Statistical Thinking for Non-Statisticians in Drug Regulation

Richard Kay

Consultant in Statistics for the Pharmaceutical Industry
Great Longstone
Derbyshire, UK

John Wiley & Sons, Ltd

Statistical Thinking for Non-Statisticians in Drug Regulation

Statistical Thinking for Non-Statisticians in Drug Regulation

Richard Kay

Consultant in Statistics for the Pharmaceutical Industry
Great Longstone
Derbyshire, UK

John Wiley & Sons, Ltd

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.wiley.com

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, Canada

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

Anniversary Logo Design: Richard J. Pacifico

Library of Congress Cataloging in Publication Data

Kay, R. (Richard), 1949–
Statistical thinking for non-statisticians in drug regulation / Richard Kay.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-31971-0 (cloth : alk. paper)
1. Clinical trials—Statistical methods. 2. Drugs—Testing—Statistical methods. 3. Drug approval—Statistical methods. 4. Pharmaceutical industry—Statistical methods. I. Title. [DNLM: 1. Clinical Trials—methods. 2. Statistics. 3. Drug Approval. 4. Drug Industry. QV 771 K23s 2007]
R853.C55K39 2007
615.5′8′0724—dc22

2007022438

British Library Cataloguing in Publication Data

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

ISBN 978-0-470-31971-0

Typeset in 10.5/13pt Minion by Integra Software Services Pvt. Ltd, Pondicherry, India
Printed and bound in Great Britain by Antony Rowe Ltd., Chippenham, Wiltshire
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.

To Jan, Matt, Sally and Becci

Contents

Sections marked with an asterisk refer to some more challenging sections of the book.

Preface xiii

Abbreviations xvii

1 Basic ideas in clinical trial design 1

1.1 Historical perspective 1
1.2 Control groups 2
1.3 Placebos and blinding 3
1.4 Randomisation 4
1.4.1 Unrestricted randomisation 5
1.4.2 Block randomisation 5
1.4.3 Unequal randomisation 6
1.4.4 Stratified randomisation 7
1.4.5 Central randomisation 8
1.4.6 Dynamic allocation and minimisation 9
1.4.7 Cluster randomisation 10
1.5 Bias and precision 11
1.6 Between- and within-patient designs 12
1.7 Cross-over trials 14
1.8 Signal and noise 15
1.8.1 Signal 15
1.8.2 Noise 15
1.8.3 Signal-to-noise ratio 15
1.9 Confirmatory and exploratory trials 16
1.10 Superiority, equivalence and non-inferiority trials 17
1.11 Data types 18
1.12 Choice of endpoint 20
1.12.1 Primary variables 20
1.12.2 Secondary variables 21
1.12.3 Surrogate variables 21
1.12.4 Global assessment variables 22
1.12.5 Composite variables 23
1.12.6 Categorisation 23

2 Sampling and inferential statistics 25
2.1 Sample and population 25
2.2 Sample statistics and population parameters 26
2.2.1 Sample and population distribution 26
2.2.2 Median and mean 27
2.2.3 Standard deviation 28
2.2.4 Notation 29
2.3 The normal distribution 29
2.4 Sampling and the standard error of the mean 32
2.5 Standard errors more generally 35
2.5.1 The standard error for the difference between two means 35
2.5.2 Standard errors for proportions 38
2.5.3 The general setting 38

3 Confidence intervals and *p*-values 39
3.1 Confidence intervals for a single mean 39
3.1.1 The 95 per cent confidence interval 39
3.1.2 Changing the confidence coefficient 41
3.1.3 Changing the multiplying constant 41
3.1.4 The role of the standard error 43
3.2 Confidence intervals for other parameters 44
3.2.1 Difference between two means 44
3.2.2 Confidence intervals for proportions 45
3.2.3 General case 46
3.3 Hypothesis testing 47
3.3.1 Interpreting the *p*-value 47
3.3.2 Calculating the *p*-value 49
3.3.3 A common process 52
3.3.4 The language of statistical significance 55
3.3.5 One-tailed and two-tailed tests 55

4 Tests for simple treatment comparisons 57
4.1 The unpaired t-test 57
4.2 The paired t-test 58
4.3 Interpreting the t-tests 61
4.4 The chi-square test for binary data 63
4.4.1 Pearson chi-square 63
4.4.2 The link to a signal-to-noise ratio 66
4.5 Measures of treatment benefit 67
4.5.1 Odds ratio (OR) 67
4.5.2 Relative risk (RR) 68
4.5.3 Relative risk reduction (RRR) 69
4.5.4 Number needed to treat (NNT) 69
4.5.5 Confidence intervals 70
4.5.6 Interpretation 71
4.6 Fisher's exact test 71
4.7 The chi-square test for categorical and ordinal data 73
4.7.1 Categorical data 73

4.7.2 Ordered categorical (ordinal) data 75
4.7.3 Measures of treatment benefit for categorical and ordinal data 76
4.8 Extensions for multiple treatment groups 77
4.8.1 Between-patient designs and continuous data 77
4.8.2 Within-patient designs and continuous data 78
4.8.3 Binary, categorical and ordinal data 79
4.8.4 Dose ranging studies 79
4.8.5 Further discussion 80

5 Multi-centre trials 81
5.1 Rationale for multi-centre trials 81
5.2 Comparing treatments for continuous data 82
5.3 Evaluating homogeneity of treatment effect 84
5.3.1 Treatment-by-centre interactions 84
5.3.2 Quantitative and qualitative interactions 87
5.4 Methods for binary, categorical and ordinal data 88
5.5 Combining centres 88

6 Adjusted analyses and analysis of covariance 91
6.1 Adjusting for baseline factors 91
6.2 Simple linear regression 92
*6.3 Multiple regression 94
6.4 Logistic regression 96
6.5 Analysis of covariance for continuous data 97
6.5.1 Main effect of treatment 97
6.5.2 Treatment-by-covariate interactions 99
*6.5.3 A single model 101
6.5.4 Connection with adjusted analyses 102
6.5.5 Advantages of analysis of covariance 102
6.6 Binary, categorical and ordinal data 104
6.7 Regulatory aspects of the use of covariates 106
*6.8 Connection between ANOVA and ANCOVA 109
6.9 Baseline testing 109

7 Intention-to-treat and analysis sets 111
7.1 The principle of intention-to-treat 111
7.2 The practice of intention-to-treat 115
7.2.1 Full analysis set 115
7.2.2 Per-protocol set 117
7.2.3 Sensitivity 117
7.3 Missing data 118
7.3.1 Introduction 118
7.3.2 Complete cases analysis 119
7.3.3 Last observation carried forward (LOCF) 119
7.3.4 Success/failure classification 120
7.3.5 Worst case/best case imputation 120
7.3.6 Sensitivity 121
7.3.7 Avoidance of missing data 121
7.4 Intention-to-treat and time-to-event data 122
7.5 General questions and considerations 124

8 Power and sample size 127
8.1 Type I and type II errors 127
8.2 Power 128
8.3 Calculating sample size 131
8.4 Impact of changing the parameters 134
8.4.1 Standard deviation 134
8.4.2 Event rate in the control group 135
8.4.3 Clinically relevant difference 135
8.5 Regulatory aspects 136
8.5.1 Power > 80 per cent 136
8.5.2 Powering on the per-protocol set 137
8.5.3 Sample size adjustment 137
8.6 Reporting the sample size calculation 138

9 Statistical significance and clinical importance 141
9.1 Link between *p*-values and confidence intervals 141
9.2 Confidence intervals for clinical importance 143
9.3 Misinterpretation of the *p*-value 144
9.3.1 Conclusions of similarity 144
9.3.2 The problem with 0.05 145

10 Multiple testing 147
10.1 Inflation of the type I error 147
10.2 How does multiplicity arise 148
10.3 Regulatory view 148
10.4 Multiple primary endpoints 149
10.4.1 Avoiding adjustment 149
10.4.2 Significance needed on all endpoints 149
10.4.3 Composite endpoints 150
10.4.4 Variables ranked according to clinical importance 150
10.5 Methods for adjustment 152
10.6 Multiple comparisons 153
10.7 Repeated evaluation over time 154
10.8 Subgroup testing 155
10.9 Other areas for multiplicity 157
10.9.1 Using different statistical tests 157
10.9.2 Different analysis sets 158

11 Non-parametric and related methods 159
11.1 Assumptions underlying the t-tests and their extensions 159
11.2 Homogeneity of variance 160
11.3 The assumption of normality 160
11.4 Transformations 163
11.5 Non-parametric tests 166
11.5.1 The Mann–Whitney U-test 166
11.5.2 The Wilcoxon signed rank test 168
11.5.3 General comments 169
11.6 Advantages and disadvantages of non-parametric methods 169
11.7 Outliers 170

12 Equivalence and non-inferiority 173
12.1 Demonstrating similarity 173
12.2 Confidence intervals for equivalence 175
12.3 Confidence intervals for non-inferiority 176
12.4 A *p*-value approach 178
12.5 Assay sensitivity 180
12.6 Analysis sets 182
12.7 The choice of Δ 182
12.7.1 Bioequivalence 183
12.7.2 Therapeutic equivalence 183
12.7.3 Non-inferiority 184
12.7.4 The 10 per cent rule for cure rates 185
12.7.5 Biocreep and constancy 186
12.8 Sample size calculations 187
12.9 Switching between non-inferiority and superiority 189

13 The analysis of survival data 193
13.1 Time-to-event data and censoring 193
13.2 Kaplan–Meier (KM) curves 195
13.2.1 Plotting KM curves 195
13.2.2 Event rates and relative risk 196
13.2.3 Median event times 196
13.3 Treatment comparisons 197
13.4 The hazard ratio 200
13.4.1 The hazard rate 200
13.4.2 Constant hazard ratio 201
13.4.3 Non-constant hazard ratio 201
13.4.4 Link to survival curves 202
*13.4.5 Calculating KM curves 203
*13.5 Adjusted analyses 204
13.5.1 Stratified methods 204
13.5.2 Proportional hazards regression 204
13.5.3 Accelerated failure time model 207
13.6 Independent censoring 208
13.7 Sample size calculations 209

14 Interim analysis and data monitoring committees 213
14.1 Stopping rules for interim analysis 213
14.2 Stopping for efficacy and futility 214
14.2.1 Efficacy 214
14.2.2 Futility and conditional power 215
14.2.3 Some practical issues 216
14.2.4 Analyses following completion of recruitment 217
14.3 Monitoring safety 218
14.4 Data Monitoring Committees 219
14.4.1 Introduction and responsibilities 219
14.4.2 Structure 220
14.4.3 Meetings and recommendations 222

14.5 Adaptive designs 223
14.5.1 Sample size re-evaluation 223
14.5.2 Flexible designs 224

15 Meta-analysis 229
15.1 Definition 229
15.2 Objectives 231
15.3 Statistical methodology 232
15.3.1 Methods for combination 232
15.3.2 Confidence Intervals 233
15.3.3 Fixed and random effects 234
15.3.4 Graphical methods 234
15.3.5 Detecting heterogeneity 236
15.3.6 Robustness 236
15.4 Ensuring scientific validity 237
15.4.1 Planning 237
15.4.2 Publication bias and funnel plots 238
15.5 Meta-analysis in a regulatory setting 240
15.5.1 Retrospective analyses 240
15.5.2 One pivotal study 241

16 The role of statistics and statisticians 245
16.1 The importance of statistical thinking at the design stage 245
16.2 Regulatory guidelines 247
16.3 The statistics process 249
16.3.1 The Statistical Methods section of the protocol 250
16.3.2 The statistical analysis plan 250
16.3.3 The data validation plan 251
16.3.4 The blind review 251
16.3.5 Statistical analysis 252
16.3.6 Reporting the analysis 252
16.3.7 Pre-planning 253
16.3.8 Sensitivity and robustness 255
16.4 The regulatory submission 256
16.5 Publications and presentations 257

References 261

Index 267

Preface

This book is primarily concerned with clinical trials planned and conducted within the pharmaceutical industry. Much of the methodology presented is in fact applicable on a broader basis and can be used in observational studies and in clinical trials outside of the pharmaceutical sector; nonetheless the primary context is clinical trials and pharmaceuticals. The development is aimed at non-statisticians and will be suitable for physicians, investigators, clinical research scientists, medical writers, regulatory personnel, statistical programmers, senior data managers and those working in quality assurance. Statisticians moving from other areas of application outside of pharmaceuticals may also find the book useful in that it places the methods that they are familiar with, in context in their new environment. There is substantial coverage of regulatory aspects of drug registration that impact on statistical issues. Those of us working within the pharmaceutical industry recognise the importance of being familiar with the rules and regulations that govern our activities and statistics is a key aspect of this.

The aim of the book is not to turn non-statisticians into statisticians. I do not want you to go away from this book and 'do' statistics. It is the job of the statistician to provide statistical input to the development plan, to individual protocols, to write the statistical analysis plan, to analyse the data and to work with medical writing in producing the clinical report; also to support the company in its interactions with regulators on statistical issues.

The aims of the book are really three-fold. Firstly, to aid communication between statisticians and non-statisticians, secondly, to help in the critical review of reports and publications and finally, to enable the more effective use of statistical arguments within the regulatory process. We will take each of these points in turn.

In many situations the interaction between a statistician and a non-statistician is not a particularly successful one. The statistician uses terms, for example, power, odds ratio, p-value, full analysis set, hazard ratio, non-inferiority, type II error, geometric mean, last observation carried forward and so on, of which the non-statistician has a vague understanding, but maybe not a good enough understanding to be able to get an awful lot out of such interactions. Of course it is always the job of a statistician to educate and every opportunity should be taken for imparting knowledge about statistics, but in a specific context there may not be time for that. Hopefully this book will explain, in ways that are understandable,

just what these terms mean and provide some insight into their interpretation and the context in which they are used. There is also a lot of confusion between what on the surface appear to be the same or similar things; significance level and p-value, equivalence and non-inferiority, odds ratio and relative risk, relative risk and hazard ratio (by the way this is a minefield!) and meta-analysis and pooling to name just a few. This book will clarify these important distinctions.

It is unfortunately the case that many publications, including some leading journals, contain mistakes with regard to statistics. Things have improved over the years with the standardisation of the ways in which publications are put together and reviewed. For example the CONSORT statement (see Section 16.5) has led to a distinct improvement in the quality of reporting. Nonetheless mistakes do slip through, in terms of poor design, incorrect analysis, incomplete reporting and inappropriate interpretation – hopefully not all at once! It is important therefore when reading an article that the non-statistical reader is able to make a judgement regarding the quality of the statistics and to notice any obvious flaws that may undermine the conclusions that have been drawn. Ideally the non-statistician should involve their statistical colleagues in evaluating their concerns but keeping a keen eye on statistical arguments within the publication may help to alert the non-statistician to a potential problem. The same applies to presentations at conferences, posters, advertising materials and so on.

Finally the basis of many concerns raised by regulators, when they are reviewing a proposed development plan or assessing an application for regulatory approval, is statistical. It is important that non-statisticians are able to work with their statistical colleagues in correcting mistakes, changing aspects of the design, responding to questions about the data to hopefully overcome those concerns.

In writing this book I have made the assumption that the reader is familiar with general aspects of the drug development process. I have assumed knowledge of the phase I to phase IV framework, of placebos, control groups, and double-dummy together with other fundamental elements of the nuts and bolts of clinical trials. I have assumed however no knowledge of statistics! This may or may not be the correct assumption in individual cases, but it is the common denominator that we must start from, and also it is actually not a bad thing to refresh on the basics. The book starts with some basic issues in trial design in Chapter 1 and I guess most people picking up this book will be familiar with many of the topics covered there. But don't be tempted to skip this chapter; there are still certain issues, raised in this first chapter, that will be new and important for understanding arguments put forward in subsequent chapters. Chapter 2 looks at sampling and inferential statistics. In this chapter we look at the interplay between the population and the sample, basic thoughts on measuring average and variability and then explore the process of sampling leading to the concept of the standard error as a way of capturing precision/reliability of the sampling process. The construction and interpretation of confidence intervals is covered in

Chapter 3 together with testing hypotheses and the (dreaded!) p-value. Common statistical tests for various data types are developed in Chapter 4 which also covers different ways of measuring treatment effect for binary data, such as the odds ratio and relative risk.

Many clinical trials that we conduct are multi-centre and Chapter 5 looks at how we extend our simple statistical comparisons to this more complex structure. These ideas lead naturally to the topics in Chapter 6 which include the concepts of adjusted analyses, and more generally, analysis of covariance which allows adjustment for many baseline factors, not just centre. Chapters 2 to 6 follow a logical development sequence in which the basic building blocks are initially put in place and then used to deal with more and more complex data structures. Chapter 7 moves a little away from this development path and covers the important topic of Intention-to-Treat and aspects of conforming with that principle through the definition of different analysis sets and dealing with missing data. In Chapter 8, we cover the very important design topics of power and the sample size calculation which then leads naturally to a discussion about the distinction between statistical significance and clinical importance in Chapter 9.

The regulatory authorities, in my experience, tend to dig their heels in on certain issues and one such issue is multiplicity. This topic, which has many facets, is discussed in detail in Chapter 10. Non-parametric and related methods are covered in Chapter 11. In Chapter 12 we develop the concepts behind the establishment of equivalence and non-inferiority. This is an area where many mistakes are made in applications, and in many cases these slip through into published articles. It is a source of great concern to many statisticians that there is widespread misunderstanding of how to deal with equivalence and non-inferiority. I hope that this chapter helps to develop a better understanding of the methods and the issues. If you have survived so far, then Chapter 13 covers the analysis of survival data. When an endpoint is time to some event, for example death, the data are inevitably subject to what we call censoring and it is this aspect of so-called survival data that has led to the development of a completely separate set of statistical methods. Chapter 11 builds on the earlier discussion on multiplicity to cover one particular manifestation of that, the interim analysis. This chapter also looks at the management of these interim looks at the data through data monitoring committees. Meta-analysis and its role in clinical development is covered in Chapter 15 and the book finishes with a general Chapter 16 on the role of statistics and statisticians in terms of the various aspects of design and analysis and statistical thinking more generally.

It should be clear from the last few paragraphs that the book is organised in a logical way; it is a book for learning rather than a reference book for dipping into. The development in later chapters will build on the development in earlier chapters. I strongly recommend, therefore, that you start on page 1 and work through. I have tried to keep the discussion away from formal mathematics. There are

formulas in the book but I have only included these where I think this will enhance understanding; there are no formulas for formulas sake! There are some sections that are more challenging than others and I have marked with an asterisk those sections that can be safely sidestepped on a first (or even a second) run through the book.

The world of statistics is ever changing. New methods are being developed by theoreticians within university departments and ultimately some of these will find their way into mainstream methods for design and statistical analysis within our industry. The regulatory environment is ever changing as regulators respond to increasing demands for new and more effective medicines. This book in one sense represents a snapshot in time in terms of what statistical methods are employed within the pharmaceutical industry and also in relation to current regulatory requirements. Two statistical topics that are not included in this book are Bayesian Methods and Adaptive (Flexible) Designs (although some brief mention is made of this latter topic in section 14.5.2). Both areas are receiving considerable attention at the moment and I am sure that within a fairly short period of time there will be much to say about them in terms of the methodological thinking, examples of their application and possibly with regard to their regulatory acceptance but for the moment they are excluded from our discussions.

The book has largely come out of courses that I have been running under the general heading of Statistical Thinking for Non-Statisticians for a number of years. There have been several people who have contributed from time to time and I would like to thank them for their input and support; Werner Wierich, Mike Bradburn and in particular Ann Gibb who gave these courses with me over a period of several years and enhanced my understanding through lively discussion and asking many challenging questions. I would also like to thank Simon Gillis who contributed to Chapter 16 with his much deeper knowledge of the processes that go on within a pharmaceutical company in relation to the analysis and reporting of a clinical trial.

Richard Kay
Great Longstone
January 2007

Abbreviations

AIDAC	Anti-Infective Drugs Advisory Committee
ANCOVA	analysis of covariance
ANOVA	analysis of variance
AE	adverse event
ARR	absolute relative risk
AUC	area under the time concentration curve
BMD	bone mineral density
CDER	Center for Drug Evaluation and Research
CFC	Chlorofluorocarbon
CHMP	Committee for Medical Products for Human Use
CI	confidence interval
CPMP	Committee for Proprietary Medicinal Products
C_{MAX}	maximum concentration
CMH	Cochran-Mantel-Haenszel
CNS	central nervous system
CRF	Case Report Form
CR	complete response
crd	clinically relevant difference
DMC	Data Monitoring Committee
DSMB	Data and Safety Monitoring Board
DSMC	Data and Safety Monitoring Committee
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
EMEA	European Medicines Evaluation Agency
FDA	Food and Drug Administration
FEV_1	forced expiratory volume in one second
GP	General Practitioner
HAMA	Hamilton Anxiety Scale
HAMD	Hamilton Depression Scale

HER2	human epidermal growth factor receptor-2
HIV	human immunodeficiency virus
HR	Hazard Ratio
ICH	International Committee on Harmonisation
ITT	Intention-to-Treat
IVRS	Interactive Voice Response System
KM	Kaplan-Meier
LDH	lactate hehydrogenase
LOCF	last observation carried forward
MedDRA	*Med*ical *D*ictionary for *R*egulatory *A*ctivities
MH	Mantel-Haenszel
MI	myocardial infarction
NNH	number needed to harm
NNT	number needed to treat
NS	not statistically significant
OR	odds ratio
PD	progressive disease
PEF	peak expiratory flow
PHN	post-hepatic neuralgia
PR	partial response
RECIST	Response Evaluation Criteria in Solid Tumours
RR	relative risk
RRR	relative risk reduction
SAE	serious adverse event
SAP	Statistical Analysis Plan
SD	stable disease
sd	standard deviation
se	standard error
VAS	visual analogue scale
WHO	World Health Organisation

1

Basic ideas in clinical trial design

1.1 Historical perspective

As many of us who are involved in clinical trials will know, the randomised, controlled trial is a relatively new invention. As pointed out by Pocock (1983) and others, very few clinical trials of the kind we now regularly see were conducted prior to 1950. It took a number of high profile successes plus the failure of alternative methodologies to convince researchers of their value.

Example 1.1: The Salk Polio Vaccine trial

One of the largest trials ever conducted took place in the US in 1954 and concerned the evaluation of the Salk Polio Vaccine. The trial has been reported extensively by Meier (1978) and is used by Pocock (1983) in his discussion of the historical development of clinical trials.

Within the project there were essentially two trials and these clearly illustrated the effectiveness of the randomised controlled design.

