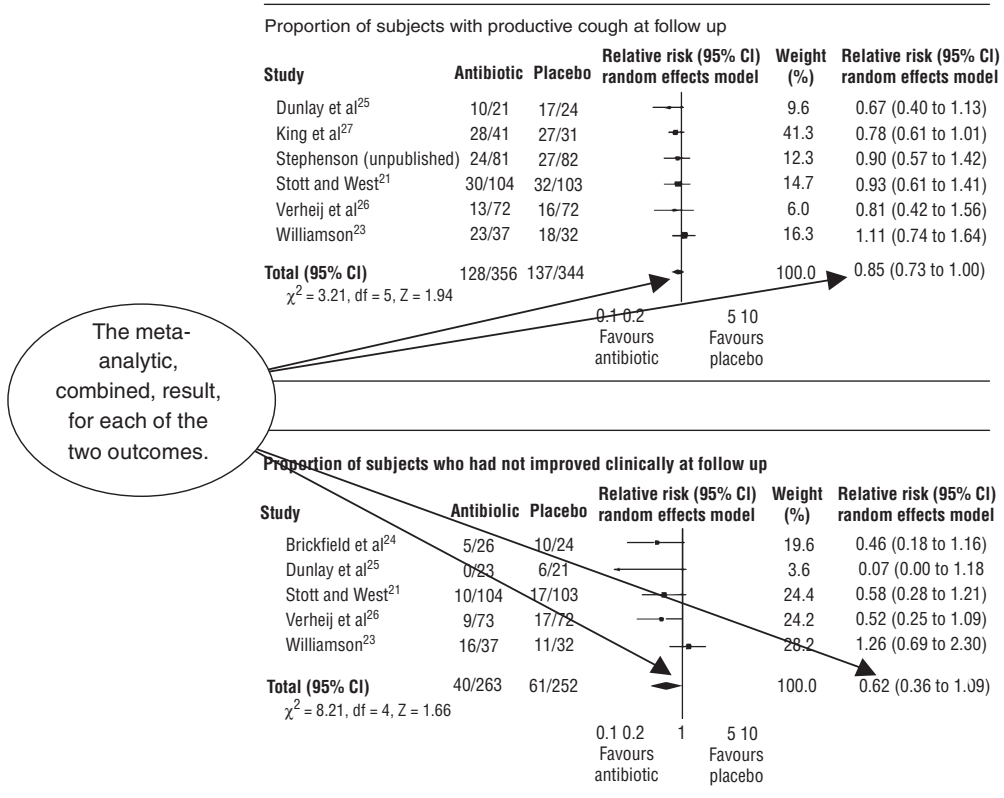


Figure 20.5 Forest plots showing relative risks (risk ratios) for two specific outcomes; *Productive cough*, and *No improvement at follow-up*, in a systematic review of antibiotic versus placebo for acute cough in adults. Reproduced from *BMJ* 1998, **316**: 906–10. Figure 4, p. 909. Figures 2 and 3, p. 908, courtesy of BMJ Publishing Group

both diamonds cross the line where relative risk = 1. In other words, the efficacy of antibiotic over placebo for acute cough in this population is not established by this meta-analysis.

However, look back at Figure 20.1, the forest plot for the dietary intervention and blood cholesterol meta-analysis. Here you will see at the bottom of the figure, the box representing the overall aggregated mean per cent reduction in cholesterol (labelled ‘Any diet’), which shows a reduction of 5.3 per cent. This box does not cross the per cent change line, so this is a statistically significant result, confirmed by the 95 per cent confidence interval of (4.7 per cent to 5.9 per cent).

Exercise 20.4 (a) Explain why homogeneity across studies is important before a meta-analysis is performed. (b) What methods are available for the detection of heterogeneity? (c) What advantage over the results of individual studies does a meta-analysis provide?

Appendix

Table of random numbers

23157	54859	01837	25993	76249	70886	95230	36744
05545	55043	10537	43508	90611	83744	10962	21343
14871	60350	32404	36223	50051	00322	11543	80834
38976	74951	94051	75853	78805	90194	32428	71695
97312	61718	99755	30870	94251	25841	54882	10513
11742	69381	44339	30872	32797	33118	22647	06850
43361	28859	11016	45623	93009	00499	43640	74036
93806	20478	38268	04491	55751	18932	58475	52571
49540	13181	08429	84187	69538	29661	77738	09527
36768	72633	37948	21569	41959	68670	45274	83880
07092	52392	24627	12067	06558	45344	67338	45320
43310	01081	44863	80307	52555	16148	89742	94647
61570	06360	06173	63775	63148	95123	35017	46993
31352	83799	10779	18941	31579	76448	62584	86919
57048	86526	27795	93692	90529	56546	35065	32254
09243	44200	68721	07137	30729	75756	09298	27650
97957	35018	40894	88329	52230	82521	22532	61587
93732	59570	43781	98885	56671	66826	95996	44569
72621	11225	00922	68264	35666	59434	71687	58167
61020	74418	45371	20794	95917	37866	99536	19378
97839	85474	33055	91718	45473	54144	22034	23000
89160	97192	22232	90637	35055	45489	88438	16361
25966	88220	62871	79265	02823	52862	84919	54883
81443	31719	05049	54806	74690	07567	65017	16543
11322	54931	42362	34386	08624	97687	46245	23245

Solutions to Exercises

Note: Although I have provided complete solutions to the calculating parts of the exercises, I have offered only brief comments where a commentary is required. This is deliberate, firstly because I don't want to write the book again in terms of the solutions and secondly tutors might want to tease these answers from the students themselves, perhaps as part of a wider discussion.

- 1.1 Ethnicity, sex, marital status, type of operation, smoking status, etc.
- 1.2 Apgar scale, Waterlow scale, Edinburgh Post-natal Depressions scale, Beck Depression Inventory, SF36, Apache, etc.
- 1.3 GCS produces ordinal data, which are not real numbers, so can't be added or divided.
- 1.4 Height, temp., cholesterol, body mass index, age, time, etc.
- 1.5 Number of deaths, number of angina attacks, number of operations performed, number of stillbirths, etc.
- 1.6 A continuous metric variable has an infinite or uncountable number of possible values. A discrete metric variable has a limited, countable number of possible values. (a) 7 (0, 1, 2, . . . , 6). (b) Not possible to do this, since number of possible weights is infinite.
- 1.7 VAS data is ordinal, because these are subjective judgements, which are not measured but assessed, and will probably vary from patient to patient and moment to moment. So it's not possible to calculate *average* if by this is meant adding up four values and dividing by four, because ordinal data are not real numbers.
- 1.8 Age, MC. Social class, O. No. of children, MD. Age at 1st child, MC. Age at menarche, MC. Menopausal state, O. Age at menopause, MC. Lifetime use of oral contraceptives, N. No. years taking oral contraceptives, MC. No. months breastfeeding, MC. Lifetime use of hrt, MC. Years of hrt, MC. Family history of ovarian cancer, N. Family history of breast cancer, N. Units of alcohol, MD. No. cigs per day, MD. Body mass index, MC. (key: N = nominal; O = ordinal; MD = metric discrete; MC = metric cont.).
- 1.9 Maternal age, MC, but given here in ordinal groups. Parity, MD. No. cigs daily, MD. Multiple pregnancy, N. Pre-eclampsia, N. Cesarean, N.

1.10 Age, MC. Sex, N. Number of rooms in home, MD. Length of hair, O. Colour of hair, N. Texture of hair, N. Pruritus, N. Excoriations, N. Live lice, O. Viable nits, O.

2.1

Cause of injury	Frequency (number of patients)	Relative frequency (% of patients)
Falls	46	61.33
Crush	20	26.67
Motor vehicle crash	6	8.00
Other	3	4.00

2.2

Satisfaction with nursing care	Frequency (number of patients)	Relative frequency (% of patients)
Very satisfied	121	25.5
Satisfied	161	33.9
Neutral	90	18.9
Dissatisfied	51	10.7
Very dissatisfied	52	10.9

2.3

% mortality	tally	Frequency
10.0–14.9	/// ////	9
15.0–19.9	/// ///	8
20.0–24.9	///	5
25.0–29.9	///	3
30.0–34.9	/	1

Observation: Most ICUs have percentage mortality under 20 per cent.

2.4

Parity	Frequency	% frequency
0	5	12.50
1	6	15.00
2	14	35.00
3	10	25.00
4	3	7.5
5	1	2.5
6	0	0
7	0	0
8	1	2.5

Most women have a parity between 1 and 3, with the largest percentage of women (35 per cent) having a parity of 1.

2.5 (a)

GCS score	Frequency (no. of patients)	Cumulative frequency (cumulative no. of patients)	Relative frequency (% of patients)	Cumulative relative frequency. (Cumulative % of patients)
3	10	10	6.49	6.49
4	5	15	3.25	9.74
5	6	21	3.90	13.64
6	2	23	1.30	14.94
7	12	35	7.79	22.73
8	15	50	9.74	32.47
9	18	68	11.69	44.16
10	14	82	9.09	53.25
11	15	97	9.74	62.99
12	21	118	13.64	76.63
13	13	131	8.44	85.07
14	17	148	11.04	96.11
15	6	154	3.90	100.00

(b) 53.25 per cent

2.6

(a) Better to have parity as the columns and diagnosis as the rows.

Diagnosis	Parity (no.)		totals
	≤ 2	> 2	
Benign	22	10	32
Malignant	4	4	8
totals	26	14	40

(b)

Diagnosis	Parity (%)	
	≤ 2	> 2
Benign	84.6	71.4
Malignant	15.4	28.6
totals	100.0	100.0

(c) Only 15.4 per cent of those with a parity of 2 or less had a malignant diagnosis, compared to nearly twice as many with a parity of 3 or more. Low levels of parity seem to favour a benign diagnosis.

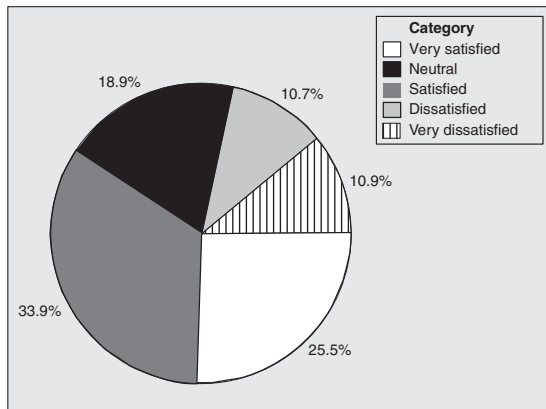
2.7

OCP	Cases (n = 106)	Controls (n = 226)
Yes	38	61
No	62	39
totals	100	100

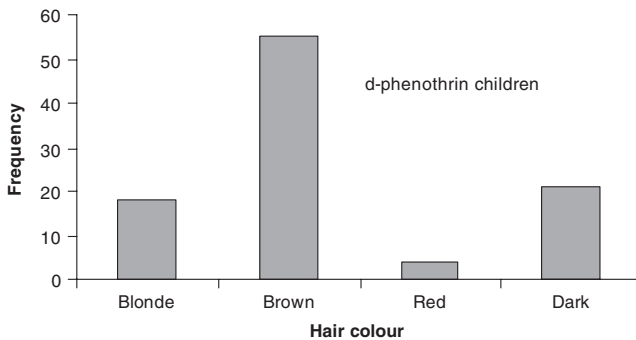
Comment: Only 38 per cent of those receiving a malignant diagnosis (the cases) had at some time used OCP, whereas 61 per cent of the controls (receiving a benign diagnosis), had used OCP. This suggests that a woman who had used OCP is more likely to receive a benign diagnosis. This is not a contingency table. There are two distinct groups of patients, those with a malignant diagnosis and those with a benign diagnosis.