Trial 1: Original design; observed control
1.08 million children from selected schools were included in this first trial. The second graders in those schools were offered the vaccine while the first and third graders would serve as the control group. Parents of the second graders were approached for their consent and it was noted that the consenting parents tended to have higher incomes. Also, this design was not blinded so that both parents and investigators knew which children had received the vaccine and which had not.

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

Example 1.1: (Continued)

Trial 2: Alternative design; randomised control
A further 0.75 million children in other selected schools in grades one to three were to be included in this second trial. All parents were approached for their consent and those children where consent was given were randomised to receive either the vaccine or a placebo injection. The trial was double-blind with parents, children and investigators unaware of who had received the vaccine and who had not.

The results from the randomised control trial were conclusive. The incidence of paralytic polio for example was 0.057 per cent in the placebo group compared to 0.016 per cent in the active group and there were four deaths in the placebo group compared to none in the active group. The results from the observed control trial, however, were less convincing with a smaller observed difference (0.046 per cent versus 0.017 per cent). In addition, in the cases where consent could not be obtained, the incidence of paralytic polio was 0.036 per cent in the randomised trial and 0.037 per cent in the observed control trial, event rates considerably lower than those amongst placebo patients and in the untreated controls respectively. This has no impact on the conclusions from the randomised trial, which is robust against this absence of consent; the randomised part is still comparing like with like. In the observed control part however the fact that the 'no consent' (grade 2) children have a lower incidence that those children (grades 1 and 3) who were never offered the vaccine potentially causes some confusion in a non-randomised comparison; does it mean that grade 2 children naturally have lower incidence than those in grades 1 and 3? Whatever the explanation, the presence of this uncertainty reduced confidence in other aspects of the observed control trial.

The randomised part of the Salk Polio Vaccine trial has all the hallmarks of modern day trials; randomisation, control group, blinding and it was experiences of these kinds that helped convince researchers that only under these conditions can clear, scientifically valid conclusions be drawn.

1.2 Control groups

We invariably evaluate our treatments by making comparisons; active compared to control. It is very difficult to make absolute statements about specific treatments and conclusions regarding the efficacy and safety of a new treatment are made relative to an existing treatment or placebo.

ICH E10 (2001): 'Note for Guidance on Choice of Control Group in Clinical Trials'

'Control groups have one major purpose: to allow discrimination of patient outcomes (for example, changes in symptoms, signs, or other morbidity) caused by the test treatment from outcomes caused by other factors, such as the natural progression of the disease, observer or patient expectations, or other treatment.'

Control groups can take a variety of forms, here are just a few examples of trials with alternative types of control group:

- Active versus placebo
- Active A versus active B (versus active C)
- Placebo versus dose level 1 versus dose level 2 versus dose level 3 (dose-finding)
- Active A + active B versus active A + placebo (add-on)

The choice will depend on the objectives of the trial.

Open trials with no control group can nonetheless be useful in an exploratory, maybe early phase setting, but it is unlikely that such trials will be able to provide confirmatory, robust evidence regarding the performance of the new treatment.

Similarly, external or historical controls (groups of subjects external to the study either in a different setting or previously treated) cannot provide definitive evidence. Byar (1980) provides an extensive discussion on these issues.

1.3 Placebos and blinding

It is important to have blinding of both the subject and the investigator wherever possible to avoid unconscious bias creeping in, either in terms of the way a subject reacts psychologically to a treatment or in relation to the way the investigator influences or records subject outcome.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Blinding or masking is intended to limit the occurrence of conscious or unconscious bias in the conduct and interpretation of a clinical trial arising from the influence which the knowledge of treatment may have on the recruitment and allocation of subjects, their subsequent care, the attitudes of subjects to the treatments, the assessment of the end-points, the handling of withdrawals, the exclusion of data from analysis, and so on.'

Ideally the trial should be *double-blind* with both the subject and the investigator being blind to the specific treatment allocation. If this is not possible for the investigator, for example, then the next best thing is to have an independent evaluation of outcome, both for efficacy and for safety. A *single-blind* trial arises when either the subject or investigator, but not both, is blind.

An absence of blinding can seriously undermine the validity of an endpoint in the eyes of regulators and the scientific community more generally, especially when the evaluation of that endpoint has an element of subjectivity. In situations where blinding is not possible it is essential to use hard, unambiguous endpoints.

The use of placebos and blinding go hand in hand. The existence of placebos enable trials to be blinded and account for the placebo effect; the change in a patient's condition that is due to the act of being treated, but is not caused by the active component of that treatment.

1.4 Randomisation

Randomisation is clearly a key element in the design of our clinical trials. There are two reasons why we randomise subjects to the treatment groups:

- To avoid any bias in the allocation of the patients to the treatment groups
- To ensure the validity of the statistical test comparisons

Randomisation lists are produced in a variety of ways and we will discuss several methods later. Once the list is produced the next patient entering the trial receives the next allocation within the randomisation scheme. In practice this process is managed by 'packaging' the treatments according to the pre-defined randomisation list.

There are a number of different possibilities when producing randomisation lists:

- Unrestricted randomisation
- Block randomisation
- Unequal randomisation
- Stratified randomisation
- Central randomisation
- Dynamic allocation and minimisation
- Cluster randomisation

1.4.1 Unrestricted randomisation

Unrestricted (or simple) randomisation is simply a random list of, for example, As and Bs. In a moderately large trial, with say $n = 200$ subjects, such a process will likely produce approximately equal group sizes. There is no guarantee however that this will automatically happen and in small trials, in particular, this can cause problems.

1.4.2 Block randomisation

To ensure balance in terms of numbers of subjects, we usually undertake *block randomisation* where a randomisation list is constructed by randomly choosing from the list of potential blocks. For example, there are six ways of allocating two As and two Bs in a 'block' of size four:

AABB, ABAB, ABBA, BAAB, BABA, BBAA

and we choose at random from this set of six blocks to produce our randomisation list, for example:

ABBA BAAB ABAB ABBA, …

Clearly if we recruit a multiple of four patients into the trial we will have perfect balance, and approximate balance (which is usually good enough) for any sample size.

In large trials it could be argued that block randomisation is unnecessary. In one sense this is true, overall balance will be achieved by chance with an unrestricted randomisation list. However, it is usually the case that large trials will be multicentre trials and not only is it important to have balance overall it is also important to have balance within each centre. In practice therefore we would allocate several blocks to each centre, for example five blocks of size four if we are planning to recruit 20 patients from each centre. This will ensure balance within each centre and also overall.

How do we choose block size? There is no magic formula but more often than not the block size is equal to two times the number of treatments.

What are the issues with block size?

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Care must be taken to choose block lengths which are sufficiently short to limit possible imbalance, but which are long enough to avoid predictability towards the end of the sequence in a block. Investigators and other relevant staff should generally be blind to the block length . . . '

Shorter block lengths are better at producing balance. With two treatments a block length of four is better at producing balance than a block length of 12. The block length of four gives perfect balance if there is a multiple of four patients entering, whereas with a block length of 12, perfect balance is only going to be achieved if there are a multiple of 12 patients in the study. The problem, however, with the shorter block lengths is that this is an easy code to crack and inadvertent unblinding can occur. For example suppose a block length of four was being used in a placebo controlled trial and also assume that experience of the active drug suggests that many patients receiving that drug will suffer nausea. Suppose the trial begins and the first two patients suffer nausea. The investigator is likely to conclude that both these patients have been randomised to active and that therefore the next two allocations are to placebo. This knowledge could influence his willingness to enter certain patients into the next two positions in the randomisation list, causing bias in the mix of patients randomised into the two treatment groups. Note the comment in the ICH guideline regarding keeping the investigator (and others) blind to the block length. While in principle this comment is sound, the drug is often delivered to a site according to the chosen block length, making it difficult to conceal information on block size. If the issue of inadvertent unblinding is going to cause problems then more sophisticated methodologies can be used, such as having the block length itself varying; perhaps randomly chosen from two, four or six.

1.4.3 Unequal randomisation

All other things being equal, having equal numbers of subjects in the two treatment groups provides the maximum amount of information (the greatest power) with regard to the relative efficacy of the treatments. There may, however, be issues that override statistical efficiency:

- It may be necessary to place more patients on active compared to placebo in order to obtain the required safety information.
- In a three group trial with active A, active B and placebo(P), it may make sense to have a 2:2:1 randomisation to give more power for the A versus B comparison as that difference is likely to be smaller then the A versus P and B versus P differences.

Unequal randomisation is sometimes needed as a result of these considerations. To achieve this, the randomisation list will be designed for the second example above with double the number of A and B allocations compared to placebo.

For unequal randomisation we would choose the block size accordingly. For a 2:1 randomisation to A or P we could randomly choose from the blocks:

AAP, APA, PAA

1.4.4 Stratified randomisation

Block randomisation therefore forces the required balance in terms of the numbers of patients in the treatment groups, but things can still go wrong. For example, let's suppose in an oncology study with time to death as the primary endpoint that we can measure baseline risk (say in terms of the size of the primary tumour) and classify patients as either high risk (H) or low risk (L) and further suppose that the groups turn out as follows:

A: HHLHLHHHHLLHHHLHHLHHH (H=15, L=6)
B: LLHHLHHLLHLHLHLHHLLHLL (H=10, L=12)

Note that there are 15 patients (71 per cent) high risk and six (29 per cent) low risk patients in treatment group A compared to a split of 10 (45 per cent) high risk and 12 (55 per cent) low risk patients in treatment group B.

Now suppose that the mean survival times are observed to be 21.5 months in A and 27.8 months in group B. What conclusions can we draw? It is very difficult; the difference we have seen could be due to treatment differences or could be caused by the imbalance in terms of differential risk across the groups, or a mixture of the two. Statisticians talk in terms of *confounding* (just a fancy way of saying 'mixed up') between the treatment effect and the effect of baseline risk. This situation is very difficult to unravel and we avoid it by *stratified randomisation* to ensure that the 'case mix' in the treatment groups is comparable.

This simply means that we produce separate randomisation lists for the high risk and the low risk patients, the strata in this case. For example the following lists (which are block size four in each case):

H: ABBAAABBABABABABBBAAABBAABABBBAA
L: BAABBABAAABBBAABABABBBAABBAABAAB

will ensure firstly that we end up with balance in terms of group sizes but also secondly that both the high and low risk patients will be equally split across those groups, that is balance in terms of the mix of patients.

Having separate randomisation lists for the different centres in a multi-centre trial to ensure 'equal' numbers of patients in the treatment groups within each centre is using 'centre' as a stratification factor; this will ensure that we do not end up with treatment being confounded with centre.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'It is advisable to have a separate random scheme for each centre, i.e. to stratify by centre or to allocate several whole blocks to each centre. Stratification by important prognostic factors measured at baseline (e.g. severity of disease, age, sex, etc.) may sometimes be valuable in order to promote balanced allocation within strata . . . '

Where the requirement is to have balance in terms of several factors, a stratified randomisation scheme using all combinations of these factors to define the strata would ensure balance. For example if balance is required for sex and age, then a scheme with four strata:

- Males, < 50 years
- Females, < 50 years
- Males, ≥ 50 years
- Females, ≥ 50 years

will achieve the required balance.

1.4.5 Central randomisation

In *central randomisation* the randomisation process is controlled and managed from a centralised point of contact. Each investigator makes a telephone call through an Interactive Voice Response System (IVRS) to this centralised point when they have identified a patient to be entered into the study and is given the next allocation, taken from the appropriate randomisation list. Blind can be preserved by simply specifying the number of the (pre-numbered) pack to be used to treat the particular patient; the computerised system keeps a record of which packs have been used already and which packs contain which treatment. Central randomisation has a number of practical advantages:

- It can provide a check that the patient about to be entered satisfies certain inclusion/exclusion criteria thus reducing the number of protocol violations.
- It provides up-to-date information on all aspects of recruitment.
- It allows more efficient distribution and stock control of medication.
- It provides some protection against biased allocation of patients to treatment groups in trials where the investigator is not blind; the investigator knowing the next allocation could (perhaps subconsciously) select

patients to include or not include based on that knowledge; with central randomisation the patient is identified and information given to the system before the next allocation is revealed to them.

- It gives an effective way of managing multi-centre trials.
- It allows the implementation of more complex allocation schemes such as minimisation and dynamic allocation.

Earlier we discussed the use of stratified randomisation in multi-centre trials and where the centres are large this is appropriate. With small centres however, for example in GP trials, this does not make sense and a stratified randomisation with 'region' defining the strata may be more appropriate. Central randomisation would be essential to manage such a scheme.

Stratified randomisation with more than a small number of strata would be difficult to manage at the site level and the use of central randomisation is then almost mandatory.

1.4.6 Dynamic allocation and minimisation

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Dynamic allocation is an alternative procedure in which the allocation of treatment to a subject is influenced by the current balance of allocated treatments and, in a stratified trial, by the stratum to which the subject belongs and the balance within that stratum. Deterministic dynamic allocation procedures should be avoided and an appropriate element of randomisation should be incorporated for each treatment allocation.'

Dynamic allocation moves away from having a pre-specified randomisation list and the allocation of patients evolves as the trial proceeds. The method looks at the current balance, in terms of the mix of patients and a number of pre-specified factors, and allocates the next patient in an optimum way to help redress any imbalances that exist at that time.

For example, suppose we require balance in terms of sex and age (≥ 65 versus < 65) and part way through the trial we see a mix of patients as in Table 1.1.

Table 1.1 Current mix of patients

	A	B
Total	25	25
Male	12/25	10/25
Age ≥ 65	7/25	8/25

Treatment group A contains proportionately more males (12 out of 25 versus 10 out of 25) than treatment group B but fewer patients over 65 (7 out of 25 versus 8 out of 25). Further suppose that the next patient to enter is male and aged 68 years. In terms of sex we would prefer that this patient be placed in treatment group B while for age we would prefer this patient to enter in group A. The greater imbalance however is in relation to sex so our overall preference would be for treatment group A to help 'correct' for the current imbalance. The method of *minimisation* would simply put this patient in group B. ICH E9 however recommends that we have a 'random element' to that allocation and so for example we would allocate this patient to treatment group A with say probability 0.7. Minimisation is a special case of dynamic allocation where the random assignment probability (0.7 in the example) is equal to one. Of course with a small number of baseline factors, for example centre and two others, stratified randomisation will give good enough balance and there is no need to consider the more complex dynamic allocation. This technique, however, has been proposed when there are more factors involved.

Since the publication of ICH E9 there has been considerable debate about the validity of dynamic allocation, even with the random element. There is a school of thought which has some sympathy within regulatory circles that supports the view that the properties of standard statistical methodologies, notably *p*-values and confidence intervals, are not strictly valid when such allocation schemes are used. As a result regulators are very cautious:

CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'

'. . . techniques of dynamic allocation such as minimisation are sometimes used to achieve balance across several factors simultaneously. Even if deterministic schemes are avoided, such methods remain highly controversial. Thus applicants are strongly advised to avoid such methods.'

So if you are planning a trial then stick with stratification and avoid dynamic allocation. If you have an ongoing trial which is using dynamic allocation then continue, but be prepared at the statistical analysis stage to supplement the standard methods of calculating *p*-values with more complex methods which take account of the dynamic allocation scheme. These methods go under the name of *randomisation tests*.

See Roes (2004) for a comprehensive discussion of dynamic allocation.

1.4.7 Cluster randomisation

In some cases it can be more convenient or appropriate not to randomise individual patients, but to randomise groups of patients. The groups for example

could correspond to GPs so that each GP enters say four patients and it is the 100 GPs that are randomised, 50 giving treatment A and 50 giving treatment B. Such methods are used but are more suited to phase IV than the earlier phases of clinical development.

Bland (2004) provides a review and some examples of cluster randomised trials while Campbell, Donner and Klar (2007) give a comprehensive review of the methodology.

1.5 Bias and precision

When we are evaluating and comparing our treatments we are looking for two things:

- An unbiased, correct view of how effective (or safe) the treatment is
- An accurate estimate of how effective (or safe) the treatment is

As statisticians we talk in terms of *bias* and *precision*; we want to eliminate bias and to have high precision. Imagine having 10 attempts at hitting the bull's eye on a target board as shown in Figure 1.1. Bias is about hitting the bull's eye on average; precision is about being consistent.

These aspects are clearly set out in ICH E9.

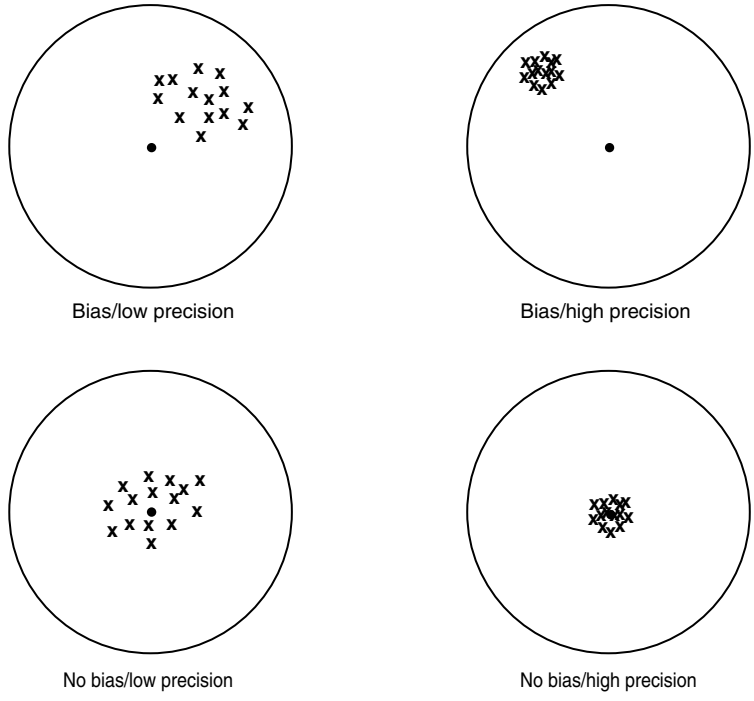


Figure 1.1 Bias and precision

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Many of the principles delineated in this guidance deal with minimising bias and maximising precision. As used in this guidance, the term "bias" describes the systematic tendency of any factors associated with the design, conduct, analysis and interpretation of the results of clinical trials to make the estimate of a treatment effect deviate from its true value.'

What particular features in the design of a trial help to eliminate bias?

- Concurrent control group as the basis for a 'comparison'
- Randomisation to avoid bias in allocating subjects to treatments
- Blinding of both the subject and the investigator
- Pre-specification of the methods of statistical analysis

What particular features in the design of a trial help to increase precision?

- Large sample size
- Measuring the endpoints in a precise way
- Standardising aspects of the protocol which impact on patient-to-patient variation
- Collecting data on key prognostic factors
- Choosing a homogeneous group of patients
- Choosing the most appropriate design (for example using a cross-over design rather than a parallel group design where this is appropriate)

Several of the issues raised here may be unclear at this point, simply be aware that eliminating bias and increasing precision are the key issues that drive our statistical thinking from a design perspective. Also be aware that if something should be sacrificed then it is precision rather than bias. High precision in the presence of bias is of no value. First and foremost we require an unbiased view; increasing precision is then a bonus. Similar considerations are also needed when we choose the appropriate statistical methodology at the analysis stage.

1.6 Between- and within-patient designs

The simplest trial design of course is the *parallel group design* assigning patients to receive either treatment A or treatment B. While this is a valid and effective design

it is important to recognise some inherent drawbacks. For example, suppose we have a randomised parallel group design in hypertension with 50 patients per group and that the mean fall in diastolic blood pressure in each of the two groups is as follows:

$$\begin{aligned} \text{A}: &\quad \bar{x}_1 = 4.6\,\text{mmHg} \\ \text{B}: &\quad \bar{x}_2 = 7.1\,\text{mmHg} \end{aligned}$$

It would be easy simply to conclude in light of the data that B is a more effective treatment than A, but is that necessarily the case? One thing we have to remember is that the 50 patients in group A are a different group of patients from the 50 in group B and patients respond differently, so in fact the observed difference between the treatments could simply be caused by patient-to-patient variation.

As we will see later, unravelling whether the observed difference is reflective of a real treatment difference or simply a chance difference caused by patient-to-patient variation with identical treatments is precisely the role of the *p*-value; but it is not easy.

This design is what we refer to as a *between-patient design*. The basis of the treatment comparison is the comparison between two independent groups of patients.

An alternative design is the *within-patient design*. Such designs are not universally applicable but can be very powerful under certain circumstances. One form of the within-patient design is the *paired design*:

- In ophthalmology; treatment A in the right eye, treatment B in the left eye
- In a volunteer study in wound care; 'create' a wound on each forearm and use dressing of type A on the right forearm and dressing of type B on the left forearm

Here the 50 subjects receiving A will be the same 50 subjects who receive B and the comparison of A and B in terms of say mean healing time in the second example is a comparison based on identical 'groups' of subjects. At least in principle, drawing conclusions regarding the relative effect of the two treatments and accounting for the patient-to-patient variation may be easier under these circumstances.

Another example of the within-patient design is the *cross-over design*. Again each subject receives each of the treatments but now sequentially in time with some subjects receiving the treatments in the order A followed by B and some in the order B followed by A.

In both the paired design and the cross-over design there is, of course, randomisation; in the second paired design example above, it is according to which forearm receives A and which receives B and randomisation is to treatment order, A/B or B/A, in the cross-over design.

1.7 Cross-over trials

The cross-over trial was mentioned in the previous section as one example of a within-patient design. In order to discuss some issues associated with these designs we will consider the simplest form of cross-over trial; two treatments A and B and two treatment periods I and II.

The main problem with the use of this design is the possible presence of the so-called *carry-over effect*. This is the residual effect of one of the treatments in period I influencing the outcome on the other treatment in period II. An extreme example of this would be the situation where one of the treatments, say A, was very efficacious, so much so that many of the patients receiving treatment A were cured of their disease, while B was ineffective and had no impact on the underlying disease. As a consequence many of the subjects following the A/B sequence would give a good response at the end of period I (an outcome ascribed to A) but would also give a good response at the end of period II (an outcome ascribed to B) because they were cured by A. These data would give a false impression of the A versus B difference. In this situation the B data obtained from period II is contaminated and the data coming out of such a trial are virtually useless.

It is important therefore to only use these designs when you can be sure that carry-over effects will not be seen. Introducing a washout period between period I and period II can help to eliminate carry-over so that when the subject enters period II their disease condition is similar to what it was at the start of period I. Cross-over designs should not be used where there is the potential to affect the underlying disease state. ICH E9 is very clear on the use of these designs.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Crossover designs have a number of problems that can invalidate their results. The chief difficulty concerns carryover, that is, the residual influence of treatments in subsequent treatment periods . . . When the crossover design is used it is therefore important to avoid carryover. This is best done by selective and careful use of the design on the basis of adequate knowledge of both the disease area and the new medication. The disease under study should be chronic and stable. The relevant effects of the medication should develop fully within the treatment period. The washout periods should be sufficiently long for complete reversibility of drug effect. The fact that these conditions are likely to be met should be established in advance of the trial by means of prior information and data.'

The cross-over design is used extensively in phase I trials in healthy volunteers to compare different formulations in terms of their bioequivalence (where there is no underlying disease to affect). They can also be considered in diseases, for

example asthma, where the treatments are being used simply to relieve symptoms; once the treatments are removed the symptoms return to their earlier level.

1.8 Signal and noise

1.8.1 Signal

Consider the example in Section 1.6 comparing treatments A and B in a parallel group trial. The purpose of this investigation is to detect differences in the mean reductions in diastolic blood pressure between the two groups. The observed difference between $\bar{x}_1 = 4.6\,\text{mmHg}$ and $\bar{x}_2 = 7.1\,\text{mmHg}$ is $2.5\,\text{mmHg}$. We will refer to this difference as the *signal* and this captures in part the evidence that the treatments truly are different. Clearly, if the observed difference was larger then we would likely be more inclined to conclude differences. Large differences give strong signals while small differences give weak signals.

1.8.2 Noise

The signal, however, is not the only aspect of the data that plays a part in our conclusions. A difference of 2.5 mmHg based on group sizes of 100 is much stronger evidence than a difference of 2.5 mmHg based on groups of size 10. Further, if we were to see a large amount of patient-to-patient variation then we would be less inclined to conclude differences than if all the patients in treatment group A had reductions tightly clustered around 4.6 mmHg while those in treatment group B had values tightly clustered around 7.1 mmHg. As can be seen in Figure 1.2 the evidence for a real treatment difference in situation I is much stronger than the evidence seen in situation II although the mean values for both groups are actually the same in both cases.

The sample size and patient-to-patient variability are the key elements of the *noise*. A small sample size and a large amount of patient-to-patient variability contribute to a large amount of noise. Both increasing the sample size and reducing the patient-to-patient variability will have the effect of reducing the noise and make it much easier to conclude that the treatments are different.

1.8.3 Signal-to-noise ratio

These concepts of signal and noise are taken from the field of engineering, but provide a way of thinking for statistical experiments. In declaring differences we look for strong signals and small amounts of noise, that is a large *signal-to-noise ratio*. If either the signal is weak or the noise is large, or both, then this ratio will be small and we will have little evidence on which to 'declare' differences.

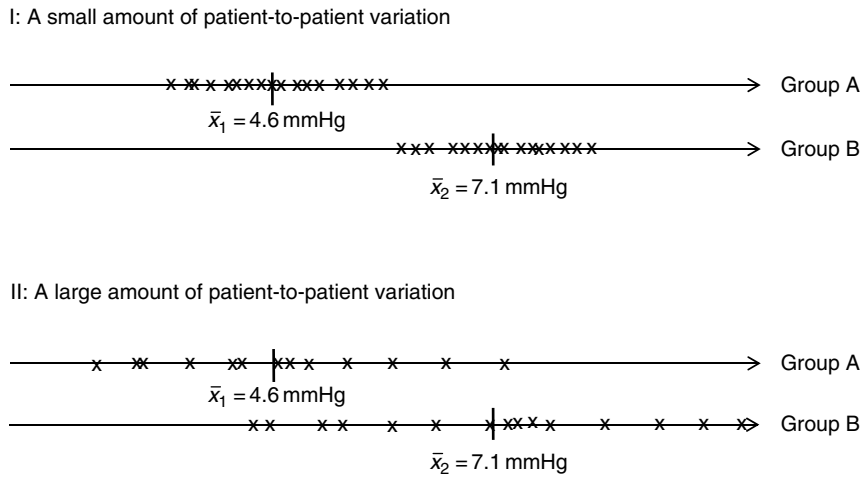


Figure 1.2 Differing degrees of patient-to-patient variation

In one sense the signal is out of our control, it will depend entirely on what the true treatment difference is. Similarly there is little we can do about the patient-to-patient variability, although we can reduce this by having, for example, precise measures of outcome or a more homogeneous group of patients. The sample size however is very much under our control and common sense tells us that increasing this will provide a more reliable comparison and make it easier for us to detect treatment differences when they exist.

Later, in Chapter 8, we will discuss power and sample size and see how to choose sample size in order to meet our objectives. We will also see in Section 3.3 how, in many circumstances, the calculation of the *p*-value is directly based on the signal-to-noise ratio.

1.9 Confirmatory and exploratory trials

ICH E9 makes a very clear distinction between *confirmatory* and *exploratory* trials. From a statistical perspective this is an important distinction as certain aspects of the design and analysis of data depend upon this confirmatory/exploratory distinction.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'A confirmatory trial is an adequately controlled trial in which the hypotheses are stated in advance and evaluated. As a rule, confirmatory trials are needed to provide firm evidence of efficacy or safety.'

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'The rationale and design of confirmatory trials nearly always rests on earlier clinical work carried out in a series of exploratory studies. Like all clinical trials, these exploratory studies should have clear and precise objectives. However, in contrast to confirmatory trials, their objectives may not always lead to simple tests of pre-defined hypotheses.'

Typically, later phase trials tend to have the confirmatory elements while the earlier phase studies; proof of concept, dose-finding etc. are viewed as exploratory. Indeed an alternative word for confirmatory is pivotal. It is the confirmatory elements of our trials that provide the pivotal information from a regulatory perspective.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Any individual trial may have both confirmatory and exploratory aspects.'

Usually it is the primary endpoint(s) that provides the basis of the confirmatory claims and exploratory aspects are relegated to the secondary (or exploratory) endpoints.

1.10 Superiority, equivalence and non-inferiority trials

A clear distinction should also be made between superiority, equivalence and non-inferiority trials.

In a *superiority* trial our objective is to demonstrate either that our treatment works by demonstrating superiority over placebo or that we are superior to some reference or standard control treatment.

In an *equivalence* trial we are looking to show that we are similar to some reference treatment, bioequivalence trials are the most common examples of this type of trial.

Finally, in a *non-inferiority* trial we are trying to demonstrate that we are 'at least as good as' or 'not worse than' some active control reference treatment in some pre-defined sense.

In therapeutic equivalence trials and in non-inferiority trials we are often looking to demonstrate efficacy of our test treatment indirectly. It may be that for ethical or practical reasons it is not feasible to show efficacy by undertaking a superiority trial against placebo. In such a case we compare our test treatment to a control treatment that is known to be efficacious and demonstrate either strict

equivalence or 'at least as good as' (non-inferiority). If we are successful then we can be confident that our test treatment works.

Alternatively there may be commercial reasons why we want to demonstrate the non-inferiority of our treatment against an active control. Maybe our treatment potentially has fewer side effects than the active control and we are prepared to pay a small price for this safety advantage in relation to efficacy. If this were the case then of course we would need to show advantages in terms of a reduction in side effects but we would also need to demonstrate that we do not lose much with regard to efficacy.

Non-inferiority trials are becoming more and more common as time goes on. This in part is due to the constraints imposed by the revised Helsinki Declaration (2004) and the increasing concern in some circles regarding the ethics of placebo use. These trials however require very careful design and conduct and we will discuss this whole area in a subsequent chapter.

1.11 Data types

It is useful to classify the types of data that we see in our clinical investigations.

The most common kind of data that we see is *continuous* data. Examples include cholesterol level, exercise duration, blood pressure, FEV_1 and so on. Each of these quantities is based on a continuum of potential values. In some cases, of course, our measurement technique may only enable us to record to the nearest whole number (blood pressure for example), but that does not alter the basic fact that the underlying scale is continuous.

Probably the second most common data type is *binary*. Examples of binary data include cured/not cured, responder/non-responder, died/survived. Here the measure is based on a dichotomy.

Moving up from binary is *categorical* data where there are more than two categories that form the basis of the 'measurement'. The following are examples of categorical variables:

- Death from cancer causes/death from cardiovascular causes/death from respiratory causes/death from other causes/survival
- Pain: none/mild/moderate/severe/very severe

The categories are non-overlapping and each patient is placed into one and only one of the outcome categories. Binary data is a special case where the number of categories is just two.

These examples are also different in another regard; in the first example the categories are unordered while in the second example there is a complete ordering

across the defined categories. In the latter case we term the data type either *ordered categorical* or *ordinal*.

Ordinal data arises in many situations. In oncology (solid tumours) the RECIST criteria record outcome in one of 4 response categories (National Cancer Institute, www.cancer.gov):

- CR (complete response) = disappearance of all target lesions
- PR (partial response) = 30 per cent decrease in the sum of the longest diameter of target lesions
- PD (progressive disease) = 20 per cent increase in the sum of the longest diameter of target lesions
- SD (stable disease) = small changes that do not meet the above criteria

When analysing data it is important of course that we clearly specify the appropriate order and in this case it is CR, PR, SD, PD.

Other data arise as *scores*. These are frequently as a result of the need to provide a measure of some clinical condition such as depression or anxiety. The Hamilton Depression (HAM-D) scale and the Hamilton Anxiety (HAM-A) scale provide measures in these cases. These scales contain distinct items which are scored individually and then the total score is obtained as the sum of the individual scores. For the Hamilton Depression scale there are usually 17 items; depressed mood, self-depreciation and guilt feelings, etc., each scored on a three-point scale or on a five-point scale. The five-point scales are typically scores 0 = absent, 1 = doubtful to mild, 2 = mild to moderate, 3 = moderate to severe and 4 = very severe while the three-point scales are typically 0 = absent, 1 = probable or mild and 3 = definite.

Finally, data can arise as *counts* of items or events; number of epileptic seizures in a 12-month period, number of asthma attacks in a three-month period and number of lesions are just a few examples.

As we shall see later the data type to a large extent determines the class of statistical tests that we undertake. Commonly for continuous data we use the t-tests and their extensions; analysis of variance and analysis of covariance. For binary, categorical and ordinal data we use the class of chi-square tests (Pearson chi-square for categorical data and the Mantel-Haenszel chi-square for ordinal data) and their extension, logistic regression.

Note also that we can move between data types depending on the circumstances. In hypertension we might be interested in:

- the fall in diastolic blood pressure (continuous) or
- success/failure with success defined as a reduction of at least 10 mm Hg in diastolic blood pressure and diastolic below 90 mmHg (binary) or

- complete success/partial success/failure with complete success = reduction of at least 10 mmHg and diastolic below 90 mmHg, partial success = reduction of at least 10 mmHg but diastolic 90 mmHg or above and failure = everything else (ordinal)

There are further links across the data types. For example, from time to time we group continuous, score or count data into ordered categories and analyse using techniques for ordinal data. For example, in a smoking cessation study we may reduce the basic data on cigarette consumption to just four groups (Table 1.2). accepting that there is little reliable information beyond that.

Table 1.2 Categorisation

Group	Cigarettes per day
1	0
2	1–5
3	6–20
4	> 20

We will continue this discussion in the next section on endpoints.

1.12 Choice of endpoint

1.12.1 Primary variables

Choosing a single primary endpoint is part of a strategy to reduce multiplicity in statistical testing. We will leave discussion of the problems arising with multiplicity until Chapter 10 and focus here on the nature of endpoints both from a statistical and a clinical point of view.

Generally the primary endpoint should be that endpoint that is the clinically most relevant endpoint from the patients' perspective.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'The primary variable ("target" variable, primary endpoint) should be that variable capable of providing the most clinically relevant and convincing evidence directly related to the primary objective of the trial.'

This choice should allow, amongst other things, a clear quantitative measure of benefit at the individual patient level. As we will see, identifying new treatments is not just about statistical significance, it is also about clinical importance and the

importance of the clinical finding can only ever be evaluated if we can quantify the clinical benefit for patients.

Usually the primary variable will relate to efficacy, but not always. If the primary objective of the trial concerns safety or quality of life then a primary variable(s) relating to these issues would be needed.

The primary endpoint should not be confused with a summary measure of the benefit. For example, the primary endpoint may be a binary endpoint, survival beyond two years/death within two years, while the primary evaluation is based upon a comparison of two year survival rates between two treatments. The primary endpoint *is not* the proportion surviving two years, it is binary outcome survival beyond two years/death within two years, the variable measured at the patient level.

The primary endpoint must be pre-specified in a confirmatory trial as specification after unblinding could clearly lead to bias. Generally, there would be only one primary endpoint, but in some circumstances more than one primary endpoint may be needed in order to study the different effects of a new treatment. For example in acute stroke it is generally accepted that two primary endpoints are used; one relating to survival, free of disability and a second relating to improvement in neurological outcome. See CPMP (2001) 'Note for Guidance on Clinical Investigation of Medicinal Products for the Treatment of Acute Stroke' for further details on this.

1.12.2 Secondary variables

Secondary variables may be defined which support a more detailed evaluation of the primary endpoints or alternatively such endpoints may relate to secondary objectives. These variables may not be critical to a claim but may help in understanding the nature of the way the treatment works. In addition, data on secondary endpoints may help to embellish a marketing position for the new treatment.

If the primary endpoint gives a negative result then the secondary endpoints cannot generally recover a claim. If, however, the primary endpoint has given a positive result, then additional claims can be based on the secondary endpoints provided these have been structured correctly within the confirmatory strategy. In Chapter 10 we will discuss hierarchical testing as a basis for such a strategy.

1.12.3 Surrogate variables

Surrogate endpoints are usually used when it is not possible within the timeframe of the trial to measure true clinical benefit. Many examples exist as seen in Table 1.3.

Unfortunately many treatments which have shown promise in terms of surrogate endpoints have been shown not to provide subsequent improvement in terms

Table 1.3 Surrogate variable and clinical endpoints

Disease	Surrogate variable	Clinical endpoint
congestive heart failure	exercise tolerance	mortality
osteoporosis	bone mineral density	fractures
HIV	CD4 cell count	mortality
hypercholesterolemia	cholesterol level	coronary heart disease

of the clinical outcome. Fleming and DeMetss (1996) provide a number of examples where we have been disappointed by surrogate endpoints and provide in each of these cases possible explanations for this failure of the surrogate. One common issue in particular is that a treatment may have an effect on a surrogate through a particular pathway that is unrelated to the underlying disease process or the clinical outcome.

Example 1.2: Bone Mineral Density (BMD) and Fracture Risk in Osteoporosis

Li, Chines and Meredith (2004) quote three clinical trials evaluating the effectiveness of alendronate, risedronate and raloxifene in increasing BMD and reducing fracture risk in osteoporosis. These treatments are seen to reduce fracture risk by similar amounts; 47 per cent, 49 per cent and 46 per cent respectively, yet their effects on increasing BMD are somewhat different; 6.2 per cent, 5.8 per cent and 2.7 per cent respectively. Drawing conclusions on the relative effectiveness of these treatments based solely in terms of the surrogate BMD would clearly be misleading.

Treatment effects on surrogate endpoints therefore do not necessarily translate into treatment effects on clinical endpoints and the validity of the surrogate depends not only on the variable itself but also on the disease area and the mode of action of the treatment. Establishing new valid surrogates is very difficult. Fleming and DeMets conclude that surrogates are extremely valuable in phase II 'proof of concept' studies but they question their general use in phase III confirmatory trials.

1.12.4 Global assessment variables

Global assessment variables involve an investigator's overall impression of improvement or benefit. Usually this is done in terms of an ordinal scale of categories. While the guidelines allow such variables, experience shows that they

must at the very least be accompanied by objective measures of benefit. Indeed both the FDA and the European regulators tend to prefer the use of the objective measures only, certainly at the primary variable level.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'If objective variables are considered by the investigator when making a global assessment, then those objective variables should be considered as additional primary, or at least important secondary, variables.'

1.12.5 Composite variables

In some circumstances it may be necessary to combine several events/endpoints to produce a combined or composite endpoint. The main purpose in doing this is to again avoid multiple testing and more will be said about this in Chapter 10. In addition, combining endpoints/events will increase the absolute numbers of events observed and this can increase sensitivity for the detection of treatment effects.

1.12.6 Categorisation

In general, a variable measured on a continuous scale contains more information and is a better reflection of the effect of treatment than a categorisation of such a scale. For example, in hypertension the clinical goal may be to reduce diastolic blood to below 90 mmHg; that is not to say that a reduction down to 91 mmHg is totally unacceptable while a reduction down to 89 mmHg is a perfect outcome. Having a binary outcome which relates to achieving 90 mmHg is clearly only a very crude measure of treatment benefit.

Nonetheless, categorisation can be of benefit under some circumstances. In an earlier section we discussed the categorisation of number of cigarettes to a four-point ordinal scale, accepting that measures on the original scale may be subject to substantial error and misreporting; the additional information contained in the number of cigarettes smoked is in a sense spurious precision.

There may also be circumstances where a categorisation combines responses measured, on different domains of action for example to give a single dichotomous responder/non-responder outcome. There are connections here with global assessment variables. This approach is taken in Alzheimer's disease where the effect of treatment is ultimately expressed in terms of the 'proportion of patients who achieve a meaningful benefit (response)'; see the CPMP (1997) 'Note for Guidance on Medicinal Products in the Treatment of Alzheimer's Disease'. In oncology,

the RECIST criteria may be used simply to give the proportion of patients who achieve a CR or PR. This reduces the sensitivity of the complete scale but may make it easier to quantify the clinical benefit. For an interesting exchange on the value of dichotomisation see Senn (2003) and Lewis (2004).

Finally, a few words about the use of the Visual Analogue Scale (VAS). A value on this 10 mm line gives a continuous measure (the distance between the left-hand end and the marked value) and these are used successfully in a number of therapeutic settings. Their advantage over an ordinal four- or five-point scale, however, is questionable as again there is an argument that the additional 'precision' provided by VAS scales is of no value. A study by Jensen *et al.* (1989) in the measurement of post-operative pain showed that information relating to pain was best captured using an 11-point scoring scale (0,1, 2,. . . ,10) – sometimes referred to as a *Likert scale*, or a verbal rating scale with five points; mild, discomforting, distressing, horrible, excruciating. In addition around 10 per cent of the patients were unable to understand the requirement for completion of the VAS for pain. These ordered categorical scales may well be as precise, or more precise than the VAS and at the same time prove to be more effective because patients understand them better.

2

Sampling and inferential statistics

2.1 Sample and population

Consider the comparison of a new treatment A to an existing treatment B for lowering blood pressure in mild to moderate hypertension in the context of a clinical trial conducted across Europe. The characteristics of the *population* of mild to moderate hypertensive patients to be studied will be defined by the inclusion (and exclusion) criteria and may well contain several millions of individuals. In another sense this population will be infinite if we also include those patients satisfying the same inclusion/exclusion criteria in the future. Our clinical trial will, for example, also involve selecting a *sample* of say 200 individuals from this population and randomly assigning 100 to treatment group A and 100 to the treatment group B.

Each subject in the sample will be observed and provide a value for the fall in diastolic blood pressure, the primary endpoint. The mean falls in blood pressure in groups A and B will then be computed and compared. Suppose that:

$$\overline{x}_1 = 8.6\,\text{mmHg}$$

$$\overline{x}_2 = 3.9\,\text{mmHg}$$

The conclusion we draw will be based on this comparison of the means and in general there are three possibilities in relation to what we conclude:

- Treatment A is better than treatment B
- Treatment B is better than treatment A
- There are no differences

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

Suppose in this case we conclude on the basis of the data that treatment A is better than treatment B. This statement of course is correct in terms of what we have seen on average in the sample data, but the statement we are making is in fact stronger than that; it is a statement about the complete population. On the basis of the data we are concluding that treatment A will, on average, work better than treatment B in the complete population; we are extrapolating from the sample to the population. Statisticians talk in terms of making *inferences*. On the basis of the sample data we are inferring things about the complete population.

In one sense moving from the sample to the population in this way is a leap of faith! However, it should work providing the sample is representative of the complete population. If it is not representative, but if we can assume that the treatment difference is homogeneous across the population as a whole, then we would still obtain a valid estimate of that treatment difference. It is the randomisation that protects us in this regard.

In order to make any progress in understanding how inferential statistics works we need to understand what happens when we take samples from populations. In a later section we will explore this through a computer simulation and see how all of this comes together in practical applications.

2.2 Sample statistics and population parameters

2.2.1 Sample and population distribution

The *sample histogram* in Figure 2.1 provides a visual summary of the distribution of total cholesterol in a group of 100 patients at baseline (artificial data). The x-axis is divided up into intervals of width 0.5 mmol/l and the y-axis counts the number of individuals with values within those intervals.

We will sometimes refer to the sample histogram as the *sample distribution*.

These data form the sample and they have been taken from a well-defined population, which sits in the background. We can also envisage a corresponding histogram for the complete population and this will have a smooth shape as a result of the size of that population; we use the terms *population histogram* or *population distribution*. Figure 2.2 shows the population histogram superimposed onto the sample histogram. Providing the sample is representative of the population then the sample and population histograms should be similar. In practice remember that we only see the sample histogram, the population histogram is hidden to us; indeed we want to use the sample distribution to tell us about the distribution in the population.