3.1 Most common type of stroke is non-disabling large-artery in both groups. Second most common is disabling large artery in both groups.

3.2



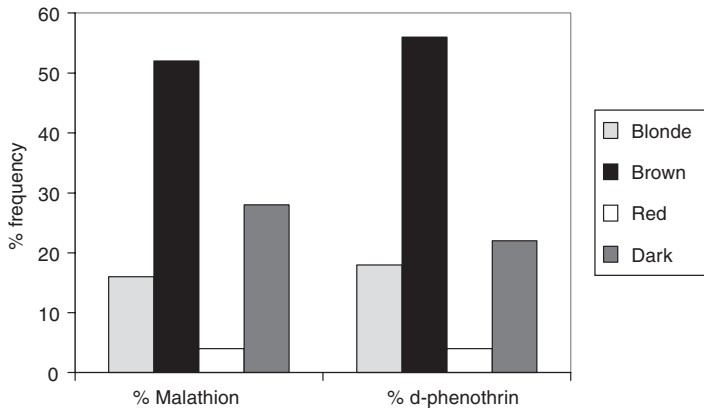
3.3



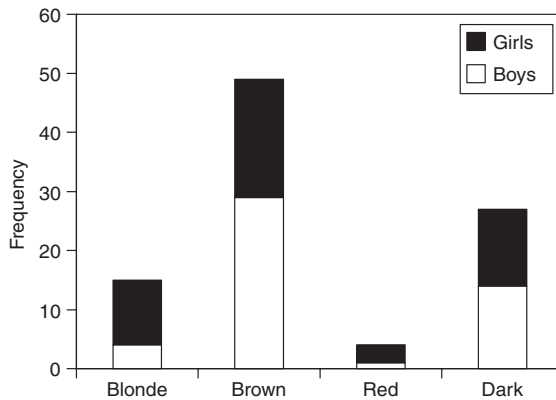
3.4



3.5



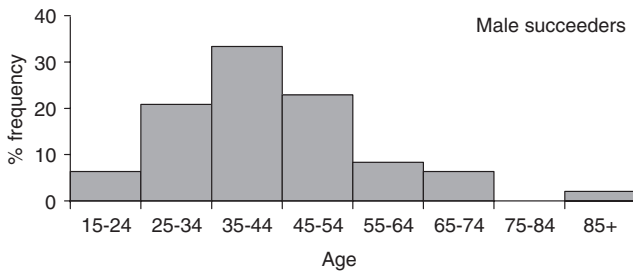
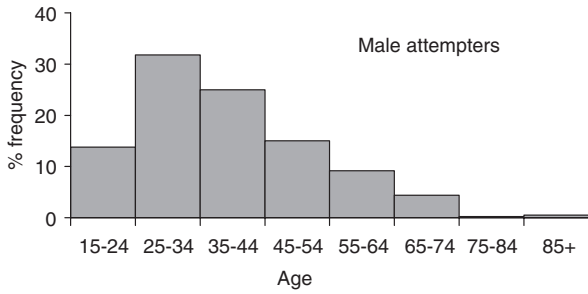
3.6 Stacked bar chart



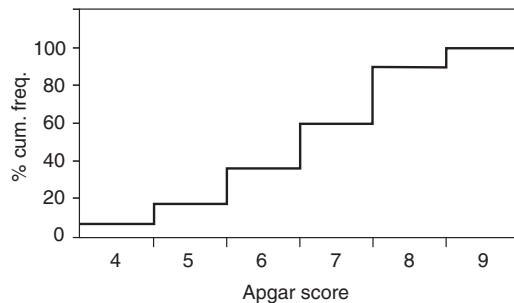
3.7 Schools have very few cases, most only one (20 schools). The majority of the rest have under 10 cases. One school exceptionally has 23 cases.

3.8 Most men have SP levels between four and four and a half, with progressively fewer and fewer men with less and more SP than this. There is a longish tail of higher values (up towards six).

3.9

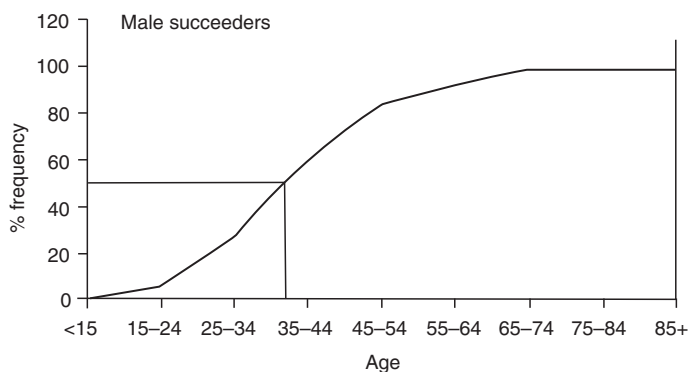
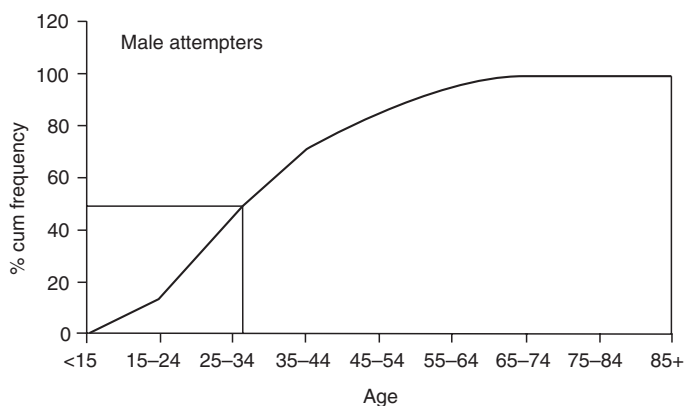


3.10



3.11 (a) In both groups minimum cholesterol levels are about 4 mmol/l, maximum levels about 11 mmol/l, but the control group showed slightly higher cholesterol levels throughout. About half the patients had a cholesterol level of 6 mmol/l and half more.

(b)



About 26 and 33

Comment: although this data is grouped we can see that half of the male attempters were younger than the youngest half of the male succeeders.

4.1 (a) highest is 70–79, (b) lowest is <15.

4.2 Less skewed.

4.3 (a) Negative. (b) The distribution is positively skewed, but only shows the lowest 95 per cent of values.

4.4 For attempters, the majority of both men and women are aged between 25 and 35. For succeeders, the majority are between 25 and 54. In all cases the distributions are positively skewed.

5.1 (a) Proportion breast fed = $67/149 = 0.4497$; percentage = $0.4497 \times 100 = 44.97\%$.

(b) Proportion bottle fed = $93/182 = 0.5110$; percentage = $0.5110 \times 100 = 51.10\%$.

5.2 (a) Prevalence of genital chlamydia = $(23/890) \times 100 = 2.58\%$.

(b) Incidence of SIDS per year = $10/10000$.

Incidence rate *per thousand* live births per year = $10/10 = 1$ SIDS death per 1000 live births per year.

5.3 (a) Cases and controls, modal class = II. (b) Satisfied. (c) PSF = 0.

5.4 Falls.

5.5 (a) Putting the percentage mortality values in ascending order gives:

11.2	12.8	13.5	13.6	13.7	14.0	14.3	14.7	14.9	15.2	16.1	16.3	17.7
1	2	3	4	5	6	7	8	9	10	11	12	13
18.2	18.9	19.3	19.3	20.2	20.4	21.1	22.4	22.8	26.7	27.2	29.4	31.3
14	15	16	17	18	19	20	21	22	23	24	25	26

Since there is an even number of values, the median percentage mortality is the average of the two 'middle' values, i.e. the average of the 13th (17.7) and 14th (18.2) values, i.e. the 13.5th value. The median is thus = $(17.7 + 18.2)/2 = 17.95\%$. Or you could have used the formula, median = $\frac{1}{2}(n + 1)$ th value; or $\frac{1}{2}(26 + 1) = \frac{1}{2} \times 27 = 13.5$ th value, as before.

(b) Attempters. (i) Men. 412 men. So median will be the average of the 206th and 207th values, which are in the 35–44 age group. (ii) Women. 562 women. So median is the average of the 281th and 282th values, which are in the 35–44 age group.

Succeeders.

(i) Men. 48 men, so median will be average of the 24th and 25th values, so the median must be in the 35–44 age group. (ii) Women. 55 women, so median is value of the middle, 28th, value, so the median must be in the 35–44 age group. You might want to repeat this exercise using the formula.

5.6 (a) mean > median; because of long tail of values to the right (positive skewness). (b) mean > median; positively skewed.

5.7 Mean percentage mortality = 18.66%, compared to median of 17.95%. These values are quite similar which suggests that the distribution might be reasonably symmetric (which you could check with a histogram).

5.8 (a) With outliers, mean = 720.4, median = 500, standard deviation = 622.2. (b) Without outliers, mean = 610.6, median = 500, standard deviation = 319.8.

5.9 Using 25th percentile is $\frac{1}{4}(n + 1)$ th value, then the 25th percentile = 14.23%. Using 75th percentile is $\frac{3}{4}(n + 1)$ th value, then the 75th percentile = 21.43%. So a quarter of the ICUs have a mortality of less than 14.23%, and a quarter have a mortality above 21.43%.

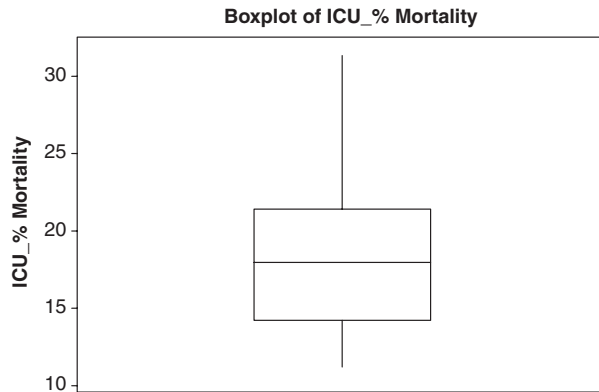
5.10 Breast fed, range = 20 to 28 years; bottle fed, range = 20 to 27 years.

5.11 Interquartile range of percentage mortality = (14.23 to 21.43)%. This means that the range of the middle half (50 per cent) of the ICU % mortality rates is from 14.23 per cent to 21.43 per cent.

5.12 Median (iqr) pain = 51 (23.8 to 87.8). The median pain level is 51 out of a maximum of 100, so 50 per cent of subjects had pain levels below 51 and half above 51. The interquartile range indicates that the middle 50 per cent of pain levels lay between 23.8 and 87.8.

5.13 Q2, the median = 6mmol/l; Q1 = 5.5mmol/l; Q3 = 7.0mmol/l; iqr = (5.5 to 7.0) mmol/l.

5.14



Seems to have a long positive tail and positively skewed (or outliers).

5.15 Median percentage DNA damage higher in the control group - about 12 compared to about eight in survivors. Interquartile range is also slightly larger. Max value much larger in controls (about 25 compared to 15). Minimums similar.

5.16 You can think of this, roughly speaking, as indicating that the average distance of all of these cord platelet count values is $69 \times 10^9/l$ from the mean value of $306 \times 10^9/l$.

5.17 $SD = 5.36\%$. With a mean of 18.66 per cent (Exercise 5.6), this suggests that the ICU's percentage mortality rates are, on average, 5.36 per cent away from this mean value.

5.18 For data to be Normally distributed, we need to be able to fit in three standard deviations below the mean (and three above it). In all cases it is impossible (by a long way!) to fit three SDs below the mean value without going into negative time. This would suggest that all the distributions are positively skewed.

6.1 The target population is the population at which the research is aimed; this is too large to be studied in any way. The study population is a more attainable but nonetheless still too large to be studied. The sample is a sample, representative of the study population. Consider trying to study the population of people in the UK who are HIV+. This is large population, perhaps many hundreds of thousands. It will be impossible to identify all, or even a reasonable proportion of this population. Many of these people will be transient; many will be undiagnosed. Many will refuse to participate in any research, etc.

6.2 The principal advantage is that a random sample will be representative of the population. The principal drawback is that a sampling frame is needed to take a random sample. Practically, sample frames for any realistic population are virtually impossible to obtain.

6.3 In an observational study, the investigators do not influence in any way the recruitment, treatment or aftercare of subjects, but may simply ask questions, take measurements, observe events and so on. In an experimental study, the investigator takes any active role in some aspect of the study, giving a drug, changing nursing care, etc.

6.4 A sample to determine levels of satisfaction with an endoscopy procedure. A sample to determine the prevalence of pressure sores in elderly patients in hospitals.

6.5 (a) Case-control studies are usually quicker, cheaper and better with rare conditions, than cohort studies. They don't suffer from subject fall-out over time. (b) Selection of suitable controls is often difficult. Problems with reliance on accuracy of patient recall, and medical records. Not good when exposure to risk factor is rare.

6.6 By double-blinding.

6.7 (a) To produce two groups of subjects who are as alike as possible. This will balance all factors, known and unknown, that might differentially affect the response of the two groups to the two treatments or treatment and placebo, and includes controlling for confounders. (b) Any solution to this problem will, of course, depend on the particular set of random numbers used. My random numbers were: 2 3 1 5 (7) 5 4 (8) 5 (9) (0) 1 (8) 3 (7) 2. Since we only have six blocks we can't use the random numbers in parentheses. With blocks of four:

Block 1, CCTT; Block 2, CTCT; Block 3, CTTC;
Block 4, TCTC; Block 5, TCCT; Block 6, TTCC

The first number is 2, so the first four subjects are allocated as block 2: C, T, C and T. The next number is 3, so the next four subjects are allocated: C, T, T and C. Continue this procedure until there are 20 in each group.

6.8 (a) The authors used a cross-section study of schoolchildren who were given a skin-prick test of sensitivity to six common allergens (the outcome variable), to determine atopic status, complimented by a questionnaire completed by parents to elicit pertinent socio-economic factors (including number of siblings). Possible confounders identified by the researchers were family history of atopy, sex, socio-economic status, presence of pets, smoking, and age.

(b) The researchers used a double-blind RCT, with patients (aged 2–15 years) randomised to either CF or PM. To quote, 'A double-dummy techniques was used: patients randomly assigned to CF also received a placebo PM, and patients randomly assigned to PM also received a placebo CF. Drug allocation was determined by a computer-generated list of random numbers.' The clinical outcome variable was the presence or absence of persistent dysentery after three days, and acceptable stool quality¹ and no fever after five days. Confounding is not an issue in RCTs, since the randomisation process is supposed to produce two groups with identical characteristics.

(c) The researchers used a cohort design, following a group of 2185 pregnant women becoming pregnant and having a baby between August 1991 and May 1993. The women were divided into two groups, normotensive and hypertensive. The outcome variable was defined as a birthweight below the 10th decile of expected weight (values from reference tables). Potential confounders were parity, age, socio-economic status, ethnicity, weight and height, smoking status, and use of aspirin.

(d) The researchers used a case-control study, in which cases were women with Down syndrome children, and controls were women selected randomly, having children with no congenital

¹ Satisfying a number of criteria.

abnormalities. Controls were matched only on birth year. There were 10 controls for each case! Potential confounding factors were: maternal and paternal ages, marital status (married/unmarried), parity, alcohol consumption (yes/no), prior foetal loss, and race (white/non-white).

(e) The researchers describe their study design as a 'follow-up' study. They selected two groups of patients (and their relatives), one receiving home-based care in one part of a city, the other hospital-based care, in a different part of the city. The relatives were interviewed at 10 days, one month and one year, and given questionnaires to assess the burden they were experiencing, their satisfaction with the service, and the General Health Questionnaire. The patients were assessed after four days, and then weekly and given a number of psychiatric questionnaires (Present State Examination, Morningside Rehabilitation Scale). The results from these various questionnaires constituted the outcome measures.

(f) The researchers used a randomised cross-over design. The subjects were randomised to either the 'regular' treatment arm (two puffs of salbutamol four times daily) or the 'as needed' treatment arm (salbutamol used as needed), each arm lasting two weeks. Patients were asked to record their peak expiratory flow rate (PEFR) morning and evening before inhaler use, the number of asthma episodes, and the number of as-needed salbutamol puffs used for symptom relief.

(g) The researchers summarise their design as follows, 'All new clients referred for counselling by GPs were asked to complete a questionnaire before and after counselling'. This contained: three psychological scales to measure anxiety and depression, self-esteem, and quality of life; and questions on levels of satisfaction with the counselling service. GPs were also asked to complete a questionnaire on their level of satisfaction with the service. The prescribing of anxiolytic/hypnotic and anti-depressant drugs, and the number of referrals to other mental health services in practices with and without counsellors was compared.

7.1 (a) A population parameter is a defining characteristic of a population, for example the mean age of all men dying of lung cancer in England and Wales. The population parameter is unknown but can be estimated from a representative sample drawn from this population. (b) A sample will never have exactly the same characteristics as a population because there is always the possibility that those members of a population not included in the sample may in some way be different from those included. (c) Determining the parameters of a target population is the underlying objective, but in practice this may prove to be difficult if not impossible. The study population is the population that, in practice, can be sampled.

7.2 They may be more wealthy or poorer, or older, or ethnically more or less mixed, etc.

8.1 (a) (i) $p(\text{benign}) = 226/332 = 0.681$; (ii) $p(\text{malignant}) = 106/332 = 0.319$. Notice these two probabilities sum to 1. (b) $p(\text{postmenopausal}) = 200/332 = 0.602$; (c) $p(>3 \text{ children}) = 112/332 = 0.337$.

8.2 (a) $p(\text{age} < 30) = (0.355 + 0.206 + 0.043) = 0.604$. (b) $p(\text{age} > 29) = (0.248 + 0.148) = 0.396$.

8.3 (a) 0.99. (b) 0.165

8.4 (a) Men.

Dead	Alcohol consumption (beverages/week)		Totals
	<1	>69	
Yes	195	66	261
No	430	145	575
Totals	625	211	836

(i) Absolute risk of death if consuming <1 beverage per week = $195/625 = 0.312$. (ii) Absolute risk of death if consuming >69 beverages per week = $66/211 = 0.313$.

(b) Women.

Dead	Alcohol consumption (beverages/week)		Totals
	<1	>69	
Yes	394	1	395
No	2078	19	2097
Totals	2472	20	2492

(i) Absolute risk of death if consuming <1 beverage per week = $394/2472 = 0.159$. (ii) Absolute risk of death if consuming >69 beverages per week = $1/20 = 0.050$.

Interpretation of results. For men there is approximately the same absolute risk of death among those consuming <1 beverage per week and those consuming >69 beverages per week (0.312 versus 0.313). For women the absolute risk of death if consuming <1 beverage per week is about three times the absolute risk for those consuming >69 beverages per week (0.159 versus 0.050). This perhaps surprising result may be due to the very small numbers consuming >69 beverages per week, which makes the result very unreliable.

8.5 (a) Under 35.

Smoked	Down syndrome baby	
	Yes	No
Yes	112	1411
No	421	5214
Totals	533	6625

(i) The odds that a woman having a Down syndrome baby smoked = $112/421 = 0.2660$. (ii) The odds that a woman having a healthy baby smoked = $1411/5214 = 0.2706$.

(b) ≥ 35

Smoked	Down syndrome baby	
	Yes	No
Yes	15	108
No	186	611
Totals	201	719

(i) The odds that a woman having a Down syndrome baby, smoked = $15/186 = 0.0806$. (ii) The odds that a woman having a healthy baby, smoked = $108/611 = 0.1768$.

Interpretation of results. Among the under 35 mothers there is little difference in the odds for Down syndrome between smoking and non-smoking mothers (0.2660 versus 0.2706). Among mothers ≥ 35 , the odds for Down syndrome among smoking mothers is about a half the odds for non-smoking mothers (0.0806 versus 0.1768).

8.6 (a) $p = 0.0806/(1 + 0.806) = 0.0746$; (b) $p = 0.1768/(1 + 0.1768) = 0.1502$.

8.7 (a) Men: risk ratio of death among those drinking >69 beverages per week compared to those drinking <1 beverage per week = $0.313/0.312 = 1.003$. (b) Women: risk ratio = $0.050/0.159 = 0.314$.

Interpretation of results. For men a risk ratio very close to 1 implies that there is no increased or decreased risk of death among those drinking <1 compared to those drinking >69 beverages per week. For women, the risk of death among the heavy drinkers appears to be only about a third the risk for light (or none) drinkers. But small numbers in the sample are not reliable.

8.8 (a) Mothers <35 . Odds ratio for a woman with a Down syndrome baby having smoked, compared to a woman with a healthy baby = $0.2660/0.2706 = 0.9830$. (b) Mothers ≥ 35 . Odds ratio = $0.0806/0.1768 = 0.4558$.

Interpretation of results. In younger mothers, the odds ratio close to 1 (0.9830) implies that smoking neither increases nor decreases the odds for Down syndrome. In older mothers, the odds ratio of 0.4558, implies that mothers who smoked during pregnancy have under half the odds of having a Down syndrome baby compared to non-smoking mothers.

8.9

Death from CHD	Periodontitis		Totals
	Yes	No	
Yes	151	92	243
No	1635	3450	5085
Totals	1786	3542	5328

Absolute risk of dying from CHD with periodontitis = $151/1786 = 0.084$. Absolute risk of dying from CHD with no dental disease = $92/3542 = 0.026$. So risk reduction = $0.084 - 0.026 = 0.058$. Therefore NNT = $1/0.058 = 17.2$, i.e. 18 people.

9.1 The smaller the s.e. of the sample mean, the more precise the estimate of the population mean. In this case the sample mean vitamin E intake of 6.30 mg (non-cases), has a s.e. of 0.05 mg, so we can be 95 per cent confident that the *population* mean vitamin E intake (non-cases) is no further than two s.e.s from this mean, i.e. within ± 0.10 mg. The largest s.e., 5.06 mg, and therefore the least precise estimate of the population mean, is that for vitamin C (cases).