There are usually two aspects of these histograms that we are interested in, what is happening on average and the patient-to-patient variation or spread of the data values. The average will be used as a basis for measuring the signal and

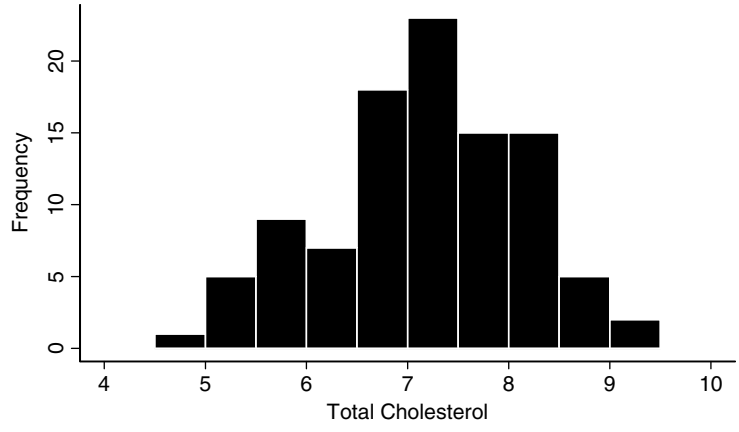


Figure 2.1 Histogram for total cholesterol ($n = 100$)

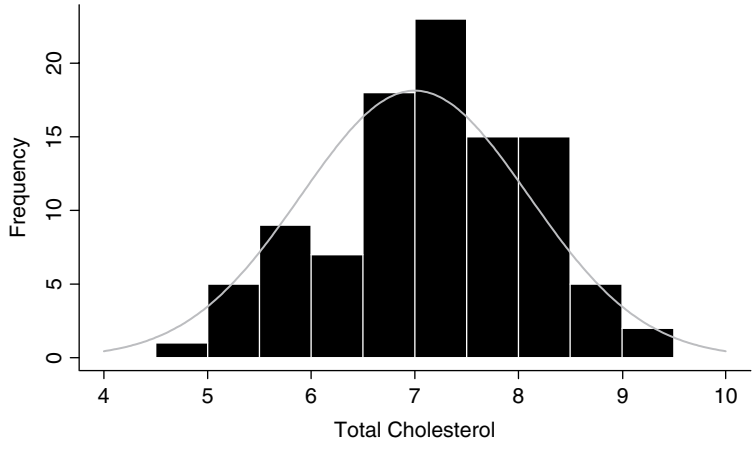


Figure 2.2 Sample histogram ($n = 100$) and population histogram

in particular we will be looking at differences between two averages as the signal, while the patient-to-patient variation is a key element of the noise as discussed earlier in Section 1.8.2.

The measures of average that we commonly use are the mean and the median while the standard deviation provides us with our measure of patient-to-patient variation.

2.2.2 Median and mean

The *median* (denoted by $\tilde{x}$) is the middle value when the data points are ordered from smallest to largest. The median can only be defined in this way when there

are an odd number of values (subjects). When the number of subjects is even, we define the median to be the average of the middle two values. For the data in Figure 2.1, $n = 100$ and the median $\tilde{x} = 7.20$ mmol/l. The *mean* (denoted by $\bar{x}$) is the arithmetic average; $\bar{x} = \frac{1}{n}\sum x$. For the data in Figure 2.1, $n = 100$ and the mean $\bar{x} = 7.16$ mmol/l.

2.2.3 Standard deviation

The *standard deviation* (denoted *s* or *sd*) is a measure of patient-to-patient variation. There are other potential measures but this quantity is used as it possesses a number of desirable mathematical properties and appropriately captures the overall amount of variability within the group of patients.

It is related to another quantity called the *variance* and

$$\text{variance} = (\text{standard deviation})^2$$

So if the standard deviation is 3, the variance is 9, if the variance is 25 then the standard deviation is 5.

The method of calculation of the standard deviation seems, at least at face value, to have some arbitrary elements to it. There are several steps:

- Calculate the mean of all values.
- Calculate the difference between each individual value and the mean and square each of those differences.
- Take the average of these squared differences, but with the 'average' calculated by dividing by $n - 1$, not $n -$ the resulting quantity is called the variance (with units mmol/l^2 for the data in our example).
- Take the square root of the variance to revert to the original units, mmol/l; this is the standard deviation.

For the example data, $n = 100$ and the standard deviation $s = 0.98$ mmol/l

People often ask; why divide by $n - 1$ rather than n? Well, the answer is a fairly technical one. It can be shown mathematically that dividing by n gives a quantity that, on average, underestimates the true standard deviation, particularly in small samples and dividing by $n - 1$ rather than n corrects for this underestimation. Of course for a large sample size it makes very little difference, dividing by 99 is much the same as dividing by 100.

Another frequent question is; why square, average and then square root, why not simply take the average distance of each point from the mean without bothering about the squaring? Well you could do this and yes you would end up with

a measure of patient-to-patient variability; this quantity is actually referred to as the *mean absolute deviation* and is indeed sometimes used as a measure of spread. The standard deviation, however, has several strong theoretical properties that we will need in our subsequent development and so we will go with that as our measure of variation.

2.2.4 Notation

In order to distinguish between quantities measured in the sample and corresponding quantities in the population we use different symbols:

The mean in the sample is denoted $\bar{x}$
The mean in the population is denoted μ

The standard deviation in the sample is denoted s or sd
The standard deviation in the population is denoted σ

Remember, $\bar{x}$ and s are quantities that we calculate from our data while μ and σ are theoretical quantities (parameters) that are unknown to us but nonetheless exist in the context of the broader population from which the sample (and therefore the data) is taken. If we had access to every single subject in the population then yes we could compute μ but this is never going to be the case. We can also think of μ and σ as the 'true' mean and 'true' standard deviation respectively in the complete population.

The calculation of mean and standard deviation only really makes sense when we are dealing with continuous, score or count data. These quantities have little relevance when we are looking at binary or ordinal data. In these situations we would tend to use proportions in the various categories as our summary statistics and population parameters of interest.

For binary data:

The sample proportion is denoted r

The population proportion is denoted θ

2.3 The normal distribution

The *normal* or *Gaussian* distribution was in fact first discovered by de Moivre, a French mathematician, in 1733. Gauss came upon it somewhat later, just after 1800, but from a completely different start point. Nonetheless, it is Gauss who has his name attached to this distribution.

The normal distribution is a particular form for a population histogram. It is symmetric around the mean μ and is bell-shaped. It has been noted empirically

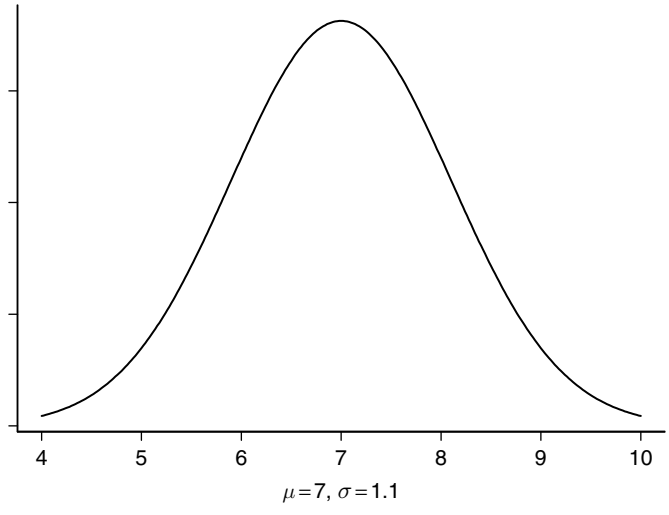


Figure 2.3 Normal distribution for total cholesterol at baseline ($\mu = 7$ mmol/l, $\sigma = 1.1$ mmol/l)

that in many situations, data, particularly when collected from random systems, gives a histogram with this 'normal distribution' shape. In fact there is a very powerful theorem called the 'Central Limit Theorem', which looks at the behaviour of data and says that under certain general conditions data behave according to this distribution. An example of a normal distribution is given in Figure 2.3.

As a consequence of both the empirical evidence and the theoretical base we often assume that the data we are collecting have been drawn from a distribution with the normal shape; we assume that our data are *normally distributed*.

One further point relating to the normal distribution in the population is that, because of the symmetry, the median and the mean take the same value. This is in fact a property of any population distribution for data which are symmetric.

Briefly returning to Gauss, one can gauge the importance of 'his' discovery by observing the old German 10-Mark banknote in Figure 2.4. Here we have Gauss and just above the '10' and to the right we can also see the normal distribution and its mathematical equation.

When the population does indeed follow this distribution then the standard deviation, σ, has a more specific interpretation. If we move σ units below the mean, to $\mu - \sigma$ and σ units above the mean, to $\mu + \sigma$, then that interval ($\mu - \sigma$, $\mu + \sigma$) will capture 68.3 per cent of the population values. This is true whatever we are considering; diastolic blood pressure, fall in diastolic blood pressure over a six-month period, cholesterol level, FEV_1 etc. and whatever the values of μ and σ; in all cases 68.3 per cent of the patients will have data values in the range $\mu - \sigma$ to $\mu + \sigma$ providing the data are normally distributed.

Several further properties hold as shown in Table 2.1.

Figure 2.4 Gauss on the Deutsche 10 Mark note

Table 2.1 Probabilities for the normal distribution

Range	Percentage of patients
$\mu - 2\sigma$ to $\mu + 2\sigma$	95.4 per cent
$\mu - 3\sigma$ to $\mu + 3\sigma$	99.7 per cent
$\mu - 1.645\sigma$ to $\mu + 1.645\sigma$	90 per cent
$\mu - 1.960\sigma$ to $\mu + 1.960\sigma$	95 per cent
$\mu - 2.576\sigma$ to $\mu + 2.576\sigma$	99 per cent

Note that the normal distribution curve has a mathematical equation and integrating the equation of this curve, for example between $\mu - 2\sigma$ and $\mu + 2\sigma$, irrespective of the values of μ and σ, will always give the answer 0.954. So 95.4 per cent of the area under the normal curve is contained between $\mu - 2\sigma$ and $\mu + 2\sigma$ and it is this area calculation that also tells us that 95.4 per cent of the individuals within the population will have data values in that range.

Example 2.1: Normal distribution (Figure 2.3)

A population of patients in a cholesterol lowering study have total cholesterol measured at baseline. Assume that total cholesterol is normally distributed with mean, $\mu = 7.0$ mmol/l and standard deviation, $s = 1.1$ mmol/l so that the variance is 1.21 ($= 1.1^2$). We write this as N(7.0, 1.21). For historical reasons we put the variance as the second parameter here. Under these assumptions the following results hold:

- 68.3 per cent of the patients have total cholesterol in the range 5.9 mmol/l to 8.1 mmol/l.

Example 2.1: (Continued)

- 90 per cent of the patients will have values in the range 5.19 mmol/l to 8.81 mmol/l.
- 95.4 per cent of the patients will have values in the range 4.8 mmol/l to 9.2 mmol/l.

2.4 Sampling and the standard error of the mean

Earlier in this chapter we spoke about the essence of inferential statistics; drawing conclusions about populations based upon samples taken from those populations. In order to understand how we do this we need to understand what happens when we take a sample from the population:

- Do we always reach the correct conclusion about the population?
- Are we sometimes misled by the sample data?
- How big a sample do we need to be confident that we will end up with a correct conclusion?

In order to gain an understanding of the sampling process we have undertaken a computer simulation. For this simulation we have set up, on the computer, a very large population of patients whose diastolic blood pressures have been recorded. The population has been structured to be normally distributed with mean 80 mmHg and standard deviation 4 mmHg; N(80, 16) as shown in Figure 2.5.

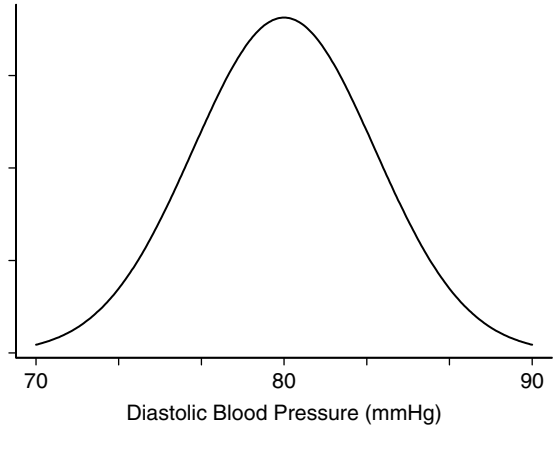


Figure 2.5 Normal distribution, N(80, 16)

Imagine that this is a real clinical trial setting and our objective is to find out the value of the mean diastolic blood pressure in the population (but remember because this is a computer simulation we know the answer!). So let's take a sample of size 50.

The mean, $\bar{x}$, from this sample turned out to be 80.218 mmHg. The one thing you will notice about this value is that it is not equal to μ, which is 80 mmHg, so we see immediately that the sampling process does not necessarily hit the truth. So let's take a second sample. The second sample gave a mean of 80.767 mmHg; again this value is not equal to the true mean. Not only that, the second sample has given a different answer to the first sample. We repeated this sampling process 100 times, going back to the population and taking further samples of size 50. The complete set of mean values is given in Table 2.2.

There are two things you will notice about this list of means. Firstly not one of them has hit the true mean value of 80 mmHg. Secondly all the values are different. The implication of this is as follows. Whenever you get an answer in a clinical trial the only thing you will know for certain is that it is the wrong answer! Not only that, if you were to repeat the trial under identical circumstances, same

Table 2.2 Mean values $\bar{x}$ from 100 samples of size 50 from N(80,16). Units are mmHg

1	80.218	26	79.894	51	79.308	76	80.821
2	80.767	27	79.629	52	79.620	77	80.498
3	78.985	28	80.233	53	80.012	78	79.579
4	81.580	29	79.738	54	80.678	79	81.051
5	79.799	30	80.358	55	80.185	80	80.786
6	80.302	31	79.617	56	79.901	81	79.780
7	79.094	32	79.784	57	79.778	82	79.802
8	80.660	33	79.099	58	79.597	83	80.510
9	79.455	34	79.779	59	79.320	84	79.592
10	79.275	35	81.438	60	79.076	85	79.617
11	80.732	36	80.066	61	80.580	86	79.587
12	79.713	37	79.591	62	79.878	87	79.124
13	79.314	38	80.254	63	79.656	88	79.520
14	80.010	39	80.494	64	79.302	89	79.587
15	79.481	40	79.259	65	80.242	90	79.544
16	79.674	41	80.452	66	78.344	91	80.054
17	80.049	42	80.957	67	80.653	92	80.458
18	79.156	43	80.113	68	79.848	93	79.895
19	80.826	44	80.043	69	80.294	94	79.293
20	80.321	45	80.436	70	80.797	95	79.376
21	79.476	46	81.220	71	79.226	96	80.296
22	80.155	47	79.391	72	78.883	97	79.722
23	79.429	48	80.552	73	79.871	98	78.464
24	80.775	49	80.422	74	80.291	99	78.695
25	80.490	50	80.265	75	79.544	100	79.692

protocol, same investigator and so on, but with a new set of patients then you would get a different answer. These are simply aspects of the sampling process; it is by no means a perfect process and we need to understand it more. This so-called *sampling variation* is fundamentally a result of patient-to-patient variation; patients behave differently and successive samples of 50 patients are going to give different results. In order to be able to work in this uncertain environment we need to quantify the extent of the sampling variation.

The standard deviation of this list of $\bar{x}$ values can be calculated using the method described earlier and gives a measure of the inherent variability in the sampling process. A large value for this standard deviation would indicate that the $\bar{x}$ values are all over the place and we are in an unreliable situation in terms of estimating where the true mean (μ) lies; a small value for this standard deviation would indicate that the $\bar{x}$ values are closely bunched and the sampling process is giving a consistent, reliable value. This standard deviation for the list of $\bar{x}$ values was calculated to be 0.626 and provides a measure of the variation inherent in the sampling process.

In practice we will never have the luxury of seeing the behaviour of the sampling process in this way; remember this is a computer simulation. However there is a way of estimating the standard deviation associated with the sampling process through a mathematical expression applied to the data from a single sample. This formula is given by $s/\sqrt{n}$, where s is the sd from the sample and n is the sample size.

So, in practice, we calculate the mean from a sample (size n) of data plus the corresponding standard deviation, s. We then divide the standard deviation by $\sqrt{n}$ and the resulting numerical value gives an estimate of the standard deviation associated with the sampling process, the standard deviation for the repeat $\bar{x}$ values had we undertaken the sampling many times.

Example 2.2: Sampling variation

In the first computer simulation, $n = 50$, $\bar{x} = 80.218$ and $s = 4.329$, the standard deviation of the 50 patient diastolic blood pressures in that sample.

The estimated standard deviation associated with the repeat $\bar{x}$ values is then given by:

$$\frac{s}{\sqrt{n}} = \frac{4.329}{\sqrt{50}} = 0.612$$

In other words, were we to repeat the sampling process, getting a list of $\bar{x}$ values by repeatedly going back to the population and sampling 50 subjects, then 0.612 gives us an estimate of the standard deviation associated with these $\bar{x}$ values.

One potentially confusing issue here is that there are two standard deviations; one measures the patient-to-patient variability from the single sample/trial while the second estimates the mean-to-mean variation that you would get by repeating the sampling exercise. To help distinguish the two, we reserve the term standard deviation for the first of these (patient-to-patient variation), and we call the second, *the standard error (*se*)* of $\bar{x}$ (mean-to-mean variation). Thus in the above example 0.612 is the standard error of $\bar{x}$ from the first computer simulation sample; an estimate of the standard deviation of those mean values under repeated sampling. Note that this value is close to 0.626, the 'standard error' calculated via the computer simulation through repetition of the sampling process.

Small standard errors tell us that we are in a reliable sampling situation where the repeat mean values are very likely to be closely bunched together; a large standard error tells us we are in an unreliable situation where the mean values are varying considerably.

It is not possible at this stage to say precisely what we mean by small and large in this context, we need the concept of the confidence interval to be able to say more in this regard and we will cover this topic in the next chapter. For the moment just look upon the standard error as an informal measure of precision; high values mean low precision and vice versa. Further if the standard error is small, it is likely that our estimate $\bar{x}$ is close to the true mean, μ. If the standard error is large, however, there is no guarantee that we will be close to the true mean.

Figure 2.6 shows histograms of $\bar{x}$ values for sample sizes of 20, 50 and 200 based on 100 simulations in each case. It is clear that for $n = 20$ there is considerable variation; there is no guarantee that the mean from a particular sample will be close to μ. For $n = 50$ things are not quite so bad, although the sample mean could still be out at 82 or at 78.2. For the sample size of 200 there is only a small amount of variability; over 250 of the 1000 mean values are within 0.1 units of the true mean. These histograms/distributions are referred to as *sampling distributions*. They are the distributions of $\bar{x}$ from the sampling process. Remember when you conduct a trial and get a mean value it is just one realisation of such a sampling process. The standard errors are the estimated standard deviations of the $\bar{x}$ values in these histograms and measure their spread.

2.5 Standard errors more generally

The standard error concept can be extended in relation to any *statistic* (i.e. quantity calculated from the data).

2.5.1 The standard error for the difference between two means

As a further example imagine a placebo-controlled cholesterol lowering trial. Generally in such trials patients in each of the groups will receive lifestyle

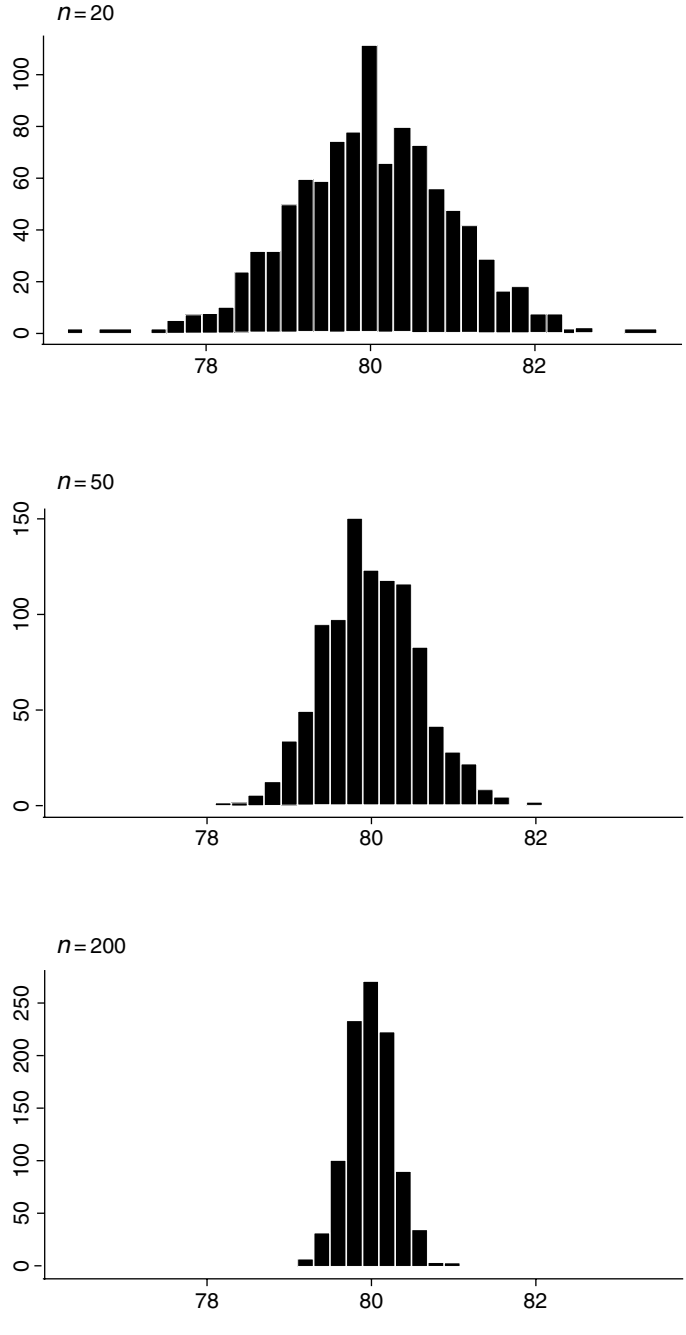


Figure 2.6 Sampling distribution of the mean $\bar{x}$; data from N(80,16), sample size n

and dietary advice plus medication, either active or placebo, according to the randomisation scheme. Let μ_1 be the true mean reduction in total cholesterol in the active treatment group and let μ_2 be the corresponding true mean reduction in the placebo group. So μ_1 is the mean reduction you would get if all patients in the population were given active treatment and μ_2 is the mean you would get if all patients were given placebo. The lifestyle and dietary advice will, of themselves, have a positive effect and coupled with the 'placebo effect' we will most likely see a mean reduction in each of the two treatment groups. The issue is, are we seeing a larger reduction in the active group compared to the placebo group. With this in mind our main interest lies, not in the individual means, but in their difference $\mu_1 - \mu_2$, the *treatment effect*.

Our best guess for the value of $\mu_1 - \mu_2$ is the observed difference in the sample means $\bar{x}_1 - \bar{x}_2$ from the trial.

Suppose that the value of $\bar{x}_1 - \bar{x}_2$ turns out to be 1.4 mmol/l. We know full well that this will not be equal to the true difference in the means, $\mu_1 - \mu_2$. We also know that if we were to repeat the trial under identical circumstances, same protocol, same investigator and so on, but of course a different sample of patients, then we would come up with a different value for $\bar{x}_1 - \bar{x}_2$.

So we need to have some measure of precision and reliability and this is provided by the standard error of $\bar{x}_1 - \bar{x}_2$. Again we have a formula for this:

$$\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\times\frac{(n_1-1)\,s_1^2+(n_2-1)\,s_2^2}{n_1+n_2-2}}$$

This expression allows us to estimate the standard deviation of the $\bar{x}_1 - \bar{x}_2$ values that we would get were we to repeat the trial.

Example 2.3: Standard error for the difference between two means

In a placebo controlled trial in cholesterol lowering we have the following (artificial) data (Table 2.3).

The standard error for the difference in the means, $\bar{x}_1 - \bar{x}_2$, is:

$$\sqrt{\left(\frac{1}{24}+\frac{1}{23}\right)\times\frac{(24-1)\times 0.92\times 0.92+(23-1)\times 1.05\times 1.05}{24+23-2}}=0.29$$

Table 2.3 Cholesterol lowering data

	n	Mean	*sd*
Active	24	2.7 mmol/l	0.92 mmol/l
Placebo	23	1.3 mmol/l	1.05 mmol/l

Small values of this standard error indicate high reliability; it is likely that the observed value, $\bar{x}_1 - \bar{x}_2$, for the treatment effect is close to the true treatment effect, $\mu_1 - \mu_2$. In contrast, a large value for the standard error tells us that $\bar{x}_1 - \bar{x}_2$ is not a reliable estimate of $\mu_1 - \mu_2$.

2.5.2 Standard errors for proportions

So far we have considered standard errors associated with means and differences between means. When dealing with binary data and proportions different formulas apply.

In Section 2.2.4 we let r denote a proportion in the sample, and θ the corresponding proportion in the population. For a single proportion r the standard error formula is $\sqrt{r(1-r)/n}$ where n is the number of subjects in the sample.

For the difference between two proportions, for example if we are looking at the difference $r_1 - r_2$ between the cure rate in the active group (group 1) and the cure rate in the placebo group (group 2), the standard error formula is $\sqrt{r_1(1-r_1)/n_1 + r_2(1-r_2)/n_2}$ where n_1 and n_2 are the numbers of subjects in groups 1 and 2 respectively.

2.5.3 The general setting

More generally, whatever statistic we are interested in, there is always a formula that allows us to calculate its standard error. The formulas change but their interpretation always remains the same; a small standard error is indicative of high precision, high reliability. Conversely a large standard error means that the observed value of the statistic is an unreliable estimate of the true (population) value. It is also always the case that the standard error is an estimate of the standard deviation of the list of repeat values of the statistic that we would get were we to repeat the sampling process, a measure of the inherent sampling variability.

As discussed in the previous section the standard error simply provides indirect information about reliability, it is not something we can use in any specific way, as yet, to tell us where the truth lies. We also have no way of saying what is large and what is small in standard error terms. We will, however, in the next chapter cover the concept of the confidence interval and we will see how this provides a methodology for making use of the standard error to enable us to make statements about where we think the true (population) value lies.

3

Confidence intervals and *p*-values

3.1 Confidence intervals for a single mean

3.1.1 The 95 per cent confidence interval

We have seen in the previous chapter that it is not possible to make a precise statement about the exact value of a population parameter, based on sample data, and that this is a consequence of the inherent sampling variation in the sampling process. The confidence interval provides us with a compromise; rather than trying to pin down precisely the value of the mean μ or the difference between two means $\mu_1 - \mu_2$, for example, we give a range of values, within which we are fairly certain that the true value lies.

We will first look at the way we calculate the confidence interval for a single mean μ and then talk about its interpretation. Later in this chapter we will extend the methodology to deal with $\mu_1 - \mu_2$ and other parameters of interest.

In the computer simulation in Chapter 2, the first sample ($n = 50$) gave data (to 2 decimal places) as follows:

$$\bar{x} = 80.22 \text{ mmHg and } s = 4.33 \text{ mmHg}$$

The lower end of the confidence interval, the *lower confidence limit*, is then given by:

$$\bar{x} - 1.96\frac{s}{\sqrt{n}} = 80.22 - \left(1.96 \times \frac{4.33}{\sqrt{50}}\right) = 79.02$$

The upper end of the confidence interval, the *upper confidence limit*, is given by:

$$\bar{x} + 1.96\frac{s}{\sqrt{n}} = 80.22 + \left(1.96 \times \frac{4.33}{\sqrt{50}}\right) = 81.42$$

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

The interval, (79.02, 81.42), then forms the *95 per cent confidence interval (CI)*.

These data arose from a computer simulation where, of course, we know that the true mean μ is 80 mmHg, so we can see that the method has worked, μ is contained within the range 79.02 to 81.42.

The second sample in the computer simulation gave the following data; $\bar{x} = 80.77$ mmHg and $s = 4.50$ mmHg and this results in the 95 per cent confidence interval as (79.52, 82.02). Again we see that the interval has captured the true mean.

Now look at all 100 samples taken from the normal population with $\mu = 80$ mmHg. Figure 3.1 shows the 95 per cent confidence intervals plotted for each of the 100 simulations. A horizontal line has also been placed at 80 mmHg to allow the confidence intervals to be judged in terms of capturing the true mean.

Most of the 95 per cent confidence intervals do contain the true mean of 80 mmHg, but not all. Sample number 4 gave a mean value $\bar{x} = 81.58$ mmHg with a 95 per cent confidence interval (80.33, 82.83), which has missed the true mean at the lower end. Similarly samples 35, 46, 66, 98 and 99 have given confidence intervals that do not contain $\mu = 80$ mmHg. So we have a method that seems to work most of the time, but not all of the time. For this simulation we have a 94 per cent (94/100) success rate. If we were to extend the simulation and take many thousands of samples from this population, constructing 95 per cent confidence intervals each time, we would in fact see a success rate of 95 per cent; exactly 95 per cent of those intervals would contain the true (population) mean value. This provides us with the interpretation of a 95 per cent confidence interval in

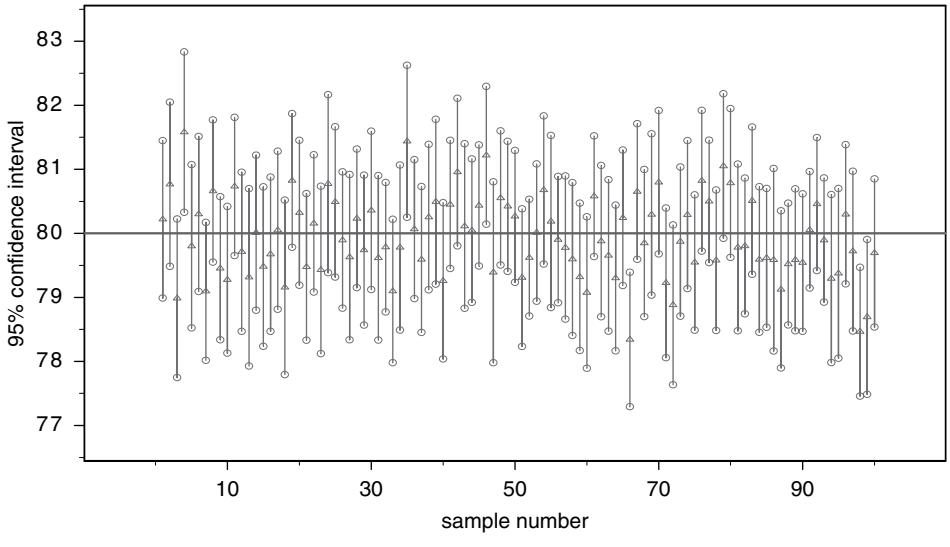


Figure 3.1 Computer simulation, 95 per cent confidence intervals, $n = 50$, mean = 80 mmHg

practice; when we construct a 95 per cent confidence interval from data then we can be 95 per cent certain that the true mean lies within the calculated range.

Why? – because 95 per cent of these intervals will indeed contain μ in the long run. Of course in any particular case we do not know whether our confidence interval is one of those 95 per cent or whether we have been unlucky and got one of the 5 per cent that do not contain the truth. In such a case we will have been misled by the data.

Just as an aside, look back at the formula for the 95 per cent confidence interval. Where does the 1.96 come from? It comes from the normal distribution; 1.96 is the number of standard deviations you need to move out to, to capture 95 per cent of the values in the population. The reason we get the so-called 95 per cent coverage for the confidence interval is directly linked to this property of the normal distribution.

3.1.2 Changing the confidence coefficient

We now have a procedure which allows us to make a statement about the value of μ with 95 per cent confidence, but we have to accept that such intervals will mislead us 5 per cent of the time. You may feel that this is too risky and instead request a 99 per cent confidence interval which will only mislead you 1 per cent of the time. That's fine, but the formula will change, and instead of using 1.96 to give 95 per cent coverage we will need to use 2.576 to give us 99 per cent coverage. The formula for the 99 per cent confidence interval is then:

$$\bar{x} - 2.576\frac{s}{\sqrt{n}} \text{ to } \bar{x} + 2.576\frac{s}{\sqrt{n}}$$

For the first sample in the computer simulation the 99 per cent confidence interval is (78.64, 81.80). This is a wider interval than the 95 per cent interval; the more confidence we require the more we have to hedge our bets. It is fairly standard to use 95 per cent confidence intervals and this links with the conventional use of 0.05 (or 5 per cent) for the cut-off for statistical significance. Under some circumstances we also use 90 per cent confidence intervals and we will mention one such situation later. In multiple testing it is also sometimes the case that we use *confidence coefficients* larger than 95 per cent, again we will discuss the circumstances where this might happen in a later chapter.

3.1.3 Changing the multiplying constant

The formula for the 95 per cent confidence interval (and also for the 99 per cent confidence interval) given above is in fact not quite correct. It is correct up to a

Table 3.1 Multiplying constants for calculating confidence intervals

Degrees of freedom (d.f.)	Confidence coefficient		
	90%	95%	99%
5	2.02	2.57	4.04
10	1.81	2.23	3.17
11	1.80	2.20	3.11
12	1.78	2.18	3.06
13	1.77	2.16	3.01
14	1.76	2.15	2.98
15	1.75	2.13	2.95
16	1.75	2.12	2.92
17	1.74	2.11	2.90
18	1.73	2.10	2.88
19	1.73	2.09	2.86
20	1.73	2.09	2.85
25	1.71	2.06	2.79
30	1.70	2.04	2.75
35	1.69	2.03	2.72
40	1.68	2.02	2.70
45	1.68	2.01	2.69
50	1.68	2.01	2.68
100	1.66	1.99	2.63
200	1.65	1.97	2.60
∞	1.645	1.960	2.576

point, in that it will work for very large sample sizes. For smaller sample sizes we need to change the multiplying constant according to the values in Table 3.1.

The reason for this is again a technical one but relates to the uncertainty associated with the use of the sample standard deviation (s) in place of the true population value (σ) in the formula for the standard error. When σ is known, the multiplying constants given earlier apply. When σ is not known (the usual case) we make the confidence intervals slightly wider in order to account for this uncertainty. When n is large of course s will be close to σ and so the earlier multiplying constants apply approximately.

Multiplying factors are given here for 90 per cent, 95 per cent and 99 per cent confidence intervals. Note that the constants 1.960 and 2.576, those used for 95 per cent and 99 per cent confidence intervals previously, appear at the foot of the columns. The column on the left hand side, termed degrees of freedom, is closely linked to sample size. When calculating a confidence interval for a mean, as in this section, we use the row corresponding to sample size $(= n) - 1$, so degrees of freedom for a single sample $= n - 1$. Do not agonise over the label, degrees of

freedom, just think in terms of getting the appropriate multiplying constant by going into the row; sample size −1.

A more complete table can be found in many standard statistics textbooks. Alternatively most statistics packages will contain a function that will give the multiplying constants for any value of degrees of freedom.

So if we were calculating a confidence interval for a mean μ from a sample of size 16 then we would look in row 15 for the multiplying constant and use 2.13 in place of 1.960 in the calculation of the 95 per cent confidence interval and 2.95 in place of 2.576 for the 99 per cent confidence interval.

Example 3.1: Confidence intervals for a single mean

In an asthma trial comparing two short acting treatments, the following (hypothetical) data were obtained for the increase in FEV_1 (Table 3.2):

Table 3.2 Asthma data

Treatment	n	$\bar{x}$	s
A	18	54.6	14.6
B	21	38.8	12.9

95 per cent and 99 per cent confidence intervals for μ_1 and μ_2, the population mean increases in FEV_1 in treatment groups A and B, are calculated as follows (Table 3.3):

Table 3.3 Confidence intervals for asthma data

Treatment	Multiplying constants (95%/99%)	$\frac{s}{\sqrt{n}}$	95% CI	99% CI
A	2.12/2.92	3.54	(47.1, 62.1)	(44.3, 64.9)
B	2.09/2.86	2.88	(41.3, 53.3)	(39.1, 55.5)

3.1.4 The role of the standard error

Note the role played by the standard error in the formula for the confidence interval. We have previously seen that the standard error of the mean provides an indirect measure of the precision with which we have calculated the mean. The confidence interval has now translated the numerical value for the standard error into something useful in terms of being able to make a statement about where μ lies. A large standard error will lead to a wide confidence interval reflecting the imprecision and resulting poor information about the value of μ. In contrast a

small standard error will produce a narrow confidence interval giving us a very definite statement about the value of μ.

For sample sizes beyond about 30 the multiplying constant for the 95 per cent confidence interval is approximately equal to two. Sometimes for reasonably large sample sizes we may not agonise over the value of the multiplying constant and simply use the value two as a good approximation. This gives us an approximate formula for the 95 per cent confidence interval as $(\overline{x} - 2se, \overline{x} + 2se)$.

Finally, returning again to the formula for the standard error, $s/\sqrt{n}$, we can, at least in principle, see how we could make the standard error smaller; increase the sample size n and reduce the patient-to-patient variability. These actions will translate into narrower confidence intervals.

3.2 Confidence intervals for other parameters

3.2.1 Difference between two means

At the end of the previous chapter we saw how to extend the idea of a standard error for a single mean to a standard error for the difference between two means. The extension of the confidence interval is similarly straightforward. Consider the placebo controlled trial in cholesterol lowering described in Example 2.3 in Chapter 2. We had an observed difference in the sample means $\overline{x}_1 - \overline{x}_2$ of 1.4 mmol/l and a standard error of 0.29. The formula for the 95 per cent confidence interval for the difference between two means $(\mu_1 - \mu_2)$ is:

$$(\overline{x}_1 - \overline{x}_2) - (\text{constant} \times se) \text{ to } (\overline{x}_1 - \overline{x}_2) + (\text{constant} \times se)$$

This expression is essentially the same as that for a single mean: statistic $\pm$ (constant $\times$ *se*). The rules for obtaining the multiplying constant however are slightly different. For the difference between two means we use Table 3.1 as before, but now we go into that table at the row $n_1 + n_2 - 2$, where n_1 and n_2 are the sample sizes for treatment groups 1 and 2 respectively.

So for our data ($n_1 = 24$ and $n_2 = 23$), the multiplying constant (from row 45) is 2.01 and the calculation of the 95 per cent confidence interval is as follows:

$$\text{lower confidence limit} = 1.4 - (2.01 \times 0.29) = 0.8 \,\text{mmol/l}$$
$$\text{upper confidence limit} = 1.4 + (2.01 \times 0.29) = 2.0 \,\text{mmol/l}$$

The interpretation of this interval is essentially as before; we can be 95 per cent confident that the true difference in the (population) means, $\mu_1 - \mu_2$, is between 0.8 and 2.0. In other words the data are telling us that the mean reduction μ_1 in

the active group is greater than the corresponding mean reduction μ_2 in the placebo group by between 0.8 mmol/l and 2.0 mmol/l.

3.2.2 Confidence intervals for proportions

The previous sections in this chapter are applicable when we are dealing with means. As noted earlier these parameters are relevant when we have continuous, count or score data. With binary data we will be looking to construct confidence intervals for rates or proportions plus differences between those rates.

Example 3.3: Trastuzumab in HER2-positive breast cancer

The following data (Table 3.4) are taken from Piccart-Gebhart *et al.* (2005) who compared trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer with observation only. The binary outcome here is one or more serious adverse events (SAEs) versus no SAEs during the one year trial. The rate in the observation only group provides the background incidence of SAEs.

Table 3.4 Trastuzumab data

	≥ 1 SAE	No SAEs	Total
Trastuzumab	117	1560	1677
Observation	81	1629	1710
Total	198	3189	3387

This display is termed a 2×2 *contingency table*.
The incidence rates in the test treatment and control groups respectively are:

$$r_1 = \frac{117}{1677} = 0.070 \quad r_2 = \frac{81}{1710} = 0.047$$

For example 3.3, if we label the true SAE incidence rates in the population as a whole as θ_1 (assuming all patients in the population received trastuzumab) and θ_2 (assuming all patients were only observed), then we would be interested in confidence intervals for the individual rates θ_1 and θ_2 and also the difference in those rates $\theta_1 - \theta_2$.

In Section 2.5.2 we set down the formulas for the standard errors for both individual rates and the difference between two rates. These lead naturally to expressions for the confidence intervals.

For the trastuzumab group the 95 per cent confidence interval for θ_1 is given by:

$$0.070 \pm 1.96\sqrt{\frac{0.070(1-0.070)}{1677}} = (0.058, 0.082)$$

The 95 per cent confidence interval for $\theta_1 - \theta_2$, the difference in the SAE rates, is given by:

$$\begin{aligned}
&(r_1 - r_2) \pm 1.96\sqrt{\frac{r_1(1-r_1)}{n_1} + \frac{r_2(1-r_2)}{n_2}} \\
&= (0.070 - 0.047) \pm 1.96\sqrt{\frac{0.070(1-0.070)}{1677} + \frac{0.047(1-0.047)}{1710}} \\
&= (0.007, 0.039)
\end{aligned}$$

So with 95 per cent confidence we can say that the absolute difference in SAE rates between trastuzumab and observation only is between 0.7 per cent and 3.9 per cent.

Note that for binary data and proportions the multiplying constant is 1.96, the value used previously when we first introduced the confidence interval idea. Again this provides an approximation, but in this case the approximation works well except in the case of very small sample sizes.

3.2.3 General case

In general, the calculation of the confidence interval for any statistic, be it a single mean, the difference between two means, a median, a proportion, the difference between two proportions and so on, always has the same structure:

$$\text{statistic} \pm (\text{constant} \times se)$$

where the *se* is the standard error for to the statistic under consideration.

There are invariably rules for how to obtain the multiplying constant for a specific confidence coefficient, but as a good approximation and providing the sample sizes are not too small, using the value 2 for the 95 per cent confidence interval and 2.6 for the 99 per cent confidence interval would get you very close.

This methodology applies whenever we are looking at statistics based on single treatment groups or those relating to differences between treatment groups. When we are dealing with ratios, such as the Odds Ratio or the Hazard Ratio the methodology is changed slightly. We will cover these issues in a later chapter (see Chapter 4.5.5 for example for confidence intervals for the Odds Ratio).

3.3 Hypothesis testing

In our clinical trials we generally have some very simple questions:

- Does the drug work?
- Is treatment A better than treatment B?
- Is there a dose response?
- Are treatments A and B clinically equivalent?

and so on.

In order to evaluate the truth or otherwise of these statements we begin by formulating the questions of interest in terms of hypotheses. The simplest (and most common) situation is the comparison of two treatments, for example in a placebo controlled trial, where we are trying to detect differences and demonstrate that the drug works.

Assume that we are dealing with a continuous endpoint, for example, fall in diastolic blood pressure, and we are comparing means. If μ_1 and μ_2 denote the mean reductions in groups 1 and 2 respectively then our basic question is as follows:

$$\text{is } \mu_1 = \mu_2 \quad \text{or is} \quad \mu_1 \neq \mu_2?$$

We formulate this question in terms of two competing hypotheses:

$$H_0 : \mu_1 = \mu_2 \quad \text{termed the } \textit{null hypothesis}$$

and

$$H_1 : \mu_1 \neq \mu_2 \quad \text{termed the } \textit{alternative hypothesis}$$

We base our conclusion regarding which of these two statements (hypotheses) we 'prefer' on data and the method that we use to make this choice is the *p*-value.

3.3.1 Interpreting the *p*-value

The '*p*' in *p*-value stands for probability and as such it therefore lies between 0 and 1. I am sure we all know that if the *p*-value falls below 0.05 we declare statistical significance and conclude that the treatments are different, that is $\mu_1 \neq \mu_2$. In contrast if the *p*-value is above 0.05 then we talk in terms of non-significant differences. We will now explore just how this *p*-value is defined and later we will see the principles behind its calculation.

In the context of the comparison of an active treatment (A) with a placebo treatment (B) in lowering diastolic blood pressure assume that we have the following data:

$$\bar{x}_1 = 9.6\,\text{mmHg (active)}$$
$$\bar{x}_2 = 4.2\,\text{mmHg (placebo)}$$

with a difference, $\bar{x}_1 - \bar{x}_2 = 5.4\,\text{mmHg}$

Suppose that the *p*-value turns out to be 0.042. What does this *p*-value actually measure? We can see of course that it is ≤ 0.05 and so would give statistical significance, but what does the probability 0.042 refer to. What is it the probability of?

Usually people give one of two responses to this question:

- Proposed definition 1: *There is a 4.2 per cent probability that* $\mu_1 = \mu_2$
- Proposed definition 2: *There is a 4.2 per cent probability that the observed difference of 5.4 mmHg is due to chance*

One of these definitions is correct and one is incorrect? Which way round is it?

Well, the second definition is the correct one. The first definition is not only incorrect, it is also the common mistake that many people make. We will explore later in Section 9.3.1 why this definition causes so many problems and misunderstandings. For the moment, however, we will explore in more detail the correct definition. It is worthwhile expanding on the various components of the definition:

- There is a 4.2 per cent probability that
- the observed difference *or a bigger difference in either direction (A better than B or B better than A)*
- is a chance finding *that has occurred with equal treatments* ($\mu_1 = \mu_2$ *or when the null hypothesis is true*)

Commit this definition to memory, it is important!

To complete the logic we then consider this statement and argue as follows:

There is only a 4.2 per cent chance of seeing a difference as big as the one observed with equal treatments. This is a small probability and it is telling us that these data are not at all likely to have occurred with equal treatments and it is on this basis that we do not believe that the treatments are equal. We declare statistically significant differences between the treatment means.

In contrast if the *p*-value had been, say, 0.65 then the definition says that there is a 65 per cent probability of seeing a difference as big (or bigger) than the one observed with equal treatments. Now 65 per cent is quite a high probability

and what we are seeing in this case is a difference that is entirely consistent with $\mu_1 = \mu_2$, it is the kind of difference you would expect to see with equal treatments and therefore we have no reason to doubt the equality of the means.