9.2 (a) Cases. Sample mean age = 61.6 y, sample s.d. = 10.9 y, $n = 106$. Thus $\text{s.e.}(\bar{x}) = 10.9/\sqrt{106} = 1.059$. The 95 per cent confidence interval is therefore: $(61.6 \pm 2 \times 1.059)$, or (59.582 to 63.718) years. (b) Controls. Sample mean age = 51.0 y, sample s.d. = 8.5 y, $n = 226$. Thus $\text{s.e.}(\bar{x}) = 8.5/\sqrt{226} = 0.565$. The 95 per cent confidence interval is therefore: $(51.0 \pm 2 \times 0.565)$, or (49.870 to 52.13) years. The fact that the two CIs don't overlap means that we can be 95 per cent confident that the two population mean ages are significantly different.

9.3 For the integrated care group, over 12 months the sample mean number of admissions is 0.15. The 95 per cent confidence interval means we can be 95 per cent confident that the interval from 0.11 to 0.19 will contain the population mean number of visits for the population of which this is a representative sample. For the conventional care group the sample mean number of visits is lower, 0.11, and the 95 per cent confidence interval means we can be 95 per cent confident that the interval from 0.08 to 0.15 will contain the population mean number of visits.

$$\mathbf{9.4} \quad p = 0.290, \text{ s.e.}(p) = \sqrt{\frac{0.29(1 - 0.29)}{226}} = 0.030. \text{ 95 \% CI is:}$$

$$(0.290 - 2 \times 0.030 \text{ to } 0.290 + 2 \times 0.030) = (0.230 \text{ to } 0.350)$$

So we can be 95 per cent confident that the interval from 0.230 to 0.350 (or 23.0 to 35.0 per cent), will contain the population proportion of women who are pre-menopausal.

9.5 For all three time periods the median differences in pain levels are reasonably similar (38, 31 and 35), as are the 95 per cent confidence intervals, which all overlap, indicating no statistically significant difference between the two groups at any time period.

10.1 Three of the confidence intervals include zero, so there is no statistically significant difference in population mean infant weights between non-smoking and smoking mothers. The confidence interval for the difference in the mean weight of non-smoking mothers and mothers smoking 1–9 cigarettes per day, (–118 to –10) g, for boys, does *not* include zero, so this difference in population mean weights is statistically significant.

10.2 That for the radius, which has the narrowest confidence interval.

10.3 Because overlapping confidence intervals imply that the difference is not statistically significant.

10.4 The difference in sample median alcohol intakes is 5.4 g. The 95 per cent confidence interval of (1.2 to 9.9) g, does not include zero, so we can be 95 per cent confident that the population difference in median alcohol intake is statistically significant and lies somewhere between 1.2 g and 9.9 g.

11.1 For *gingivitis*, the confidence intervals for both CHD and *mortality* contain 1, so difference in risk compared to no disease is not statistically significant. For *periodontitis* neither confidence interval includes 1, so the difference in risk is statistically significant. For *no teeth*, the confidence interval for CHD includes 1, so not statistically significant, but for *mortality*,

the confidence interval does not include 1, so the difference in risk compared to no disease is statistically significant.

11.2 (a) Age and sex are notorious as confounders of many other variables, and adjustment for them is nearly always advisable. (b) With no exercise taken as the referent state, the odds ratio for all three age groups are less than 1, suggesting perhaps that exercise at any age reduces the odds for a stroke. However, only exercise taken between 15 and 40 has a statistically significant effect, since the confidence interval for the 40–55 year-old group, (0.3 to 1.5), includes 1. Note, by the way, that a 25-year-old and a 40-year-old individual could each be allocated to either one of two groups. The groups are not well defined.

11.3 The following risk factors are statistically significant for increasing the risk of thromboembolic events: being aged ≤ 19 ; having any parity other than 1; smoking ≥ 10 cigarettes per day; having multiple pregnancy; having pre-eclampsia; having a cesarean. The latter two appear to increase the risk the most.

12.1 (a) Is the proportion of women using the clinic same as proportion of men, i.e. 0.5? (b) $H_0: \pi = 0.5$ (π is population proportion of women using clinic). (c) Yes, reject because the p -value is less than 0.05. The proportion of women is *not* 0.5, i.e. not the same as men. (d) No, don't reject because the p -value is *not* less than 0.05. The proportion of women using the clinic is the same as men.

12.2 Since both p -values (0.25 and 0.32) exceed 0.05, then there is no statistically significant difference in the two means.

12.3 Mean age, mean age at menopause, and mean body mass index are statistically significant, since their p -values are all less than 0.05. The other four variables show no statistically significant difference since their p -values are all greater than 0.05.

12.4 (a) A false positive is when the null hypothesis is rejected when it shouldn't have been, because it is true, i.e. an effect is detected when there isn't one. (b) A false negative is when the null hypothesis is not rejected when it should have been, because it is false, i.e. a real effect is not detected.

12.5 (a) We want to minimise the probability of a type I error, i.e. a false positive. For example, we might have a test, the results of which, if positive, will lead to an unnecessary intrusive intervention. (b) Because if α is made very small, β would become unacceptably large because of the trade-off between the two measures.

12.6 (a) (i) $n = (2 \times 12^2/10^2) \times 10.5 = 31$; (ii) $n = (2 \times 12^2/10^2) \times 14.9 = 43$; (iii) $n = (2 \times 12^2/10^2) \times 11.7 = 34$. (b) (i) $n = [(0.4 \times 0.6 + 0.20 \times 0.80)/0.20^2] \times 10.5 = 105$; (ii) $n = [(0.4 \times 0.6 + 0.20 \times 0.80)/0.20^2] \times 14.9 = 149$; (iii) $n = [(0.4 \times 0.6 + 0.20 \times 0.80)/0.20^2] \times 11.7 = 117$.

12.7 $P_a = 0.70$. $P_b = 0.80$, so $(P_b - P_a) = -0.10$. Therefore, (a) $n = [(0.70 \times 0.30 + 0.80 \times 0.20)/-0.10^2] \times 7.8 = 289$; (b) $n = [(0.70 \times 0.30 + 0.80 \times 0.20)/-0.10^2] \times 14.9 = 551$.

13.1 There are only two statistically significant risk factors, both of which show higher risks for the alteplase patients (i.e. $rr < 1$); CAPG, $rr = 0.884$, p -value = 0.049, see table footnote for meaning of CAPG; and a Killip classification > 1 ; $rr = 0.991$, p -value = 0.026. Anaphylaxis is a complication which is almost statistically significant ($rr = 0.376$, p -value = 0.052, and we might want to consider it so).

13.2 In the model with the seven variables shown, all are statistically significant except passive smoking from husband, and at work. With only the first five variables included, plus passive smoking from husband and/or at work, makes this last variable statistically significant (p -value = 0.049).

14.1 Expected values:

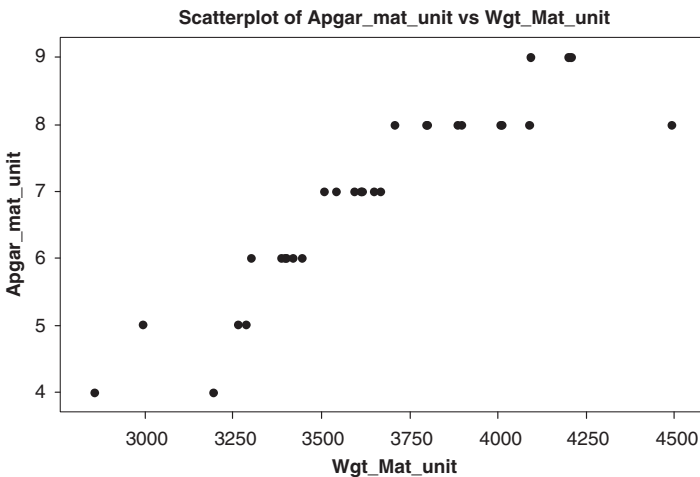
		Apgar < 7		
		Yes	No	Totals
Mother smoked	Yes	3.667	6.333	10
	No	7.333	12.667	20
	Totals	11	19	30

14.2 The test statistic = $\sqrt{\{(8 - 3.667)^2/3.667 + (3 - 7.333)^2/7.33 + (2 - 6.333)^2/6.33 + (17 - 12.67)^2/12.667\}} = \sqrt{12.109} = 3.480$. Since we have a 2×2 table, then we are in the first row of Table 14.3, because $(2 - 1) \times (2 - 1) = 1 \times 1 = 1$, and the critical chi-squared value which must be exceeded to reject the null hypothesis is 3.85. The test statistic value of 3.480 does not exceed this value, so the evidence is *not* strong enough for us to reject the null hypothesis of equal proportions of smokers in both Apgar groups.

The null hypothesis of equal proportions is equivalent to a null hypothesis of independent variables. Since we have rejected the former we have also rejected the latter, so these variables are independent.

14.3 (i) No trend across categories of social class, p -value = 0.094; (ii) statistically significant trend across the two categories (yes/no) of oral contraceptive use, p -value = 0.000; (iii) no trend across categories of alcohol consumption, p -value = 0.927; (iv) no trend across categories of cigarette consumption, p -value = 0.383.

15.1



Association seems to be strong and positive.

15.2 The association seems to be strong and positive.

15.3 The association appears to be strong and positive, but does not appear to be linear.

15.4 (a) All are statistically significant. (b) 0.896 for mothers less than two years from birth date. (c) 0.632 for mothers where the baby concerned was \geq 3rd born.

16.1 Yes. No.

16.2 Contingency table:

		Observer 1		
		<16	\geq 16	Totals
Observer 2	<16	5	2	7
	\geq 16	0	9	9
Totals		5	11	16

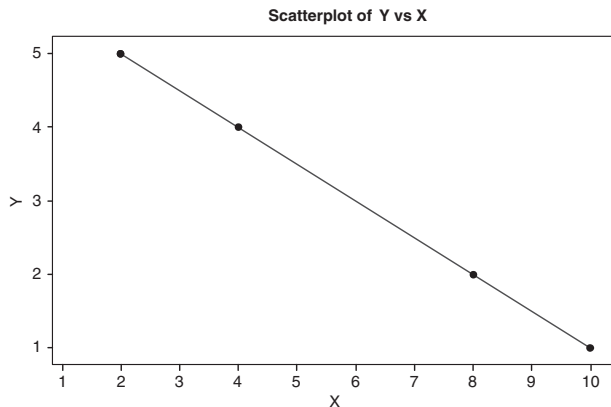
(a) Observed proportional agreement = $(5 + 9)/16 = 0.875$.

(b) Expected values are as follows:

		Observer 1	
		<16	\geq 16
Observer 2	<16	2.19	4.81
	\geq 16	2.81	6.19

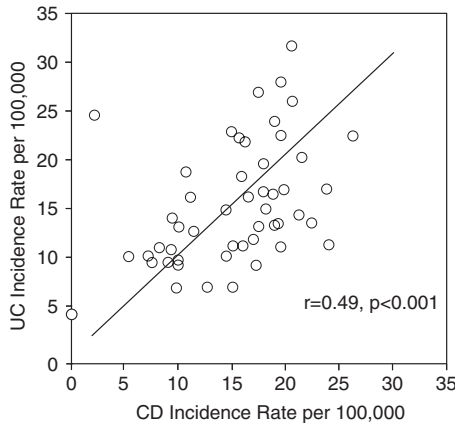
Expected agreement = $(2.19 + 6.19)/16 = 0.523$. So kappa = $(0.875 - 0.523)/(1 - 0.523) = 0.738$. From Table 16.3, chance adjusted agreement is very good.

17.1 Scatterplot.



Equation is: $Y = 6.0 - 0.5X$

17.2 (a) best straight line by eye:



Equation is: $UC = 1 + 0.85 \times CD$. By $+0.85$.

(b) $\%M = 46.886 - 0.620E$. A decrease is $\%$ mortality of 0.620 $\%$. (c) $\%$ exposed at work = $12 + 0.92 \times \%$ current smokers. 22 $\%$.

17.3 mean BMI = 41.902 kg/m².

17.4 (a) All p -values < 0.05 so all statistically significant. (b) Will decrease bmi by 0.025 for each 1 year increase. (c) Adjusted R^2 was 0.635, now 0.638, so marginal improvement. (i) 18.42; (ii) 10.95.

17.5

Subject	Age	D_1	D_2
1	50	1	0
2	55	0	0
3	35	0	1

17.6 (a) Severity of disability; mental disorders; respiratory system disorders; numbers of residents in private residential homes (all p -values < 0.05). (b) (i) natural log of utilisation time increases by 0.006, or 1.006 minutes (taking antilog). (ii) increase of 0.043 in natural log, or 1.044 minutes. (c) About 11 per cent (see R^2 in table footnote).

17.7 (a) Age; age squared; family history of hypertension; calcium intake. (b) We can be 95 per cent confident that the population regression parameter on age is between 0.28 and 0.64. (c) The blood lead model (largest age coefficient value).

17.8 See text.

18.1 Using the formula, odds = probability/(1 - probability) from Chapter 8. When $P(Y = 1) = 0.4286$ when OCP = 0, then odds = 0.7501. When $P(Y = 1) = 0.2247$, when OCP = 1, odds = 0.2898. The odds ratio = $0.2898/0.7501 = 0.386$.

18.2 (a) Because there are only two values for the dependent variable. It would be better to group the variables first and plot proportions in each group. (b) Yes, the confidence interval for odds ratio of (1.08 to 1.14) does not include 1. (c) $P(Y = 1) = e^{(-6.4672 + 0.10231 \times \text{age})}$.

(d)(i) 0.1343. (ii) 0.2707. (e) 0.8657, 0.7299. Odds ratio = 0.4182. A woman aged 45 has only about 41 per cent the odds of a malignant diagnosis as a woman aged 50. (f) The antilog_e of 0.10231 equals 1.108 (rounded to 1.11 by Minitab). (g) $10 \times 0.10231 = 1.0231$. antilog_e of 1.0231 = 2.78. In other words an increase in age of 10 years increases the odds ratio by 2.78.

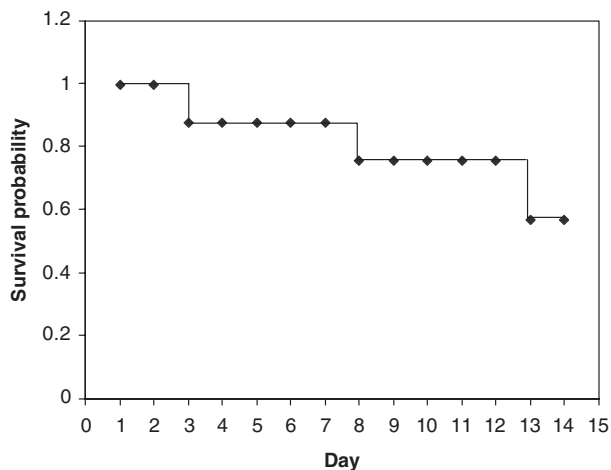
18.3 BMI is statistically significant since the p -value is < 0.05 and confidence interval does not include 1.

18.4 OCP is not statistically significant; p -value 0.278 is > 0.05 ; and confidence interval includes 1. Age and BMI both statistically significant; p -values are < 0.05 and neither confidence interval includes 1.

18.5 The null hypothesis is that the goodness-of-fit is good. The p -value here is 0.958, which is not less than 0.05, so we cannot reject the null hypothesis and conclude that the fit is good.

19.1

1 Day	2 Number still in study at start of day t	3 Withdrawn prematurely up to day t	4 Deaths in day t	5 Number at risk in day t	6 Probability of death in day t	7 Probability of surviving day t	8 Cumulative probability of surviving to day t
t	n	w	d	r	d/r	$p = 1 - d/r$	S
3	8	0	1	8	$1/8 = 0.125$	0.875	0.875
8	7	0	1	6	$1/6 = 0.167$	0.833	0.758
13	6	1	1	4	$1/4 = 0.25$	0.75	0.569



19.2 Raltitrexed; about 5 months. Lokich; about $5\frac{1}{2}$ months. de Gramont; about 6 months.

19.3 Since the p -value is < 0.05 , then a null hypothesis of no difference in survival times can be rejected.

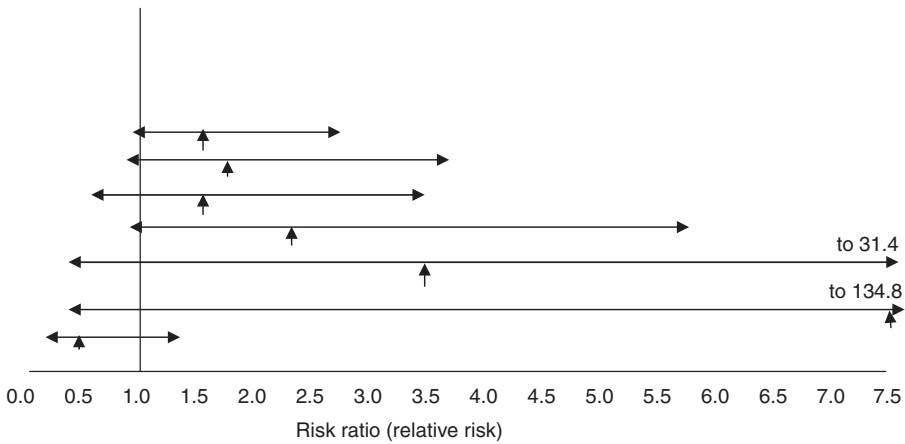
19.4 The log-rank test p -value is 0.03. Since this is < 0.05 we can assume that there is a statistically significant difference between the treatments. The combination seems to work best since it shows the lowest percentage treatment failure.

19.5 All confidence intervals include the value 1 so none are statistically significant.

19.6 In the multivariate (adjusted) results, the first five are all statistically significant since none of the confidence intervals includes 1. This is the same as for the five univariate analyses. The last, cigarette smoking at enrolment, is not statistically significant since this confidence interval does include 1; which is also not statistically significant in the univariate analysis. None of the other variables are statistically significant in the univariate analyses.

20.1 See the text.

20.2 Risk ratio (relative risk) shown by \blacktriangle . Size of sample not indicated in this figure.



20.3 (a) See the list in the section headed 'Publication and other biases'. (b) On the question of publication bias and this funnel plot the authors comment, 'Visual assessment shows some asymmetry, which indicates that there was selective non-publication of smaller trials with less sizeable benefits. However, in formal statistical analysis the degree of asymmetry is found to be small and non-significant. Bias does not therefore seem to have distorted the findings from this meta-analysis.'

20.4 (a) If studies are not similar in objective, in outcome measure, in design, have similar subjects and so on then it is not sensible to combine them. (b) L'Abbé plots; Mantel-Haenszel test; chi-squared test. (c) A larger combined sample is likely to be more reliable (precise) than a number of smaller samples.

References

- Altman, D.G. (1991) *Practical Statistics for Medical Research*. London: Chapman & Hall.
- ASSENT-2 (Assessment of the Safety and Efficacy of a New Thrombolytic) Investigators. (1999) *Lancet*, **354**, 716–21.
- Blanchard, J.F., Bernstein, C.N., Wajda, A. and Rawsthorne, P. (2001) Small-area variations and sociodemographic correlates for the incidence of Crohn's disease and ulcerative colitis. *Amer. J. Epid.*, **154**, 328–33.
- Bland, J.M. and Altman, D.G. Statistical methods for assessing agreement between two clinical measurements. *Lancet*, **I**, 307–10.
- Bland, M. (1995) *An Introduction to Medical Statistics*. Oxford: Oxford University Press.
- Brueren, M.M., Schouten, H.J.A., de Leeuw, P.W., van Montfrans, G.A. and van Ree JW. (1998) A series of self-measurements by the patient is a reliable alternative to ambulatory blood pressure measurement. *Brit. J. General Practice*, **48**, 1585–9.
- Chapman, K.R., Kesten, S. and Szalai, J.P. (1994) Regular vs as-needed inhaled salbutamol in asthma control. *Lancet*, **343**, 1379–83.
- Cheng, Y., Schartz, J., Sparrow, D. *et al.* (2001) Bone lead and blood lead levels in relation to baseline blood pressure and the prospective development of hypertension. *Amer. J. Epid.*, **153**, 164–71.
- Chi-Ling, C., Gilbert, T.J. and Daling, J.R. (1999) Maternal smoking and Down syndrome: the confounding effect of maternal age. *Amer. J. Epid.*, **149**, 442–6.
- Chosidow, O., Chastang, C., Brue, C. *et al.* (1994) Controlled study of Malathion and *d*-phenothrin lotions for *Pediculus humanus* var *capitis*-infested schoolchildren. *Lancet*, **344**, 1724–6.
- Conley, L.J., Ellerbrock, T.V., Bush, T.J. *et al.* (2002) HIV-1 infection and risk of vulvovaginal and perianal condylomata acuminata and intraepithelial neoplasia: a prospective cohort study. *Lancet*, **359**, 108–14.
- Conter, V., Cortinovis, I., Rogari, P. and Riva, L. (1995) Weight growth in infants born to mothers who smoked during pregnancy. *BMJ*, **310**, 768–71.
- DeStafano, F., Anda, R.F., Kahn, H.S., Williamson, D.F. and Russell, C.M. (1993) Dental disease and risk of coronary heart disease and mortality. *BMJ*, **306**, 688–91.
- Dunne, M.W., Bozzette, S., McCutchan, J.A. *et al.* (1999) Kemper Class activity, Havlir D, for the California Collaborative Treatment Group. Efficacy of Azithromycin in prevention of *Pneumocystis carinii* pneumonia: a randomised trial. *Lancet*, **354**, 891–5
- Egger, M. and Davey Smith, G. (1998) Bias in location and selection of studies. *BMJ*, **316**, 61–6.
- Fahey, T., Stocks, N. and Thomas, T. (1998) Quantitative systematic review of randomised controlled trials comparing antibiotic with placebo for acute cough in adults. *BMJ*, **316**, 906–10.