Another way of thinking about the *p*-value is as a measure of how consistent the difference is with equal treatments (or equivalently with the null hypothesis). A low *p*-value says that the difference is not consistent with equal treatments, a high value says that the difference is consistent with equal treatments. The conventional cut-off between 'low' and 'high' is 0.05.

Many people ask at this stage: why 0.05? Well it is in one sense an arbitrary choice, the cut-off could easily have been 0.04 or 0.075 but 0.05 has become the agreed value, the convention. We will explore the implications of this choice later when we look at type I and type II errors.

The way that the hypotheses are set up is that we always structure H_1 to be our 'objective'. H_1 represents the desirable outcome; we want to come out of the clinical trial concluding in favour of H_1. The *p*-value measures how consistent the data are with H_0, and if the *p*-value is small, the data are not consistent with H_0 and we declare statistical significance and decide in favour of H_1. In this way we are essentially trying to disprove H_0. This is the *scientific method* with its roots in philosophical reasoning; the way science advances is by disproving things rather than by proving them. For example, proving that 'all swans are white' is very difficult but you only have to see one black swan to disprove that statement.

3.3.2 Calculating the *p*-value

We will start with a very simple situation to see how we actually calculate *p*-values. Suppose we want to know whether a coin is a fair coin; by that we mean that when we flip the coin, it has an equal chance of coming down heads (H) or tails (T).

Let pr(H) denote the probability of the coin coming down heads. We can then formulate null and alternative hypotheses as follows:

$$H_0 : \text{pr(H)} = \frac{1}{2} \text{ (fair coin)} \quad H_1 : \text{pr(H)} \neq \frac{1}{2} \text{ (unfair coin)}$$

We now need some data on which to evaluate the hypotheses. Suppose we flip the coin 20 times and end up with 15 heads and 5 tails. Without thinking too much about probabilities and *p*-values what would your intuition lead you to conclude? Would you say that the data provide evidence that the coin is not fair or are the data consistent with the coin being fair?

We will now be a little more structured about this. Because this is such a simple situation we can write down everything that could have happened in this

experiment and fairly easily calculate the probabilities associated with each of those outcomes *under the assumption that the coin is fair*. These are all contained in Table 3.5.

Note that we have included a column H − T; this is the number of heads minus the number of tails. This is done in order to link with what we do when we are comparing treatments where we use 'differences' to measure treatment effects.

So with a fair coin, getting 12 heads and 8 tails, for example, will happen on 0.120 (12 per cent) of occasions. The most likely outcome with a fair coin, not surprisingly, is 10 heads and 10 tails and this will happen 17.6 per cent of the time. The extreme outcomes are not at all likely but even 20 heads and 0 tails can still occur and we will see this outcome 0.000095 per cent of the time!

Our data were 15 heads and 5 tails, so how do we calculate the *p*-value? Well, remember the earlier definition and translate that into the current setting: *the probability of getting the observed data or more extreme data in either direction with a fair coin*. To get the *p*-value we add up the probabilities (calculated when the null hypothesis is true – coin fair) associated with our data (15 heads, 5 tails) and more extreme data (a bigger difference between the number of heads and the number of tails) in either direction:

$$= (0.00000095 + 0.000019 + 0.00018 + 0.0011 + 0.0046 + 0.015) \times 2$$

$$= 0.0417998 \text{ or } 0.0418 = p$$

This means that there is only a 4.18 per cent probability of seeing the 15/5 split or a more extreme split (either way) with a fair coin. This probability is below the magical 5 per cent, we have a statistically significant result and the evidence suggests that the coin is not fair.

Had we seen the 14/6 split, however, the *p*-value would have increased to $0.0417998 + 2 \times 0.037 = 0.1157998$, a non-significant result; the 14/6 split is not sufficiently extreme for us to be able to reject the null hypothesis (according to the conventional cut-off at 5 per cent). The 15/5 split (H − T = 10 or − 10) therefore is the smallest split that just achieves statistical significance.

It is useful to look at this visually. Figure 3.2 plots each of the outcomes on the *x*-axis with the corresponding probabilities, calculated when the null hypothesis is true, on the *y*-axis. Note that the *x*-axis has been labelled according to 'heads – tails' (H − L), the number of heads minus the number of tails. This identifies each outcome uniquely and allows us to express each data value as a difference. More generally we will label this the *test statistic*; it is the statistic that the *p*-value calculation is based. The graph and the associated table of probabilities are labelled the *null distribution (of the test statistic)*.

Using the plot we calculate the *p*-value, firstly by identifying the outcome we saw in our data, and secondly by adding up all those probabilities associated with

Table 3.5 Outcomes and probabilities for 20 flips of a *fair* coin (See below for the method of calculation for the probabilities)

Heads (H)	Tails (T)	H − T	Probability (when coin fair)
20	0	20	0.00000095
19	1	18	0.000019
18	2	16	0.00018
17	3	14	0.0011
16	4	12	0.0046
15	5	10	0.015
14	6	8	0.037
13	7	6	0.074
12	8	4	0.120
11	9	2	0.160
10	10	0	0.176
9	11	−2	0.160
8	12	−4	0.120
7	13	−6	0.074
6	14	−8	0.037
5	15	−10	0.015
4	16	−12	0.0046
3	17	−14	0.0011
2	18	−16	0.00018
1	19	−18	0.000019
0	20	−20	0.00000095

Calculating the probabilities for a fair coin:

Suppose we flip the coin just three times. The possible combinations are written below. Because the coin is fair these are all equally likely. And so each has probability $(1/2)^3 = 0.125$ of occurring.

	Probability
H H H	0.125
H H T	0.125
H T H	0.125
T H H	0.125
T T H	0.125
T H T	0.125
H T T	0.125
T T T	0.125

In terms of numbers of heads (H) and numbers of tails (T) there are just four possibilities and we simply add up the probabilities corresponding to the individual combinations.

Heads (H)	Tails (T)	Probability (when coin fair)
3	0	$0.125 = 1 \times (1/2)^3$
2	1	$0.375 = 3 \times (1/2)^3$
1	2	$0.375 = 3 \times (1/2)^3$
0	3	$0.125 = 1 \times (1/2)^3$

For 20 flips we get the probabilities by multiplying $(1/2)^{20}$ by the number of combinations that give rise to that particular outcome, so for example with 12 heads and 8 tails this is $20! \div (12! \times 8!)$ where n! denotes $n \times (n-1) \times \ldots \times 2 \times 1$.

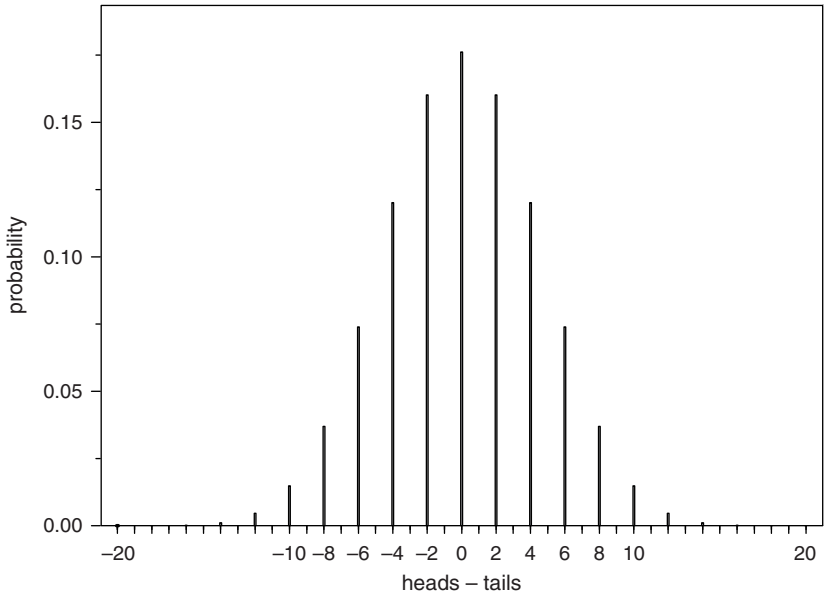


Figure 3.2 Null distribution for 20 flips of a fair coin

that outcome and the more extreme outcomes ('bigger' differences in terms of the test statistic) in both directions (positive and negative). Returning to the H − T difference of 10 or −10 (the 15/5 split); this outcome is at the boundary between $p \leq 0.05$ and $p > 0.05$ and just achieves statistical significance. We call this value for the test statistic, the *critical value*.

3.3.3 A common process

In the previous section we have calculated the *p*-value in a very simple situation. Nonetheless in more complex situations the steps that are required to calculate *p* are basically the same. Just briefly returning to the coin, all possible outcomes were expressed in terms of the difference, H − T. This quantity is what we referred to in Section 1.7.1 as the signal. Large differences are giving strong signals, strong evidence that the coin is not fair, small differences in contrast are giving weak signals.

More generally the test statistic is constructed as the signal/noise (signal-to-noise) ratio or something akin to this. We will develop this methodology in relation the comparison of two independent means for a between-patient design. The resulting test is known as the unpaired t-test or the two-sample t-test.

The null and alternative hypotheses are specified as follows:

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

The signal is captured by the observed difference $\bar{x}_1 - \bar{x}_2$. Earlier, in Section 1.7.2, we had said that the noise would be made up of two components, the sample size and the patient-to-patient variation. In fact it is the standard error of the difference between the two means that provides our measure of noise. For discussion purposes consider the formula for this standard error in the special case of equal group sizes ($n_1 = n_2 = n$); $se = \sqrt{s_1^2 + s_2^2/n}$, where s_1 and s_2 are the separate standard deviations within each of the treatment groups. Note that the formula contains both the sample size and also the measures, s_1 and s_2, of patient-to-patient variation. An increase in the sample size will reduce the noise as will a reduction in the patient-to-patient variation and vice versa.

Consider the earlier example comparing two treatments for the reduction of blood pressure and suppose in this case $n_1 = 20$ and $n_2 = 20$ are the sample sizes in the two treatment groups. The means were:

$$\bar{x}_1 = 9.6\,\text{mmHg (active)}$$
$$\bar{x}_2 = 4.2\,\text{mmHg (placebo)}$$

The observed difference $\bar{x}_1 - \bar{x}_2 = 5.4\,\text{mmHg}$ is the signal and suppose also by applying the formula the standard error calculation has given $se = 2.57$.

The value of the signal-to-noise ratio is signal/noise $= 5.4/2.57 = 2.10$. The value of this ratio reflects the evidence in favour of treatment differences, the larger the value, positive or negative, the stronger the evidence. Large values of this ratio will come from strong signals where the noise is controlled. Small values of this ratio will come from either weak signals or stronger signals, but where the noise is large, casting doubt on the 'reliability' of the strong signal. We have previously defined the *p*-value as the probability of seeing a difference between the means greater than 5.4 mmHg. Getting a signal greater than 5.4 mmHg is equivalent to getting a signal-to-noise ratio greater than 2.10, assuming that the noise is fixed, and it is this calculation that will give us the *p*-value.

It turns out that the signal-to-noise ratio, under the assumption that the two treatment means are the same, has a predictable behaviour (we will say more about this in the next chapter) and the probabilities associated with values of this ratio are given by a particular distribution, the t-distribution. Figure 3.3 displays these probabilities for the example we are considering. Note that we have labelled this the t-distribution on 38 degree of freedom again we will say more about where the 38 comes from in the next chapter.

Using computer programs we can add up all the probabilities (shown as 'prob' on the figure) associated with the observed value 2.10 of the test statistic and

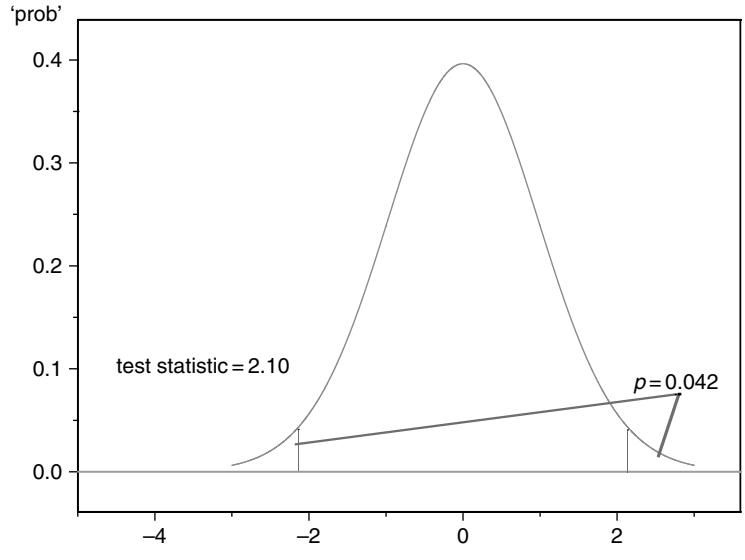


Figure 3.3 The t-distribution on 38 d.f.

more extreme values, in both directions (A better than B and B better than A) as seen in the identified areas under the t-distribution to give the *p*-value, 0.042.

This calculation of the *p*-value comes from the probabilities associated with these signal-to-noise ratios and this forms a common theme across many statistical test procedures. In general, the signal-to-noise ratio is again referred to as the test statistic. The distribution of the test statistic when the null hypothesis is true (equal treatments) is termed the *null distribution*.

In general we can think of the *p*-value calculation as a series of four steps as follows:

1. Formulate null and alternative hypotheses. In all cases the alternative hypothesis represents the 'desirable' outcome. In a superiority trial this means that the null hypothesis is equality (or no effect/no change/no dependence) of whatever is being compared while the alternative hypothesis is inequality (there is an effect/a change/a dependence).

2. Calculate the value of the test statistic (usually = signal/noise). The formula for the test statistic will be based on a standard approach determined by the data type, the design of the trial (between- or within-patient) and the hypotheses of interest. Mathematics has provided us with optimum procedures for all the common (and not so common) situations and we will see numerous examples in subsequent sections.

3. Determine the null distribution of the chosen test statistic, that is, what are the probabilities associated with all the potential values of the test statistic when

the null hypothesis is true? Again mathematical theory has provided us with solutions to this and all of these null distributions are known to us; we simply have to look them up.

4. Obtain the p-value by adding up the probabilities associated with the calculated value of the test statistic and more extreme values when the null hypothesis is true. This will correspond to adding up the probabilities associated with the observed signal or more extreme (larger) values of that signal.

Finally 'step 5' is to draw conclusions; if $p \leq 0.05$ then declare statistical significance; if $p > 0.05$ then the differences are not statistically significant.

3.3.4 The language of statistical significance

There is a fair amount of language that we wrap around this process. We talk in terms of a *test of significance*. If $p \leq 0.05$, we declare statistical significance, and *reject the null hypothesis at the 5 per cent level*. We call 5 per cent the *significance level*, it is the level at which we declare statistical significance. If $p > 0.05$, then we say we have a non-significant difference and we are *unable to reject the null hypothesis at the 5 per cent level*.

Our conventional cut-off for statistical significance is 5 per cent but we also use other levels, notably 1 per cent and 0.1 per cent. If $p \leq 0.01$ then the evidence is even stronger that there are differences and we have highly significant differences. If $p \leq 0.001$ then the evidence is even stronger still and we have very highly significant differences.

There is often quite a bit of discussion when we see $0.05 < p \leq 0.10$; 'almost significant', 'a trend towards significance', 'approaching significance' and other imaginative phrases! I have some sympathy with such comments. One thing we have to remember is that the p-value scale is a continuum and the strict cut-off at 0.05 is in a sense unrealistic. There really is little difference, from a strength of evidence point of view, between $p = 0.048$ and $p = 0.053$, yet one gives statistical significance and one does not. Unfortunately many practitioners (including regulators) seem to have a strict demarcation at 0.05. In one sense this is understandable; having a strict cut-off at 0.05 removes any ambiguity.

3.3.5 One-tailed and two-tailed tests

The p-value calculation detailed in the previous section gives what we call a *two-tailed* or a *two-sided test* since we calculate p by taking into account values of the test statistic equal to, or more extreme, than that observed, in both directions. So

for example with the coin we look for movement away from 'coin fair' both in terms of 'heads more likely than tails' and 'tails more likely than heads'.

In part this is because of the way we set up the hypotheses; in our earlier discussion we asked 'is the coin fair?' or 'is the coin not fair?' We could have asked a different set of questions; 'is the coin fair?' or 'are heads more likely than tails?' in which case we could have been justified in calculating the *p*-value only in the 'tail' corresponding to 'heads more likely than tails'. This would have given us a *one-tailed* or a *one-sided test*. Under these circumstances, had we seen 17 tails and 3 heads then this would not have led to a significant *p*-value, we would have discounted that outcome as a chance finding, it is not in the direction that we are looking for.

Clearly, one-sided *p*-values are of interest to sponsors, firstly they are smaller and more likely to give a positive result and secondly many sponsors would argue that they are only interested in departures from the null hypothesis in one particular direction; the one that favours their drug. While this may be an argument a sponsor might use, the regulators (and the scientific community more generally) unfortunately would not support it. Regulators are interested in differences both ways and insist that generally *p*-values are two-sided. In some circumstances they may be comfortable with one-sided *p*-values, but when this is the case they also state that the significance level used is 0.025 rather than 0.05. Now because most situations are symmetric, the two-sided p is usually equal to $2 \times$ the one-sided p, so it actually makes no difference mathematically! Also to avoid any confusion the decision to use one-sided tests should be made prospectively.

ICH E9: 'Note for Guidance on Statistical Principles for Clinical Trials'

'It is important to clarify whether one- or two-sided tests of statistical significance will be used, and in particular to justify prospectively the use of one-sided tests . . . The approach of setting type I errors for one-sided tests at half the conventional type I error used in two-sided tests is preferable in regulatory settings.'

4

Tests for simple treatment comparisons

4.1 The unpaired t-test

In Section 3.3.3 we introduced the general structure for a significance test with the comparison of two means in a parallel group trial. This resulted in a procedure which goes under the general heading of the two-sample (or unpaired) t-test. This test was developed for continuous data, although it is applicable more widely and, in particular, is frequently used for score and count data.

The test was developed almost 100 years ago by William Sealy Gosset. Gosset was in fact a chemist by training and was employed by the Guinness brewery, initially in Dublin, Ireland, but subsequently at the Guinness brewery in London. He became interested in statistics and in particular in the application of statistics to the improvement of quality within the brewing process. Gosset's work was based on a combination of mathematics and empirical experience (trial and error), but the procedures he came up with have certainly stood the test of time; the unpaired t-test is undoubtedly the most commonly used (although not always correctly) statistical test of them all.

The calculation of the *p*-value in the example in Section 3.3.3 consisted of adding up the probabilities, associated with values of the signal-to-noise ratio greater than the observed value of 2.10, given by the t-distribution. It turns out that these probabilities depend upon the number of patients included in the trial. There are an infinite number of t-distributions; t_1, t_2, t_3, . . . and the one we choose is based on calculating the total sample size (both groups combined) and subtracting two. We will see that this t-distribution is used in other settings where the rule for choosing the particular distribution is different, but the rule for the unpaired t-test is $n_1 + n_2 - 2$.

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

There is a connection with what we are seeing here and the calculation of the confidence interval in Chapter 3. Recall Table 3.1 within Section 3.1.3, 'Changing the multiplying constant'. It turns out that *p*-values and confidence intervals are linked and we will explore this further in a later chapter. The confidence coefficients for d.f. = 38 are 2.02 for 95 per cent confidence and 2.71 for 99 per cent confidence. If we were to look at the t_{38} distribution we would see that ± 2.02 cuts off the outer 5 per cent probability while ± 2.71 cuts off the outer 1 per cent of probability.

Having calculated the *p*-value we would also calculate the 95 per cent confidence interval for the difference $\mu_1 - \mu_2$ to give us information about the magnitude of the treatment effect. For the data in the example in Section 3.3.3 this confidence interval is given by:

$$(5.4 \pm 2.02 \times 2.57) = (0.2, 10.6)$$

So with 95 per cent confidence we can say that the true treatment effect $(\mu_1 - \mu_2)$ is somewhere in the region 0.2 mmHg to 10.6 mmHg.

4.2 The paired t-test

The *paired t-test* also known as the *one-sample t-test* was also developed by Gosset. This test is primarily used for the analysis of data arising from within-patient designs, although we also see it applied when comparing a baseline value with a final value within the same treatment group.

Consider a two-period, two-treatment cross-over trial in asthma comparing an active treatment (A) and a placebo treatment (B) in which the following PEF (l/min) data, in terms of the value at the end of each period, were obtained (Table 4.1).

Patients 1 to 16 received treatment A followed by treatment B while patients 17 to 32 received treatment B first.

Table 4.1 Data from a cross-over trial in asthma (hypothetical)

Patient	A	B	Difference (A − B)
1	395	362	33
2	404	385	19
3	382	386	−4
.			
.			
.			
32	398	344	54

The final column above has calculated the A − B differences and as we shall see, the paired t-test works entirely on the column of differences. Again we will follow through several steps for the calculation of the *p*-value for the A versus B comparison:

1. Let μ be the population mean value for the column of differences. The null and alternative hypotheses are expressed in terms of this quantity:

$$H_0: \mu = 0 \quad H_1: \mu \neq 0$$

   A non-zero value for μ will reflect treatment differences; a positive value in particular is telling us that the active treatment is effective.

2. Again the test will be based on the signal-to-noise ratio. In this case the signal is the observed mean $\bar{x}$ of the column of differences while the noise is the standard error associated with that mean. For these data:

$$\bar{x} = 28.4\text{l/min} \quad se(\text{of } \bar{x}) = 11.7\text{l/min}$$

   The standard error here is obtained from the standard deviation of the differences divided by the square root of 32, the number of patients.

   The test statistic = signal/noise = 28.4/11.7 = 2.43 captures the evidence for treatment differences. Larger values, either positive or negative, are an indication of treatment differences.

3. The probabilities associated with the values that this signal-to-noise ratio can take when the treatments are the same are again given by the *t* shape; in this case the appropriate t-distribution is t_{31}, the t-distribution on 31 degrees of freedom. Why t_{31}? The appropriate t-distribution is indexed by the number of patients minus one.

4. Our computer programs now calculate the *p*-value; the probability associated with getting a value for the signal-to-noise ratio at least as large as 2.43, in either direction, when the null hypothesis is true. This value turns out to be 0.021 (see Figure 4.1). This value also reflects, under the assumption of constant noise, the probability of seeing a mean difference at least as big as 28.4 l/min by chance with equal treatments. This signal is sufficiently strong for us to conclude a real treatment effect.