- Fall, C.H.D., Vijayakumar, M., Barker, D.J.P., Osmond, C. and Duggleby, S. (1995) Weight in infancy and prevalence of coronary heart disease in adult life. *BMJ*, **310**, 17–9.
- Field, A. (2000) *Discovering Statistics Using SPSS for Windows*. London: Sage.
- FRISC II (FRagmin and Fast Revascularisation during InStability in Coronary artery disease) Investigators. (1999) Long-term low-molecular-mass heparin in unstable coronary-artery disease: FRISC II prospective randomised multicentre study. *Lancet*, **354**, 701–7.
- Goel, V., Iron, K. and Williams, J.I. (1997) Enthusiasm or uncertainty: small area variations in the use of the mammography services in Ontario, Canada. *J. Epid. Comm. Health*, **51**, 378–82.
- Goldhaber, S.Z., Visani, L. and De Rosa, M. (1999) Acute pulmonary embolism: clinical outcomes in the International Cooperative Pulmonary Embolism Registry (ICOPER). *Lancet*, **353**, 1386–9.
- Grampian Asthma Study of Integrated Care. (1994) Integrated care for asthma: a clinical, social, and economic evaluation. *BMJ*, **308**, 559–64.
- Grandjean, P., Bjerve, K.S., Weihe, P. and Steuerwald, U. (2000) Birthweight in a fishing community: significance of essential fatty acids and marine food contaminants. *Int. J. Epid.*, **30**, 1272–7.
- Griffin, S. (1998) Diabetes care in general practice: a meta-analysis of randomised control trials. *BMJ*, **317**, 390–6.
- Gronbaek, M., Deis, A., Sorensen, T.I.A. *et al.* (1994) Influence of sex, age, body mass index, and smoking on alcohol intake and mortality. *BMJ*, **308**, 302–6.
- Grun, L., Tassano-Smith, J., Carder, C. *et al.* (1997) Comparison of two methods of screening for genital chlamydia infection in women attending in general practice: cross sectional survey. *BMJ*, **315**, 226–30.
- He, Y., Lam, T.H., Li, L.S. *et al.* (1994) Passive smoking at work as a risk factor for coronary heart disease in Chinese women who have never smoked. *BMJ*, **308**, 380–4.
- Hearn, J. and Higinson, I.J., on behalf of the Palliative Care Core Audit Project Advisory Group. (1998) Development and validation of a core outcome measure for palliative care: the palliative care outcome scale. *Quality in Health Care*, **8**, 219–27.
- Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic Regression Analysis*. Chichester: John Wiley & Sons, Ltd.
- Hu, F.B., Wand, B., Chen, C., Jin, Y., Yang, J., Stampfer, M.J. and Xu X. (2000) Body mass index and cardiovascular risk factors in a rural Chinese population. *Amer. J. Epid.*, **151**, 88–97.
- Imperial Cancer Fund OXCHECK Study Group. (1995) Effectiveness of health checks conducted by nurses in primary care: final results of the OXCHECK study. *BMJ*, **310**, 1099–104.
- Inzitari, D., Eliasziw, M., Gates, P. *et al.* (2000) The causes and risk of stroke in patients with asymptomatic internal-carotid-artery stenosis. *NEJM*, **342**, 1693–9.
- Janson, C., Chinn, S., Jarvis, D *et al.*, for the European Community Respiratory Health Survey. (2001) Effect of passive smoking on respiratory symptoms, bronchial responsiveness, lung function, and total serum IgE in the European Community Respiratory Health Survey: a cross-sectional study. *Lancet*, **358**, 2103–9.
- Kavanagh, S. and Knapp, M. (1998) The impact on general practitioners of the changing balance of care for elderly people living in an institution. *BMJ*, **317**, 322–7.
- Knaus, W.A., Draper, E.A., Wagner, D.P. and Zimmerman, J.E. (1985) APACHE II: A severity of disease classification system. *Critical Care Medicine*, **13**, 818–29.
- Lacy, A.M., Garcia-Valdecasas, J.C., Delgado, S. *et al.* (2002) Laparoscopy-assisted colectomy versus open colectomy for treatment of non-metastatic colon cancer: a randomised trial. *Lancet*, **359**, 2224–30.
- Ladwig, K.H., Roll, G., Breithardt, G., Budde, T., Borggrefe, M. (1994) Post-infarction depression and incomplete recovery 6 months after acute myocardial infarction. *Lancet*, **343**, 20–3.
- Leeson, C.P.M., Kattenhorn, J.E. and Lucas, A. (2001) Duration of breast feeding and arterial disability in early adult life: a population based study. *BMJ*, **322**, 643–7.

- Lindberg, G., Binge-fors, K., Ranstam, J. and Rastam, A.M. (1998) Use of calcium channel blockers and risk of suicide: ecological findings confirmed in population based cohort study. *BMJ*, **316**, 741–5.
- Lindelov, M., Hardy, R. and Rodgers, B. (1997) Development of a scale to measure symptoms of anxiety and depression in the general UK population: the psychiatric symptom frequency scale. *J. Epid. Comm. Health*, **51**, 549–57.
- Lindqvist, P., Dahlback, M.D. and Marsal, K. (1999) Thrombotic risk during pregnancy: a population study. *Obstetrics and Gynecology*, **94**, 595–9.
- Luke, A., Durazo-Arvizu, R., Rotimi, C. *et al.* (1997) Relations between body mass index and body fat in black population samples from Nigeria, Jamaica, and the United States. *Amer. J. Epid.*, **145**, 620–8.
- Machin, D., Campbell, M.J., Fayers, P.M. and Pinol, A.P.Y. (1987) *Sample Size Tables for Clinical Studies*. Oxford: Blackwell Scientific.
- Maughan, T.S., James, R.D., Kerr, D.J. *et al.*, for the British MRC Colorectal Cancer Working Party. (2002) *Lancet*, **359**, 1555–63.
- McCreadie, R., Macdonald, E., Blacklock, C. *et al.* (1998) Dietary intake of schizophrenic patients in Nithsdale, Scotland: case-control study. *BMJ*, **317**, 784–5.
- McKee, M. and Hunter, D. (1995) Mortality league tables: do they inform or mislead? *Quality in Health Care*, **4**, 5–12.
- Medical Research Council Advanced Bladder Working Party. (1999) Neoadjuvant cisplatin, methotrexate, and vinblastine chemotherapy for muscle-invasive bladder cancer: a randomised controlled trial. *Lancet*, **354**, 533–9.
- Michelson, D., Stratakis, C., Hill, L. *et al.* (1995) Bone mineral density in women with depression. *NEJM*, **335**, 1176–81.
- Moore, R.A., Tramer, M.R., Carroll, D., Wiffen, P.J. and McQuay, H.J. (1998) Quantitative systematic review of topically applied non-steroidal anti-inflammatory drugs. *BMJ*, **316**, 333–8.
- Morris, C.R., Kato, G.J., Poljakovic, M. *et al.* (2005) Dysregulated arginine metabolism, hemolysis-associated pulmonary hypertension, and mortality in sickle cell disease. *JAMA*, **294**, 81–91.
- Nikolajsen, L., Ilkjaer, S., Christensen, J.H., Kroner, K. and Jensen, T.S. (1997) Randomised trial of epidural bupivacaine and morphine in prevention of stump and phantom pain in lower-limb amputation. *Lancet*, **350**, 1353–7.
- Nordentoft, M., Breum, L., Munck, L.K., Nordestgaard, A.H. and Bjaeldager, P.A.L. (1993) High mortality by natural and unnatural causes: a 10 year follow up study of patients admitted to a poisoning treatment centre after suicide attempts. *BMJ*, **306**, 1637–41.
- Olson, J.E., Shu, X.O., Ross, J.A., Pendergrass, T. and Robison, L.L. (1997) Medical record validation of maternity reported birth characteristics and pregnancy-related events: A report from the Children's Cancer Group. *Amer. J. Epid.*, **145**, 58–67.
- Prevots, D.R., Watson, J.C., Redd, S.C. and Atkinson, W.A. (1997) Outbreak in highly vaccinated populations: implications for studies of vaccine performance. *Amer. J. Epid.*, **146**, 881–2.
- Protheroe, D., Turvey, K., Horgan, K. *et al.* (1999) Stressful life events and difficulties and onset of breast cancer: case-control study. *BMJ*, **319**, 1027–30.
- Rainer, T.H., Jacobs, P., Ng, Y.C. *et al.* (2000) Cost effectiveness analysis of intravenous ketorolac and morphine for treating pain after limb injury: double blind randomised controlled trial. *BMJ*, **321**, 1247–51.
- Relling, M.V., Rubnitz, J.E., Rivera, G.K. *et al.* (1999) High incidence of secondary brain tumours after radiotherapy and antimetabolites. *Lancet*, **354**, 34–9.
- Rodgers, M. and Miller, J.E. (1997) Adequacy of hormone replacement therapy for osteoporosis prevention assessed by serum oestradiol measurement, and the degree of association with menopausal symptoms. *Brit. J. General Practice*, **47**, 161–5.

- Rogers, A. and Pilgrim, D. (1991) Service users views of psychiatric nurses. *Brit J Nursing*, **3**, 16–7.
- Rowan, K.M., Kerr, J.H., Major, E. *et al.* (1993) Intensive Care Society's APACHE II study in Britain and Ireland – I: Variations in case mix of adult admissions to general intensive care units and impact on outcome. *BMJ*, **307**, 972–81.
- Sainio, S., Jarvenpaa, A.-L. and Kekomaki, R. (2000) Thrombocytopenia in term infants: a population-based study. *Obstetrics and Gynecology*, **95**, 441–4.
- Schrader, H., Stovner, L.J., Helde, G., Sand, T. and Bovin, G. (2001) Prophylactic treatment of migraine with angiotensin converting enzyme inhibitor (lisinopril): randomised, placebo-controlled, cross-over study. *BMJ*, **322**, 19–22.
- Shinton, R. and Sagar, G. Lifelong exercise and stroke. *BMJ*, **307**, 231–4.
- Staessen, J.A., Byttebier, G., Buntinx, F. *et al.* (1997) Antihypertensive treatment based on conventional or ambulatory blood pressure measurement. *JAMA*, **278**, 1065–72.
- Tang, J.L., Armitage, J.M., Lancaster, T. *et al.* (1998) Systematic review of dietary intervention trials to lower blood total cholesterol in free-living subjects. *BMJ*, **316**, 1213–20.
- Thomson, A.B., Campbell, A.J., Irvine, D.S. *et al.* (2002) Semen quality and spermatozoal DNA integrity in survivors of childhood cancer: a case-control study. *Lancet*, **360**, 361–6.
- Turnbull, D., Holmes, A., Shields, N., *et al.* (1996) Randomised, controlled trial of efficacy of midwife-managed care. *Lancet*, **348**, 213–219.
- van Es, R., Jonker, J.J., Verheugt, F.W.A., Deckers, J.W. and Grobbee, D.E., for the Antithrombotics in the Secondary Prevention of Events in Coronary Thrombosis-2 (ASPECT-2) Research Group. (2002) Aspirin and coumadin after acute coronary syndromes (the ASPECT-2 study): a randomised controlled trial. *Lancet*, **360**, 109–14.
- Wannamethee, S.G., Lever, A.F., Shaper, A.G. and Whincup, P.H. (1997) Serum potassium, cigarette smoking, and mortality in middle-aged men. *Amer. J. Epid.*, **145**, 598–607.
- Yong, L.-C., Brown, C.C., Schatzkin, A. *et al.* (1997) Intake of vitamins E, C, and A and risk of lung cancer. *Amer. J. Epid.*, **146**, 231–43.
- Zoltie, N. and de Dombal, F.T., on behalf of the Yorkshire Trauma Audit Group. (1993) The hit and miss of ISS and TRISS. *BMJ*, **307**, 906–9.