   The *p*-value is less than 0.05 giving statistical significance at the 5 per cent level and we conclude, on the basis of the evidence, that treatment A (active) is better than treatment B (placebo); the active treatment works.

This test is based entirely on the column of differences; once the column of differences is calculated, the original data are no longer used. An alternative approach might have been to simply calculate the mean value on A and the

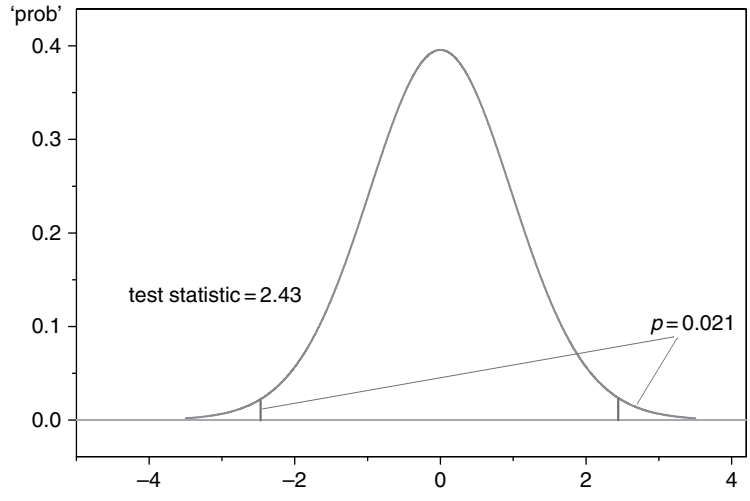


Figure 4.1 The t-distribution on 31 d.f.

mean value on B and to compare the two using the unpaired t-test. Would this have worked? In fact, no, using the unpaired t-test in this way would have been incorrect; the unpaired t-test is used to compare means across two independent samples. The paired t-test uses the data in the most efficient way; forming the column of differences links the observations on the same patient and is effectively using each patient as their own control. Calculating the A and B means separately and taking their difference would in fact have given the same signal as the mean of the column of differences, but the noise would be different. It would not reflect the patient-to-patient variability in the A − B differences, but would be based on the patient-to-patient (within-group) variability in the value of the endpoint itself.

Does it matter which way round the differences are calculated; A − B or B − A? No, as long as we are consistent across all of the patients it does not matter and will simply have the effect of changing the sign of the test statistic, but the *p*-value will remain unchanged.

As with the unpaired t-test it would be useful to calculate a confidence interval for the treatment effect, μ. This is given by:

$$(\bar{x} \pm 2.04 \times se) = (28.4 \pm 2.04 \times 11.7)$$
$$= (4.5, 52.3)$$

Here, 2.04 is the appropriate multiplication constant for 95 per cent confidence. So, we can be 95 per cent confident that the treatment difference (effect), μ, is between 4.5 l/min and 52.3 l/min.

Before we move on it should be pointed out that the paired t-test provides a valid analysis for continuous data arising from a cross-over trial only when the trial is balanced, that is when the number of patients following the A/B sequence is the same as the number of patients following the B/A sequence. When this is not true the analysis needs to be modified slightly. This is because in many settings there will also be a *period effect*, that is, irrespective of treatment, patients may respond differently in period I compared to period II. This could be caused, for example, by the temporal nature of the underlying disease, or by external conditions which may be changing over time. If a period effect is present, but the trial is balanced then there will be equal numbers of A and B patients giving responses in period I and also in period II and under fairly general conditions the period effect will cancel out in the paired t-test comparison of the treatments. When balance is not present, then this effect will not cancel out and there will be bias. The reader is referred to Senn (2002), Section 3.6 for details of how to deal with this.

4.3 Interpreting the t-tests

The following example illustrates several issues and problems associated with the interpretation of p-values arising out of the t-tests. The setting is a very common one where a variable is measured at baseline and then subsequently at the end of the treatment period and the analysis focuses on the relative effects of the two treatments.

Example 4.1: Comparison of two active treatments for the treatment of Major Depressive Disorder in a randomised control trial

The primary endpoint in this trial was the 17 point Hamilton Depression Scale (HAMD-17) and the data presented in Table 4.2 correspond to mean (*se*).

Table 4.2 Data on HAMD-17 (hypothetical)

	Baseline	Final (week 8)	Change from baseline
Active A ($n = 36$)	27.4 (1.18)	15.3 (0.92)	12.1 (0.95)
Active B ($n = 35$)	26.8 (1.22)	11.8 (1.32)	15.0 (1.17)

Example 4.1: (Continued)

There are a number of comparisons that we can undertake:

1. Unpaired t-test comparing treatment means at baseline, $p = 0.73$
2. Unpaired t-test comparing treatment means at week 8, $p = 0.030$
3. Unpaired t-test comparing the mean change from baseline in the active A group with that in the active B group, $p = 0.055$
4. Paired t-test of baseline with week 8 in the active A group, $p << 0.001$ (this means the p-value is very much less than 0.001)
5. Paired t-test of baseline with week 8 in the active B group, $p << 0.001$.

Let us consider each of these tests in turn and their interpretation:

1. Test 1 is telling us that the treatment means are comparable at baseline, which is what we would expect to see given that this is a randomised trial. Of course chance differences can sometimes occur. Indeed in a randomised trial we would expect to see $p \leq 0.05$ for such a baseline comparison 5 per cent of the time. See Section 6.9 for a further discussion on this point.
2. This test compares the treatment groups at week 8. The p-value suggests a treatment difference, but does this test necessarily provide an analysis of the data which uses all of the information? Is there a possibility that we are being misled? Note that even though the earlier comparison of baseline means gave a non-significant p-value, the mean in active group A at baseline is slightly higher than the mean in the B group. Could this have contributed to the observed difference at week 8?
3. Test 3 for the comparison of the mean change from baseline between the groups is marginally non-significant. It would appear that the difference seen in test 2 is, in part, caused by the differences already seen at baseline. Looking at change from baseline has accounted for the minor baseline imbalances and, in general, is the basis for a more appropriate and sensitive analysis than simply looking at the week 8 means. It would be inappropriate to place too much emphasis on the fact that in test 3 the p-value is technically non-significant; recall that 0.05 is a somewhat arbitrary cut-off with regard to what we define as 'statistical significance'. It is test 3 that provides the most appropriate information for evaluating the relative effects of the two treatments.

We will say quite a bit later, in the chapter on adjusted analysis and analysis of covariance (Chapter 6) about additional improvements to this kind of analysis that increase sensitivity further and also avoid the so-called potential problem of regression towards the mean. For the moment though, it is test 3 that is the best way to compare the treatments.

4. Test 4 has given a very impressive *p*-value, but what is the correct interpretation of this test? The fall in HAM-D score from 27.4 to 15.3 surely indicates that active A is an effective treatment! Well, actually it does not. The fall seen in this group could indeed have been caused by the medication, but equally it could have been caused, for example, by the ancillary counselling that all patients will be receiving or as a result of the placebo effect (the psychological impact of being in the trial and receiving 'treatment') and we have no way of knowing which of these factors is having an effect and in what combination. The only way of identifying whether active drug A is efficacious is to have a parallel placebo group and undertake test 3; this would isolate the effect due to the specific medication from the other factors that could be causing the fall.
5. Test 5 should be interpreted in exactly the same way as test 4. The fall is impressive, but is it due to the active medication? We don't know and in the absence of a placebo group we will never know.

Suppose that test 4 had given $p = 0.07$ and test 5 had given $p = 0.02$. Would that therefore mean that active B is a better treatment than active A? No, in order to evaluate the relative effect of two treatments we have to compare them! Directly that is, not indirectly through the test 4 and 5 comparisons back to baseline.

4.4 The chi-square test for binary data

4.4.1 Pearson chi-square

The previous sections have dealt with the t-tests, methods applicable to continuous data. We will now consider tests for binary data, where the outcome at the subject level is a simple dichotomy; success/failure. In a between-patient, parallel group trial our goal here is to compare two proportions or rates.

In Section 3.2.2 we presented data from a clinical trial comparing trastuzumab to observation only after adjuvant chemotherapy in HER2-positive breast cancer. The incidence rates in the test treatment and control groups were respectively 7.0 per cent and 4.7 per cent.

The rate for patients suffering SAEs in the trastuzumab group (7.0 per cent) is clearly greater than the corresponding proportion in the observation only group (4.7 per cent), but is this difference (signal) strong enough for us to conclude that there are real differences?

The chi-square test for comparing two proportions or rates was developed by Karl Pearson around 1900 and pre-dates the development of the t-tests. The steps involved in the *Pearson chi-square test* can be set down as follows:

1. The null and alternative hypotheses relate to the two true rates, θ_1 and θ_2:

$$H_0: \theta_1 = \theta_2 \quad H_1: \theta_1 \neq \theta_2$$

2. In forming a test statistic, Pearson argued in the following way. In the data as a whole we see a total of 198 patients suffering SAEs. Had there been no differences between the groups in terms of the rate of SAEs then we would have seen equal proportions of patients suffering SAEs in the two groups. This would have meant seeing $198 \times (1677/3387) = 98$ patients with SAEs in the trastuzumab group and $198 \times (1710/3387) = 100$ in the observation only group. Similarly we should have seen 1579 patients not suffering SAEs in the trastuzumab group and 1610 such patients in the observation only group. We term these values, the *expected frequencies*, and denote them by *E*; the *observed frequencies* are denoted by *O*. These observed and expected frequencies are set down in Table 4.3 where the entries in the 2×2 contingency table are $O(E)$.

   We now need a measure of how far away we are from 'equal treatments'. Clearly, if the *E*s (what we should have seen with equal treatments) are close to the *O*s (what we have actually seen) then we have little evidence of a real difference in the incidence rates between the groups. However, the further apart the *O*s are from the *E*s, the more we believe the true SAE rates are different. The test statistic is formed by looking at each of the four 'cells' of the table and firstly calculating $(O - E)$. Some of these values will be positive and some negative, for example +19 in the trastuzumab ≥ 1 SAE cell and −19 in the observation only ≥ 1 SAE cell. We then square these values (this gets rid of the sign), divide by the corresponding *E* (we will say more of this later) and add up the resulting quantities across the four cells as shown below

Table 4.3 Observed and expected O(E) frequencies for trastuzumab data

	≥ 1 SAE	No SAEs	Total
Trastuzumab	117(98)	1560(1579)	1677
Observation	81(100)	1629(1610)	1710
Total	198	3189	3387

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

$$= \frac{(117-98)^2}{98} + \frac{(1560-1579)^2}{1579} + \frac{(81-100)^2}{100} + \frac{(1629-1610)^2}{1610}$$
$$= 7.75$$

This value captures the evidence in support of treatment differences. If what we have seen (*O*s) are close to what we should have seen with 'equal' treatments (*E*s) then this statistic will have a value close to zero. If, however, the *O*s and the *E*s are well separated then this statistic will have a larger positive value; the more the *O*s and the *E*s disagree, the larger it will be. For the moment this test statistic is not in the form of a signal-to-noise ratio, but we will see later that it can be formulated in that way.

3. Pearson calculated the probabilities associated with values of this test statistic when the treatments are the same, to produce the null distribution. This distribution is called the chi-square distribution on one degree of freedom, denoted χ^2_1, and is displayed in Figure 4.2. Note that values close to zero have the highest probability. Values close to zero for the test statistic would only result when the *O*s and the *E*s agree closely, whereas large values are unlikely when the treatments are the same.

4. To obtain the *p*-value we now need to add up all of the probabilities associated with values of the test statistic at least as big as the value observed, 7.75 in our

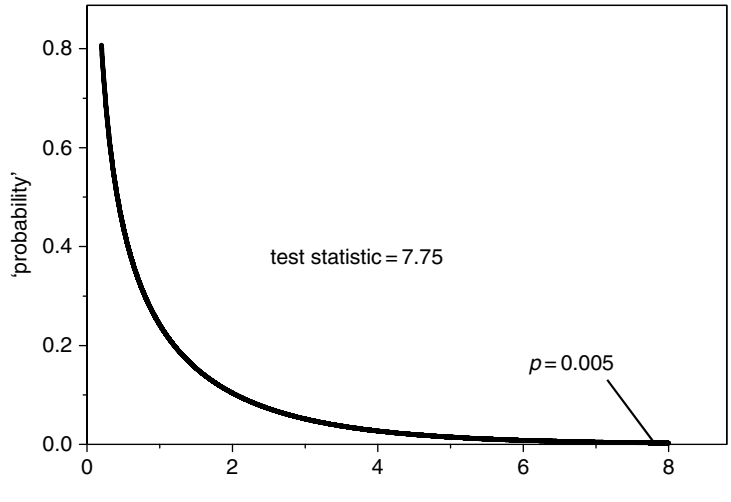


Figure 4.2 Chi-square distribution on one d.f.

data. As can be seen from Figure 4.2, this value is well out to the right of the distribution and gives, in fact, $p = 0.005$.

5. The p-value is less than 0.01 and so we have a highly significant result, a highly significant difference between the treatment groups in terms of SAE rates. Trastuzumab is associated with a significant increase in the SAE rate compared to observation only.

Several aspects of this p-value calculation deserve mention:

- The calculation of the test statistic involves division by E. This essentially weights the evidence from the different cells, so that a cell with a smaller expected frequency gets more weight, a cell with a larger expected frequency is down weighted. This makes sense since an $O - E$ difference of 19 in around 100 is much more relevant than a difference of 19 in 1600 and the test statistic is more influenced by the former than the latter.
- The $(O - E)^2$ values are all equal to 361 so algebraically the test statistic could be written:

$$(O-E)^2\left[\frac{1}{E_1}+\frac{1}{E_2}+\frac{1}{E_3}+\frac{1}{E_4}\right] = 361 \times \left[\frac{1}{98}+\frac{1}{1579}+\frac{1}{100}+\frac{1}{1610}\right]$$

- The null distribution for the t-test depended on the number of subjects in the trial. For the chi-square test comparing two proportions, and providing the sample size is reasonably large, this is not the case; the null distribution is always χ^2_1. As a consequence we become very familiar with χ^2_1. The *critical value* for 5 per cent significance is 3.841 while 6.635 cuts off the outer 1 per cent probability and 10.83 cuts off the outer 0.1 per cent.

4.4.2 The link to a signal-to-noise ratio

The formulation of the chi-square test procedure in the previous section, using observed and expected frequencies, is the standard way in which this particular test is developed and presented in most textbooks. It can be shown, however, that this procedure is akin to a development following the earlier signal-to-noise ratio approach.

In comparing two rates, the signal is provided by the observed difference $r_1 - r_2$ in the rates. The standard error for that difference as usual provides our measure of noise and this is given by the expression (see Section 2.5.2) $\sqrt{\frac{r_1(1-r_1)}{n_1}+\frac{r_2(1-r_2)}{n_2}}$. The probabilities associated with the resulting signal-to-noise ratio (the test statistic) when the true rates θ_1 and θ_2 are equal are provided by a special case of the normal distribution, N(0, 1); the normal distribution with mean zero and standard

deviation 1. For the trastuzumab example the signal takes the value 0.023 and the value of the noise is 0.0081. So to obtain the p-value based on this null distribution we compare the value of our test statistic ($= 0.023/0.0081 = 2.84$) with the N(0, 1) distribution giving a p-value of 0.0046. The Pearson chi-square test had earlier given $p = 0.0054$, a very similar value.

In general it can be shown that the approach following the signal-to-noise ratio method is mathematically very similar to the standard formulation of the chi-square test using observed and expected frequencies, and in practice they will invariably give very similar results. Altman (1991) (Section 10.7.4) provides more detail on this connection.

4.5 Measures of treatment benefit

The chi-square test has given a p-value and this provides the evidence, in general, in relation to the existence of a treatment difference. Through the confidence interval for the difference in the SAE rates calculated earlier in Section 3.2.2 (Example 3.3) we have some idea of the extent of the treatment effect in absolute terms. There are, however, other measures of treatment benefit/harm in common usage for binary data. Each of these measures; odds ratio, relative risk, relative risk reduction and number needed to treat are ways of expressing the treatment benefit. They each have their advantages and their disadvantages and I think it is fair to say that none of them is universally accepted as the 'single best approach'. We will define each of them in turn and provide an interpretation and critique of their use. For further details and discussion see Grieve (2003).

4.5.1 Odds ratio (OR)

In order to understand the odds ratio you first of all need to understand odds. For the data of Example 3.3 consider each of the treatment groups separately.

For the trastuzumab group, the odds of a patient suffering SAEs are $117/1560 = 0.075$; for every patient free from SAEs there are 0.075 patients suffering SAEs. For the observation only group the odds of a patient suffering SAEs are $81/1629 = 0.050$. In this group for every patient not suffering SAEs there are 0.050 patients who do suffer SAEs.

The *odds ratio (OR)* is then the ratio of the odds of a patient suffering one or more SAEs:

$$\text{OR} = 0.075/0.050 = 1.51$$

An odds ratio of one, or close to one, is telling us that the treatments are the same (or at least similar). An odds ratio greater than one tells us that you are worse off in the test treatment group and vice versa, but over and above that, the interpretation is not straightforward. It is the definition itself that provides this; the value 1.51 for the odds ratio indicates that the odds of suffering at least one SAE in the trastuzumab group is 1.51 times the odds of suffering at least one SAE in the observation only group. What makes this difficult to interpret is that it is a statement on the 'odds' scale and this is something that we find difficult to work with.

Usually, the odds relating to the test treatment group go on the top when calculating the ratio (the numerator), while the odds for the control group go on the bottom (the denominator). However there is no real convention regarding whether it is the odds in favour of success or the odds in favour of failure that we calculate. Had we chosen to calculate the odds in favour of no SAE, the odds ratio would have been $\frac{1/0.075}{1/0.050}$ which has the value $0.66(=1/1.51)$ so take care that when you see an odds ratio presented you are clear how the calculation has been organised.

4.5.2 Relative risk (RR)

The relative risk is defined again as a ratio, this time in relation to the risks calculated for the two treatments. For the trastuzumab group the 'risk' is the proportion of patients suffering SAEs which takes the value $117/1677 = 0.070$ while for the observation only group this is $81/1710 = 0.047$. The *relative risk (RR)* (sometimes called the *risk ratio*) is then the ratio of these risks:

$$\text{RR} = 0.070/0.047 = 1.47$$

A relative risk of one, or close to one, is again indicative of similar treatments. A relative risk above one, as here, is saying that the risk in the test treatment group is higher than that in the control group. The interpretation beyond that is a little simpler than the odds ratio. The relative risk of 1.47 is telling us that the risk in the trastuzumab group is 47 per cent higher than the risk in the observation only group.

There are also conventions with relative risk. As with the odds ratio we usually put the risk for the test treatment group as the numerator and the risk for the control group as the denominator. But now, because we are calculating risk there should be no confusion with regard to what we view as the event; we tend to calculate relative risk and not relative benefit.

Table 4.4 Active/placebo comparison, binary outcome survival (hypothetical)

	Died	Survived	Total
Active	20	80	100
Placebo	35	65	100
Total	55	145	200

4.5.3 Relative risk reduction (RRR)

Consider the data presented in table 4.4 relating to the binary outcome died/survival in a parallel group trial.

The relative risk is $0.20/0.35 = 0.57$

When the relative risk is less than one, as in this case, we often also calculate the reduction in the relative risk:

$$\textit{relative risk reduction } (RRR) = 1 - \text{relative risk}$$

In this example RRR = 0.43, there is a 43 per cent reduction in the risk (of death) in the active group compared to control.

We tend to use RRR where the intervention is having a benefit in reducing the risk. In the earlier example involving trastuzumab and the incidence of patients suffering SAEs, the active treatment was associated with an increase in risk. We could speak in terms of a relative risk increase (RRI) of $0.47(= 1.47 - 1)$, a 47 per cent increase in the risk of suffering SAEs, but this tends not to be done.

4.5.4 Number needed to treat (NNT)

In the example of the previous section (Table 4.4), 80 per cent of patients in the active group survived compared to 65 per cent in the placebo group. So out of 100 patients we would expect to see, on average, an additional 15 per cent (80 per cent – 65 per cent) surviving in the active group. The *number needed to treat (NNT)* is then 100/15 or 6.7. We need to treat on average an additional 6.7 patients with the active treatment in order to save one additional life.

A convenient formula for NNT is:

$$\text{NNT} = \frac{1}{(0.80 - 0.65)}$$

The denominator here is the difference in the survival proportions.

We usually round this up to the nearest integer, so NNT = 7. We need to treat seven patients with the active medication in order to see one extra patient survive compared to placebo.

There may be some situations where the test treatment is, in fact, harmful relative to the control treatment in terms of a particular endpoint. In these circumstances it does not make sense to take talk about number needed to treat and we refer instead to *number needed to harm (NNH)*. So, for example, if the survival rate on the test treatment were 72 per cent compared to 84 per cent in the control group then the number needed to harm would be equal to $1/(0.84 - 0.72)$ which rounds to eight.

4.5.5 Confidence intervals

We saw in the previous section methods for calculating confidence intervals for the difference in the SAE rates, or the event rates themselves. We will now look at methods for calculating a confidence interval for the odds ratio.

Calculating confidence intervals for ratios is a little more tricky than calculating confidence intervals for differences. We saw in Chapter 3 that, in general, the formula for the confidence interval is:

$$\text{statistic} \pm (\text{constant} \times se)$$

With a ratio it is not possible to obtain a standard error formula directly; however it is possible to obtain standard errors for log ratios. (Taking logs converts a ratio into a difference with $\log \text{A/B} = \log \text{A} - \log \text{B}$.) So we first of all calculate confidence intervals on the log scale. It does, in fact, not make any difference what base we use for the logs but by convention we usually use natural logarithms, denoted 'ℓn'.

The standard error for the ℓn of the OR is given by:

$$\sqrt{\frac{1}{O_1} + \frac{1}{O_2} + \frac{1}{O_3} + \frac{1}{O_4}}$$

where the Os are the respective observed frequencies in the 2×2 contingency table. In the trastuzumab example this is given by:

$$\sqrt{\frac{1}{98} + \frac{1}{1579} + \frac{1}{100} + \frac{1}{1610}} = 0.146$$

The 95 per cent confidence interval for the ℓn of the OR is then:

$$\ell\text{n } 1.51 \pm (1.96 \times 0.146) = (0.125, 0.699)$$

Finally we convert this back onto the OR scale by taking anti-logs of the ends of this interval to give a 95 per cent confidence interval for the OR as (1.13, 2.01). We can be 95 per cent confident that the OR lies within this range.