Index

- α *see* significance level
- absolute risk 100–1, 106–7
- absolute risk reduction (ARR) 106–7
- adjustment
 - confidence intervals 136, 137–8
 - confounders 81
 - goodness-of-fit 196–7, 203
 - hypothesis tests 158
- agreement 181–6
 - association 186
 - Bland-Altman charts 185–6
 - Cohen's kappa 182–4
 - continuous data 184–6
 - limits 185–6
 - ordinal data 184
 - weighted kappa 184
- analysis of variance (ANOVA) 209–11
- APACHE II scores 33–4
- Apgar scores 25–6, 38, 121, 128, 148, 174
- arithmetic mean *see* mean
- ARR *see* absolute risk reduction
- assessment bias 86
- association 171–80
 - agreement 186
 - confidence intervals 179
 - correlation coefficients 175–80, 183
 - definition 171–2
 - linear 172–3, 175
 - linear regression 190–1, 192
 - negative 172–3
 - non-linear 173, 175
 - p* values 176–9
 - positive 172–3
 - statistical significance 176–7
 - strength 174, 175–80
 - see also* scatterplots
- attitudes 77
- automated variable selection 200–2
- β *see* type II errors
- backwards elimination 202–3
- backwards selection 201
- bar charts 31–5, 41, 44–7
- beneficial risk factors 105
- bimodal distribution 47
- binary data 153, 214–15
- binomial distribution 48, 116
- Bland-Altman charts 185–6
- blinding 86
- block randomisation 85
- boxplots 41, 61–2, 63
- British Regional Heart Study 36
- case-control studies 11
 - association 177
 - confidence intervals 137
 - hypothesis tests 146, 153, 158
 - matched 81, 82, 102
 - odds ratios 105–6
 - probability 98
 - risk 102–3
 - study design 80–3
 - unmatched 81–2
- case-series studies 76
- categorical data 4–7, 10
 - agreement 184
 - association 180
 - charts 30–4, 37–8, 40–1

- categorical data (*Continued*)
 - confidence intervals 127, 131
 - frequency tables 18–20, 23–6
 - hypothesis tests 145, 151, 161–8
 - linear regression 199–200
 - numeric summary values 55, 57, 64
 - ordered 166–8
- causal relationships 77, 180, 190–1
- censored data 228
- chance-corrected proportional agreement statistic
 - 179, 182–4
- charts 29–41
 - bar charts 31–5, 41
 - boxplots 41, 61–2, 63
 - categorical data 30–4, 37–8, 40–1
 - continuous data 35–7, 40–1
 - cumulative data 37–41
 - discrete data 34–5, 37–8, 40–1
 - distribution 44–7
 - metric data 34–8, 40–1
 - nominal data 30–4, 41
 - ogives 38–40, 41, 60–1
 - ordinal data 30–4, 37–8, 40–1
 - pie charts 30–1, 41
 - step charts 37–8, 41
 - time series charts 40–1
 - see also* histograms
- chi-squared test
 - hypothesis tests 145, 151, 161–8
 - logistic regression 221, 222
 - survival 234
- clinical databases 240, 241
- clinical trials 84
- clustered bar charts 32–4
- coding design 199–200
- coefficient of determination 197
- Cohen's kappa 179, 182–4
- cohort studies
 - charts 36–7
 - probability 100–1, 102, 104, 106–7
 - study design 78–80, 83
 - survival 238
- colinearity 198
- comparative studies 13
- confidence intervals 111–18
 - agreement 183
 - association 179
 - difference between population parameters 119–31
 - hypothesis tests 117, 119–31, 149, 156, 163
 - independent populations 120–5, 127–31, 133–4
 - linear regression 195–6, 198
 - logistic regression 217
 - matched populations 125–6, 131
 - mean 112–16, 120–6, 134
 - median 117–18, 127–31
 - Minitab 120, 123, 128
 - Normal distribution 112–13, 120, 127
 - odds ratios 134–5, 137–8
 - proportions 116–17, 126–7
 - ratio of two population parameters 133–8
 - risk ratios 134–6
 - single population parameter 111–18
 - SPSS 120, 122–3
 - standard error 112–16
 - survival 237–8
 - systematic review 242–3
- confounders
 - linear regression 202–3, 204–5
 - logistic regression 222
 - risk ratios 136
 - study design 81, 84
- consecutive sampling 74–5
- constant coefficient 191
- contact sampling 74–5
- contingency tables 25–6
 - chi-squared test 162–4
 - logistic regression 221
 - risk 104
 - study design 79–80, 82
- continuous metric data 7–8, 10
 - agreement 184–6
 - association 176, 180
 - charts 35–7, 40–1
 - frequency tables 20–2
 - linear regression 193, 194, 207
 - numeric summary values 57, 64
- control groups 84, 86
- controlling for confounders 81
- correlation coefficients 175–80, 183
- counts 9
- covariates *see* independent variables
- Cox's regression model 236
- cross-over randomised control trials 86–8
- cross-section studies 12
 - association 177–80
 - study design 76–8
- cross-tabulations 25–6
- cumulative data 37–41
- cumulative frequencies 23–4
 - see also* ogives

- data, definition 3–4
- data collection *see* sampling
- databases 240, 241
- death *see* survival
- deciles 57
- decision rules 143–4
- dependent variables 72, 193, 207, 214–17
- descriptive statistics
 - charts 29–41
 - definition 17–18
 - distribution 43–9
 - frequency tables 5, 6, 17–27
 - numeric summary values 51–68
- design variables 199–200
- deviance coefficient 222
- diagnostics 205–9
- discrete metric data 9, 10
 - charts 34–5, 37–8, 40–1
 - frequency tables 23
 - numeric summary values 57, 64
- dispersion measures 52, 57–68
- distribution 43–9
 - bimodal 47
 - binomial 116
 - hypothesis tests 144–5
 - numeric summary values 55, 57, 58, 65–8
 - outliers 44
 - skew 44–5, 55, 57, 62, 64, 131
 - symmetric 44, 46
 - transformed data 66–8
 - uniform 43
 - see also* Normal distribution
- double-blind randomised control trials 86
- drop-out 89
- dummy variables 199–200

- Edinburgh Maternal Depression Scale 116
- elimination methods 202–3
- errors
 - blinding 86
 - drop-out 89
 - hypothesis tests 149–50
 - linear regression 194, 207–8, 211
 - recall bias 83
 - sampling 73, 83, 94, 112
 - selection bias 83, 84–5
- estimates 94–5
 - see also* confidence intervals
- exclusion criteria 240–1
- expected values 163–6, 182
- experimental studies 83–90
- explanatory variable *see* independent variables
- extraction of data 240–1

- false negatives/positives 150
- Fischer's exact test 145
- follow-up *see* cohort studies
- forest plots 241–3, 250
- forwards elimination 202
- forwards selection 201
- frequency matching 81–2
- frequency tables 5, 6, 17–27
 - categorical data 18–20, 23–6
 - contingency tables 25–6
 - continuous data 20–2
 - cross-tabulations 25–6
 - cumulative frequencies 23–4
 - discrete data 23
 - grouping data 20–2
 - metric data 20–3
 - nominal data 18–19
 - open-ended groups 22
 - ordinal data 20, 23–4
 - ranking data 27
 - relative frequency 19–20
- funnel plots 244–6

- GCS *see* Glasgow Coma Scale
- generalisation *see* statistical inference
- generalised linear model 209
- Glasgow Coma Scale (GCS) 5–7, 23–4
- goodness-of-fit 196–7, 203–4, 222–3
- grouped data 20–2, 35–7
- grouped frequency distributions 21–2

- hazard function 236
- hazard ratios 235–6
- heterogeneity 246–50
- histograms 35–7, 41
 - confidence intervals 120
 - distribution 44–6, 56, 65, 67
 - linear regression 211
 - numeric summary values 56, 65, 67
- homogeneity 246–50
- homoskedasticity 195

- Hosmer-Lemeshow statistic 222–3
- hypothesis tests
- chi-squared test 145, 151, 161–8, 221, 222, 234
 - confidence intervals 117, 119–31, 149, 156, 163
 - decision rules 143–4
 - difference between population parameters 141–54
 - equality of population proportions 161–8
 - errors 149–50
 - Fischer's exact test 145
 - independent populations 119–25, 127–31, 145–9, 151, 152–4, 162–3
 - Kruskal-Wallis test 145
 - logistic regression 221, 222
 - McNemar's test 145, 162
 - Mann-Whitney rank-sums test 127–31, 145, 147–9, 151
 - matched populations 125–6, 131, 145, 147, 149, 151, 162
 - matched-pairs *t* test 125–6, 145, 147, 151
 - mean 145–7
 - median 145, 147–9
 - Minitab 146, 148–9
 - Normal distribution 144–5
 - p* values 143–4, 146–9, 156–9, 164–6, 168
 - paired populations 145
 - power 150, 151–2, 168
 - procedure 143
 - proportions 161–8
 - ratio of two population parameters 155–9
 - research questions 142
 - rules of thumb 152–4
 - significance level 144, 150, 153
 - SPSS 146, 148
 - trend 166–8
 - two-sample *t* test 120–5, 145–6, 151, 222
 - Wilcoxon signed-rank test 117, 131, 145, 149, 151
- see also* null hypothesis
- incidence rate 53–4
- inclusion criteria 240–1
- independent populations
- difference 120–5, 127–31
 - hypothesis tests 145–9, 151, 152–4, 162–3
 - Mann-Whitney rank-sums test 127–31
 - ratios 133–4
 - two-sample *t* test 120–5
- independent variables
- linear regression 193, 199–201, 204, 207–8
 - logistic regression 216, 221–2
- inferences 77, 93–5
- informed guesses *see* confidence intervals
- Injury Severity Scale (ISS) 184
- intention-to-treat 89
- interquartile range (iqr) 58–61, 63–4, 232
- interval property 8
- iqr *see* interquartile range
- ISS *see* Injury Severity Scale
- journals 244
- Kaplan-Meier curves 230–1, 233–5
- Kaplan-Meier tables 228–30
- Kappa statistic 179, 182–4
- Kendal's rank-order correlation coefficient 180
- Killip scale 157
- Kruskal-Wallis test 145
- L'Abbé plots 247
- left skew *see* positive skew
- Levene's test 123
- limits of agreement 185–6
- linear association 172–3, 175
- linear regression 189–211
- analysis of variance 209–11
 - association 190–1, 192
 - assumptions 194–5, 205–9
 - causal relationships 190–1
 - coding design 199–200
 - colinearity 198
 - confounders 202–3, 204–5
 - design variables 199–200
 - diagnostics 205–9
 - goodness-of-fit 196–7, 203–4
 - Minitab 196
 - model-building 200–1
 - multiple 197–9, 203, 205–9
 - nominal independent variables 199–200
 - ordinary least squares 193–8, 205, 209
 - population regression equation 194
 - sample regression equation 193
 - SPSS 195–6
 - statistical significance 195–6
 - variable selection 200–3
 - variation 190–1
- location measures 52, 54–7, 59–61
- log-log plots 238
- log-rank test 233–5

- logistic regression 213–23
 - binary dependent variables 214–15
 - goodness-of-fit 222–3
 - maximum likelihood estimation 217–18
 - Minitab 217–20
 - model-building 221–2
 - multiple 221
 - odds ratios 217, 218–19, 220
 - regression coefficient 219
 - SPSS 217, 220–1
 - statistical inference 220–1
- longitudinal studies *see* case-control studies; cohort studies
- McNemar's test 145, 162
- Mann-Whitney rank-sums test 127–31, 145, 147–9, 151
- Mantel-Haenszel test 248–50
- manual variable selection 200, 202–3
- matched case-control studies 81, 82, 102
- matched populations 125–6, 131, 145, 147, 149, 151, 162
- matched-pairs *t* test 125–6, 145, 147, 151
- maximum likelihood estimation (MLE) 217–18
- mean
 - confidence intervals 112–16, 120–6, 134
 - hypothesis tests 145–7
 - linear regression 197–8
 - numeric summary values 55, 57
 - standard error 112–16
 - statistical inference 94
 - systematic review 242–3, 249
- measurements 8, 10
- median
 - confidence intervals 117–18, 127–31
 - hypothesis tests 145, 147–9
 - numeric summary values 54–5, 57, 59–61
 - statistical inference 94
 - survival time 231–2
- meta-analysis 239, 240, 246–50
 - homogeneity/heterogeneity 246–50
 - L'Abbé plots 247
 - Mantel-Haenszel test 248–50
- metric data 4, 7–9, 10
 - agreement 184–6
 - association 176, 180
 - charts 34–8, 40–1
 - confidence intervals 120, 126, 127, 131
 - frequency tables 20–3
 - hypothesis tests 145, 152–3
 - linear regression 193, 194, 207
 - logistic regression 222
 - numeric summary values 57, 64
 - MLE *see* maximum likelihood estimation
 - mode 54, 57
 - model-building 200–1
 - mound-shaped *see* symmetric distribution
 - multi-collinearity 198
 - multiple linear regression 197–9, 203, 205–9
 - multiple logistic regression 221
 - multivariate analysis 223, 238
- n-tiles 57
- negative
 - association 172–3
 - outcomes 244
 - skew 44–5, 55, 62
- NNT *see* number needed to treat
- nominal categorical data 4–5, 10
 - agreement 184
 - charts 30–4, 41
 - frequency tables 18–19
 - linear regression 199–200
 - numeric summary values 57, 64
- non-linear association 173, 175
 - see also* logistic regression
- non-parametric tests 127, 144–5, 151
- Normal distribution 48–9
 - association 176
 - confidence intervals 112–13, 120, 127
 - hypothesis tests 144–5
 - linear regression 194, 207, 209
 - probability 100
 - standard deviation 65–8
- null hypothesis
 - difference between population parameters 142–5, 148, 150
 - ratio of two population parameters 155–6, 158, 163–4, 168
 - survival 233–4
- number lines 6
- number needed to treat (NNT) 98, 106–7
- numeric summary values 51–68
 - dispersion measures 52, 57–68
 - distribution 55, 57, 58, 65–8
 - incidence rate 53–4
 - interquartile range 58–61, 63–4
 - location measures 52, 54–7, 59–61
 - numbers 52–3
 - ogives 60–1
 - outliers 55, 57, 58, 62
 - percentages 52–3

- numeric summary values (*Continued*)
 - percentiles 56–7
 - prevalence 53–4
 - proportions 52–3
 - quantitation 52
 - range 58
 - skew 55, 57, 62, 64
 - standard deviation 62–8
 - transformed data 66–8
- observed values 163–6
- odds 101–2, 103
- odds ratios 105–6
 - confidence intervals 134–5, 137–8
 - hypothesis tests 158–9
 - logistic regression 217, 218–19, 220
 - systematic review 245–6, 248–9
- odds 38–40, 41, 60–1
- OLS *see* ordinary least squares
- open trials 86
- open-ended groups 22
- opinions 77
- ordered categorical data 166–8
- ordering of data 5–7, 10, 18–20
- ordinal categorical data 5–7, 10
 - agreement 184
 - association 180
 - charts 30–4, 37–8, 40–1
 - confidence intervals 127, 131
 - frequency tables 19–20, 23–4
 - hypothesis tests 145
 - numeric summary values 55, 57, 64
- ordinary least squares (OLS) 193–8, 205, 209
- outcome variables *see* dependent variables
- outliers
 - distribution 44
 - frequency tables 22
 - meta-analysis 247
 - numeric summary values 55, 57, 58, 62
- OXCHECK Study Group 39
- p* values
 - association 176–9
 - hypothesis testing 143–4, 146–9, 156–9, 164–6, 168
 - linear regression 195–7, 198, 201–2
 - logistic regression 217, 220, 221–2
 - survival 234, 237–8
- Palliative Care Outcome scale (POS) 183
- parallel design 86
- parametric tests 127, 144–5, 151
- parsimony 202
- Pearson's correlation coefficient 175–7
- percentages
 - cumulative frequency 38–40
 - frequency 19–20
 - numeric summary values 52–3
- percentiles 56–7
- period prevalence 53
- pie charts 30–1, 41
- placebo bias 86
- point-biserial correlation coefficient 180
- point prevalence 53
- Poisson distribution 48–9
- population
 - correlation coefficients 175–7
 - difference between parameters 119–31, 141–54
 - logistic regression model 215–16
 - mean 112–16, 120–6, 134, 145–7
 - median 117–18, 127–31, 145, 147–9
 - odds ratios 137, 158–9
 - proportions 116–17, 126–7
 - ratio of two parameters 133–8, 155–9
 - regression equation 194
 - risk ratios 155–6
 - single parameter 111–18
 - statistical inference 93–5
 - study design 72–3
 - survival 230
 - see also* confidence intervals; hypothesis tests
- POS *see* Palliative Care Outcome scale
- positive
 - association 172–3
 - outcomes 244
 - skew 44–5, 55, 57, 62
- power of a test 150, 151–2, 168
- predictions 196
- predictors *see* independent variables
- prevalence 53–4, 77
- probability 97–100
 - calculation 99–100
 - case-control studies 98, 102–3, 105–6
 - cohort studies 100–1, 102, 104, 106–7
 - definition 98
 - logistic regression 215
 - Normal distribution 100
 - number needed to treat 98, 106–7
 - odds 101–2, 103
 - odds ratios 105–6
 - Poisson distribution 48–9

- risk 100–1, 102–3
- risk ratios 104
- survival 228–31
- proportional frequency 99
- proportional hazards 236–8
- proportions
 - confidence intervals 116–17, 126–7
 - hypothesis tests 161–8
 - numeric summary values 52–3
 - populations 161–8
 - samples 116
- prospective studies *see* cohort studies
- Psychiatric Symptom Frequency (PSF) scale 46–7
- publication bias 244, 245

- quintiles 57

- random number tables 85, 251
- randomisation 74, 84–5, 88–9
- randomised controlled trials (RCT)
 - hypothesis tests 146, 151, 156–7, 165
 - study design 85, 86–90
 - systematic review 242
- range 58
- ranked data
 - frequency tables 27
 - Kendal's rank-order correlation coefficient 180
 - log-rank test 233–5
 - Mann-Whitney rank-sums test 127–31, 145, 147–9, 151
 - Spearman's rank correlation coefficient 177–80, 183
 - Wilcoxon signed-rank test 117, 131, 145, 149, 151
- ratio property 8
- RCT *see* randomised control trials
- recall bias 83
- reference values 102
- regression *see* linear regression; logistic regression
- relative frequency 19–20
 - cumulative 38–9, 60
- relative risk *see* risk ratios
- research questions 142
- residuals 194, 207–8, 211
- response bias 86
- response variables *see* dependent variables
- retrospective studies *see* case-control studies
- review *see* systematic review
- right skew *see* negative skew
- risk 100–1, 102–3, 217
- risk ratios 100, 104
 - confidence intervals 134–6
 - hypothesis tests 155–7
 - survival 237
 - systematic review 242, 245, 248, 250
- rules of thumb 152–4

- sample
 - correlation coefficients 175–80
 - logistic regression model 216
 - mean 112, 116, 120–2, 134
 - odds ratios 137
 - percentage 94
 - proportions 116
 - regression equation 193
 - statistic 94
 - survival 230
- sampling 72, 73
 - errors 73, 83, 94, 112
 - frames 74
 - randomisation 74, 84–5, 88–9
 - statistical inference 93–5
 - types 74–5
- scatterplots 172–5, 176
 - linear regression 192, 196, 201, 208–11
 - logistic regression 214–16
- selection bias 83, 84–5, 88–9
- significance level (α) 144, 150, 153
- simple bar charts 31–2
- simple random sampling 74
- skew 44–5, 55, 57, 62, 64, 131
- slope coefficient 191
- Spearman's rank correlation coefficient 177–80, 183
- spread *see* dispersion measures
- stacked bar charts 34
- standard deviation 62–8
 - agreement 185–6
 - confidence intervals 120, 123
- standard error 112–16
- statistical inference 77, 93–5
- step charts 37–8, 41
- stepwise selection 201–2
- straight line models *see* linear regression
- stratified random sampling 74
- study design 71–90
 - blinding 86
 - case-control studies 80–3
 - case-series studies 76
 - clinical trials 84
 - cohort studies 78–80, 83

- study design (*Continued*)
 - confounders 81, 84
 - contingency tables 79–80, 82
 - cross-section studies 76–8
 - experimental studies 83–90
 - intention-to-treat 89
 - matching 81–2
 - outcome variables 72
 - populations 72–3
 - randomisation 83, 84–5, 88–9
 - randomised control trials 85, 86–90
 - sampling 72–5, 83–5, 88–9
 - types of study 75–81
- study populations 73, 75, 93–4
- sub-groups 25
- sum of squares 63
- summary values *see* numeric summary values
- surveys 76–8
- survival 227–38
 - censored data 228
 - comparison between groups 232–9
 - Cox's regression model 236
 - hazard ratios 235–6
 - Kaplan-Meier curves 230–1, 233–5
 - Kaplan-Meier tables 228–30
 - log-log plots 238
 - log-rank test 233–5
 - median 231–2
 - null hypothesis 233–4
 - probability 228–31
 - proportional hazards 236–8
 - single groups 228
- symmetric distribution 44, 46
- systematic random sampling 74
- systematic review 239–45
 - extraction of data 240–1
 - forest plots 241–3, 250
 - funnel plots 244–6
 - homogeneity/heterogeneity 246–50
 - identification of trials 240–1
 - inclusion criteria 240–1
 - L'Abbé plots 247
 - Mantel-Haenszel test 248–50
 - meta-analysis 239, 240, 246–50
 - methods 240–3
 - publication bias 244, 245
 - search strategy 241
- t* distribution 114, 120–6, 145–7, 151, 222
- tables *see* frequency tables
- target populations 73, 75, 93–4
- test statistic 164
- tests *see* hypothesis tests
- time series charts 40–1
- transformed data 66–8
- treatment bias 86
- treatment groups 84, 86
- trend 166–8
- two-sample *t* test 120–5, 145–6, 151, 222
- type I/II errors 150
- uniform distributions 43
- units 5, 7–9
- univariate analysis 238
- univariate logistic regression 222
- unmatched case-control studies 81–2
- variables
 - characteristics 9–13
 - definition 3–4
 - selection 200–3
 - types 4–9
 - see also* categorical; continuous; discrete; metric; nominal; ordinal data
- variation 190–1
- visual analogue scale (VAS) 10, 59, 118
- Wald statistic 220–1
- weighted kappa 184
- weighted mean 242–3, 249
- Wilcoxon signed-rank test 117, 131, 145, 149, 151
- z* distribution 220