Previously when we had calculated a confidence interval, for example for a difference in rates or for a difference in means, then the confidence interval was symmetric around the estimated difference; in other words the estimated difference sat squarely in the middle of the interval and the endpoints were obtained by adding and subtracting the same amount (2 × standard error). When we calculate a confidence interval for the odds ratio, that interval is symmetric only on the log scale. Once we convert back to the odds ratio scale by taking anti-logs that symmetry is lost. This is not a problem, but it is something that you will notice. Also, it is a property of all standard confidence intervals calculated for ratios.

In similar ways to the above we can obtain confidence intervals for a relative risk and for a relative risk reduction. Confidence intervals for NNT are a little more complicated; see Grieve (2003) and Altman (1998) for further details.

4.5.6 Interpretation

In large trials and with events that are rare the OR and RR give very similar values. In fact we can see this in the trastuzumab example where the OR was 1.51 and the RR was 1.47. In smaller trials and with more common events, however, this will not be the case. Comparable values for the OR and the RR arise more frequently in cohort studies where generally the sample sizes are large and the events being investigated are often rare, and these measures tend to be used interchangeably. As a result there seems to be some confusion as to the distinction and it is my experience that the OR and RR are occasionally labelled incorrectly in clinical research papers, so take care.

As mentioned previously, all of the measures; difference in event rates, OR, RR, RRR and NNT, expressed in isolation, have limitations. What we are trying to do with such quantities is to use a single measure to summarise the data. All of the information is actually contained in the two event proportions/rates r_1 and r_2 and attempting to summarise two numbers by a single number is inevitably going to lead to problems in particular cases. Beware of those limitations and revert back to r_1 and r_2, if need be, to tell the full story.

4.6 Fisher's exact test

Pearson's chi-square test is what we refer to as a large sample test; this means that provided the sample sizes are fairly large then it works well. Unfortunately when the sample sizes in the treatment groups are not large there can be problems. Under these circumstances we have an alternative test, *Fisher's exact test*.

The way this works is as follows. Consider the 2 × 2 table below.

Table 4.5 Data for fisher's exact test

	Success	Failure	Total
Group A	6	18	24
Group B	1	23	24
Total	7	41	48

Table 4.6 Probabilities for fisher's exact test

Successes on A	Successes on B	Probability
7	0	0.0047
6	1	0.0439
5	2	0.1593
4	3	0.2921
3	4	0.2921
2	5	0.1593
1	6	0.0439
0	7	0.0047

Given there are only seven successes in total we can easily write down everything that could have happened (recall the way we looked at the flipping of the coin) and calculate the probabilities associated with each of these outcomes when there really are no differences between the treatments (Table 4.6).

We observed the 6/1 split in terms of successes in our data and we can calculate the p-value by adding up the probabilities associated with those outcomes which are as extreme, or more extreme, than what we have observed, when the null hypothesis is true (equal treatments). This gives $p = 0.097$. The corresponding chi-square test applied (inappropriately) to these data would have given $p = 0.041$, so the conclusions are potentially impacted by this.

The rule of thumb for the use of Fisher's exact test is based on the expected frequencies in the 2×2 contingency table; each of these need to be at least five for the chi-square test to be used. In the example, the expected frequencies in each of the two cells corresponding to 'success' are 3.5, signalling that Fisher's exact test should be used.

In fact, Fisher's exact test could be used under all circumstances for the calculation of the p-value, even when the sample sizes are not small. Historically, however, we tend not to do this; Fisher's test requires some fairly hefty combinatorial calculations in large samples to get the null probabilities and in the past this was just too difficult. For larger samples p-values calculated using either the chi-square test or Fisher's exact test will be very similar so we tend to reserve use of Fisher's exact test for only those cases where there is a problem and use the chi-square test outside of that.

4.7 The chi-square tests for categorical and ordinal data

4.7.1 Categorical data

The Pearson chi-square test extends in a straightforward way when there are more than two outcome categories.

Consider four outcome categories labelled A, B, C and D and the comparison of two treatments in terms of the distribution across these categories. Taking the example of categorical data from Chapter 1, we might have:

A = death from cancer causes

B = death from cardiovascular causes

C = death from other causes

D = survival

Consider the following hypothetical data (Table 4.7).

Table 4.7 Observed frequencies (0)

	A	B	C	D	Total
Group 1	15	13	20	52	100
Group 2	17	20	23	40	100
Total	32	33	43	92	200

The chi-square test proceeds, as before, by calculating expected frequencies. These are given in Table 4.8

Table 4.8 Expected frequencies (E)

	A	*B*	*C*	D	Total
Group 1	16	16.5	21.5	46	100
Group 2	16	16.5	21.5	46	100
Total	32	33	43	92	200

As before we compute:

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

with the sum being over all eight cells.

The resultant test statistic is then compared to the chi-square distribution, but this time on three degrees of freedom, written χ^2_3. As we mentioned earlier,

the particular chi-square shape that we use is not determined by the number of patients, rather it depends upon the size of the contingency table. In this example we have a 2×4 table (four outcome categories) and with two treatment groups the degrees of freedom for the chi-square distribution is equal to the number of categories minus one.

In the example the test statistic value is 3.38 and Figure 4.3 illustrates the calculation of the *p*-value, which for these data turns out to be $p = 0.336$. This is a non-significant result.

Although this test provides a valid comparison of the treatment groups in relation to the outcome categories, it is not of any great value in providing a useful conclusion from a clinical perspective. The procedure provides a test of the null hypothesis:

$$H_0: \theta_{1A} = \theta_{2A} \text{ and } \theta_{1B} = \theta_{2B} \text{ and } \theta_{1C} = \theta_{2C} \text{ and } \theta_{1D} = \theta_{2D}$$

against what we call the *general alternative hypothesis*:

Where the suffices 1 and 2 relate to treatment group,

$$H_1: \text{ the opposite of } H_0$$

But suppose we see a significant *p*-value, what does it mean? Well it simply means that there are some differences somewhere across the categories, but nothing more specific than that, and that is not particularly useful.

A further issue is that we very rarely see strictly categorical outcomes in practice in our clinical trials; it is much more common to have an ordering of the categories,

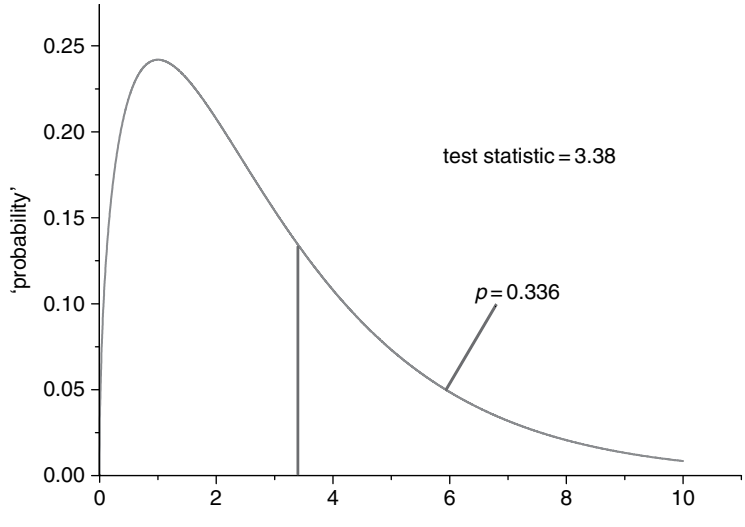


Figure 4.3 Chi-square distribution on 3 d.f.

giving us ordered categorical or ordinal data and we shall deal with this in the next section. Finally, the problem we discussed in the previous section regarding small sample sizes applies here and Fisher's exact test should be used when this is the case. The rule of thumb again is that all the expected frequencies should be at least five for the chi-square test to be valid. Computer programs tend to give both *p*-values automatically and it is no great problem therefore to pick the appropriate value depending on sample size and the rule of thumb.

4.7.2 Ordered categorical (ordinal) data

An endpoint which consists of a series of ordered categories is frequently used in our clinical trials in many different therapeutic settings.

The Pearson chi-square test is not appropriate for ordinal data as it does not take any account of the ordering of the outcome categories. The appropriate test is the *Mantel–Haenszel (MH) chi-square test* (Mantel and Haenszel (1959)). This test takes account of the ordering by scoring the ordered categories (for example improved = 1, no change = 2, worse = 3) and comparing the average score in one treatment group with the average score in the second treatment group. It is not quite as simple as this but that is the general idea! In fact the scores are chosen by the observed data patterns in the two groups combined.

The formula for the test statistic is somewhat complex, but again this statistic provides the combined evidence in favour of treatment differences. When Mantel and Haenszel developed this procedure they calculated that when the treatments are identical the probabilities associated with its values follow a χ^2_1 distribution. This is irrespective of the number of outcome categories, and the test is sometimes referred to as the *chi-square one degree of freedom test for trend*.

In example 4.1, comparison of the test statistic value with χ^2_1 gives $p = 0.006$, a very highly significant result.

A significant *p*-value coming out of this test is indicative of a shift or trend in one direction across these categories for one treatment compared to the other. In the example it is illustrative to look at the percentages in the various categories as shown in Table 4.10.

In the L + F treatment group there has been a shift towards the improvement end of the scale compared to the L + P treatment group and it is this trend that the MH chi-square test has picked up. The Pearson chi-square test does not look for such trends and would have simply compared the 19 per cent with the 27 per cent, the 68 per cent with the 65 per cent and the 13 per cent with the 7 per cent in an overall, average way. Had the ordering been ignored and the Pearson chi-square test applied (incorrectly) then the *p*-value would have been 0.023. Although this is still statistically significant indicating treatment differences, the *p*-value is very different from the correct p-value of 0.006 and in many cases such differences would lead to incorrect conclusions.

Example 4.1: Flutamide plus Leuprolide compared to Leuprolide alone in the Treatment of Prostate Cancer

The following data (Table 4.9) are taken from a randomised controlled add-on trial (Crawford *et al.* (1989)) comparing Leuprolide + placebo (L + P) with Leuprolide + Flutamide (L + F) in prostate cancer in terms of improvement in pain at week 4.

Table 4.9 Flutamide data

	Improved	No change	Worse	Total
L + P	50	180	33	263
L + F	73	174	20	267
Total	123	354	53	530

Table 4.10 Flutamide data as percentages

	Improved	No change	Worse	Total
L + P	19%	68%	13%	263
L + F	27%	65%	7%	267

As with binary and categorical data, is there an issue with small sample sizes? Well, in fact, no, there is not. The MH test is a different kind of chi-square test and is not built around expected frequencies. As a consequence it is not affected by small expected frequencies and can be used in all cases for ordinal data. There are some pathological cases where it will break down but these should not concern us in practical settings.

4.7.3 Measures of treatment benefit for categorical and ordinal data

Measures such as the difference in event rates, OR, RR, RRR and NNT do not easily translate into the categorical data context. If we want to construct such measures in these cases we would collapse the outcome categories to two, the binary case, and proceed as before. In the categorical example covered earlier this could involve collapsing categories A, B and C to produce a binary outcome death/survival.

If the categorical outcome was 'Main Reason for Discontinuation' which was classified as Adverse Event, Withdrawal of Consent, Protocol Violation, Other, there may be interest in expressing an OR in relation to Adverse Event, in which case we would collapse the other three categories and proceed as in the binary case.

For ordinal data we could follow the categorical data recommendations by collapsing adjacent outcomes. Another approach which is sometimes used is to work with the so-called *common odds ratio*. With three outcome categories, as in the example, this would involve forming two 2×2 contingency tables by firstly collapsing 'Improved' and 'No change' and secondly collapsing 'No change' and 'Worse'. In each of these two tables we calculate the odds ratio; the common odds ratio is then obtained as an 'average' of the two. In our example the two ORs are 1.77 and 1.60 so the common odds ratio would be somewhere in the middle. The averaging process is a little complex and we will not go into details here. The common odds ratio is then the odds in favour of success on average, however you define success; 'Improved' or a combination of 'Improved' and 'No change'.

4.8 Extensions for multiple treatment groups

In this section we will discuss the extension of the t-tests for continuous data and the chi-square tests for binary, categorical and ordinal data to deal with more than two treatment arms.

4.8.1 Between-patient designs and continuous data

In this setting there is a technique, termed *one-way analysis of variance (one-way ANOVA)*, which gives an overall *p*-value for the simultaneous comparison of all of the treatments. Suppose, for example, we have four treatment groups with means μ_1, μ_2, μ_3 and μ_4. This procedure gives a *p*-value for the null hypothesis:

$$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4$$

against the general alternative hypothesis:

$$H_1: \quad \text{the opposite of } H_0, \text{ that there are differences somewhere}$$

A significant *p*-value from this test would cause us to reject the null hypothesis, but the conclusion from this only tells us that there are some differences somewhere; at least two of the μs are different. At that point we would want to look to identify where those differences lie and this would lead us to pairwise comparisons of the

treatment groups and reverting to a series of unpaired t-tests. It could be argued that the question posed by the one-way analysis of variance technique is of little value and that it is more relevant to start directly with a set of structured questions relating to the comparisons of pairs of treatments.

For example, let us suppose that we have three treatment groups: test treatment (μ_1), active control (μ_2), and placebo (μ_3). In a superiority setting there are two questions of interest:

1. Does the test treatment work?

$$H_0: \ \mu_1 = \mu_3 \quad H_1: \mu_1 \neq \mu_3$$

2. Is the test treatment better than the control treatment?

$$H_0: \ \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

Both of these questions are answered by the unpaired t-test.

One small advantage, however, of one-way analysis of variance is that it uses all of the data from all of the treatment groups to give us a measure of noise, so even when we are comparing the test treatment group with the control treatment group we are using information on the patient-to-patient variation in the placebo group to help estimate the noise. We work with a pooled estimate of the standard deviation. There are ways of adapting the unpaired t-test to incorporate this broader base of information, but even then the gains are small except in very small trials where information regarding the noise is at a premium. See Julious (2005) for further details.

In summary, there is not much to be gained in using one-way analysis of variance with multiple treatment groups. A simpler analysis structuring the appropriate pairwise comparisons will more directly answer the questions of interest. One final word of caution though; undertaking multiple comparisons in this way raises another problem, that of multiplicity. For the time being we will put that issue to one side; we will, however, return to it in Chapter 10.

4.8.2 Within-patient designs and continuous data

In this context, each patient would be receiving each of the multiple treatments. In the cross-over trial with three treatments this would likely be a three-period, three-treatment design and patients would be randomised to one of the six sequences; ABC, ACB, BAC, BCA, CAB or CBA. Although there are again ways of asking a simultaneous question relating to the equality of the three treatment means through an analysis of variance approach this is unlikely to be of particular relevance; questions of real interest will concern pairwise comparisons.

Comments as above for the between-patient designs apply also for the within-patient designs and in many cases the best approach will be to focus on a sequence of pairwise comparisons using the paired t-test.

In the context of within-patient designs, for example the multi-period cross-over, there could however, be some additional considerations. Such designs are frequently used in phase I where sample sizes are small and the gains afforded by the common estimation of standard deviation could well be worthwhile, so we should not by any means dismiss these methods completely.

4.8.3 Binary, categorical and ordinal data

As with continuous data the most relevant questions for binary, categorical and ordinal data will invariably relate to pairwise comparisons of treatments. We mentioned earlier that one minor advantage was the ability to use information from all of the treatment groups to estimate patient-to-patient variation assuming a common standard deviation across the treatment groups. With binary, categorical and ordinal data, however, even this small advantage does not apply; there is no standard deviation of this kind with these data types. The recommendation again, therefore, is to focus on the chi-square procedures developed earlier in this chapter for pairwise treatment comparisons.

4.8.4 Dose ranging studies

The discussion so far in this section has assumed that the treatment groups are unordered. There are, however, situations where these multiple treatment groups correspond to placebo and then increasing dose levels of a drug. It could still be in these circumstances that we are looking to compare each dose level with placebo in order to identify, for example, the minimum effective dose and again we are back to the pairwise comparisons.

There will be, however, some circumstances where we are just interested in trends; if we increase the dose does the mean response increase?

For continuous data there is a procedure within the one-way analysis of variance methodology that is able to focus on this; we would be looking for a trend across the treatment groups.

For binary, categorical and ordinal data there is also an approach which is a further form of the Mantel–Haenszel chi-square test. You will recall that the MH test is used for ordinal responses comparing two treatments. Well, this procedure generalises to allow ordering across the treatment groups in addition, for each of

the binary, categorical and ordinal data types. More details for binary, categorical and ordinal data can be found in Stokes, Davis and Koch (1995).

4.8.5 Further discussion

For the remainder of this book as we investigate further designs and methods of analysis we will focus our developments on two treatment group comparisons. When we have more than two treatment groups, our questions are usually in relation to pairwise comparisons in any case, as discussed above, and these can be handled directly by reducing to those specific evaluations. For binary, categorical and ordinal data this is precisely the approach. For continuous data there are some advantages in efficiency in using a combined estimate of the standard deviation from the complete experiment and this is what is usually done.

Specific mention will be made, however, in multiple treatment group settings where issues arise which require considerations outside of these.

5
Multi-centre trials

5.1 Rationale for multi-centre trials

As indicated in the ICH E9 guideline there are two reasons why we conduct multicentre trials:

- To recruit sufficient numbers of patients within an appropriate timeframe
- To allow the evaluation of the homogeneity of the treatment effect and provide a basis for generalisability

The first issue is of practical importance; there is probably no other way the required numbers of patients could be recruited. The second issue has more of a statistical basis. A multi-centre structure enables us to look at treatment differences in different centres or clusters of centres to assess whether what we are seeing is a consistent effect. Without this consistency it would be very difficult to draw conclusions about the value of the treatment across a broad patient population.

From a practical perspective the randomisation scheme may well have been stratified by centre. This would certainly be the case in trials with a small number of large centres. As mentioned in Chapter 1 this avoids potential confounding between treatment and centre. The stratification also has an impact on the way the data are analysed and from a statistical perspective it is necessary to take stratification factors into account in the analysis. We will see shortly how this is done. In contrast, some trials are conducted across a large number of small centres, for example GP studies. It is unlikely in these cases that the randomisation would be stratified by centre although it may be that groupings of centres have been defined at the design stage (for example by geographical region) and the randomisation is stratified by those centre groupings. Under these circumstances

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

the groupings form pseudo-centres and these would be taken into account in the statistical analysis.

CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'

'When the number of patients within each centre is expected to be very small, it may not be practical to stratify the randomisation by centre. In that case it should be considered whether randomisation could be stratified by, for example, country or region. Such a choice might be driven by similarities in co-medication, palliative care or other factors that might make stratification advisable.'

5.2 Comparing treatments for continuous data

Consider the following hypothetical data with four centres and two treatment groups where the primary endpoint is the reduction (baseline minus final) in diastolic blood pressure (mmHg). The entries in Table 5.1 are means. These means are also plotted in Figure 5.1.

Across each of the centres, patients in treatment group A are performing better than those in treatment group B and the treatment difference is fairly consistent; the difference is largest (7.1 mmHg) in centre 3, while the difference in centre 4 (6.1 mmHg) is the smallest. There are, however, differences in overall performance; patients in centre 3 are doing the best overall while those in centre 2 are doing the worst.

Our main interest is still simply to compare the treatments, but we must recognise that it would not be unusual to see centre differences in an overall sense; different standards of ancillary care, cultural and environmental differences are just some of the things that could contribute to this.

Statistical analysis proceeds through *two-way analysis of variance (ANOVA)*. The focus in this methodology is to compare the treatment groups while recognising potential centre differences. To enable this to happen we allow the treatment means μ_A and μ_B to be different in the different centres as seen in Table 5.2.

Table 5.1 Sample means in a multi-centre trial

Centre	Treatment A	Treatment B	Difference (A − B)
1	12.4	5.8	6.6
2	9.7	3.0	6.7
3	14.6	7.5	7.1
4	10.0	3.9	6.1

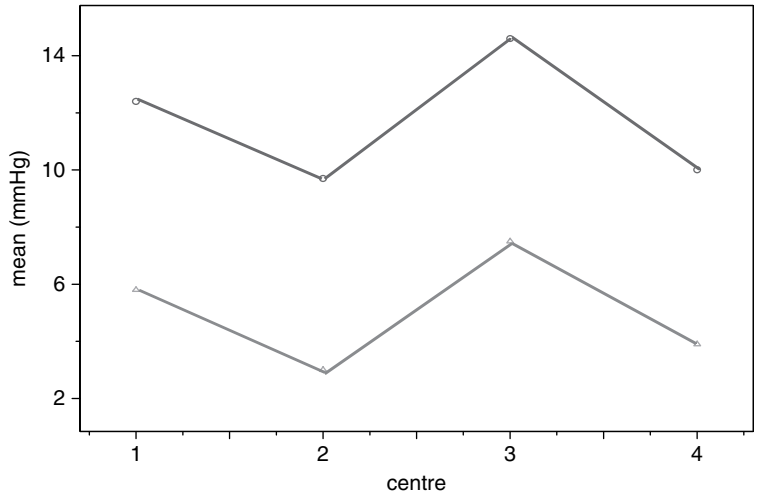


Figure 5.1 Multi-centre trial – homogeneous treatment effects

Table 5.2 True means in a multi-centre trial

Centre	Treatment A	Treatment B	Difference
1	μ_{A1}	μ_{B1}	$\mu_{A1} - \mu_{B1}$
2	μ_{A2}	μ_{B2}	$\mu_{A2} - \mu_{B2}$
3	μ_{A3}	μ_{B3}	$\mu_{A3} - \mu_{B3}$
4	μ_{A4}	μ_{B4}	$\mu_{A4} - \mu_{B4}$

The null hypothesis we evaluate is then:

$$H_0: \mu_{A1} = \mu_{B1} \text{ and } \mu_{A2} = \mu_{B2} \text{ and } \mu_{A3} = \mu_{B3} \text{ and } \mu_{A4} = \mu_{B4}$$

against the general alternative hypothesis H_1 that there are differences somewhere.

This null hypothesis is saying that the treatment means are the same within each centre, but not necessarily across centres; we are allowing the centres to be different. Two-way ANOVA gives us a p-value relating to this null hypothesis. If that p-value is significant ($p \leq 0.05$) then we reject the null hypothesis and there is evidence that the treatments are different.

Initially you may think that this approach to comparing the treatments is over elaborate; why not just compare the overall means (which happened to be 11.8 and 4.9) in an unpaired t-test? Well in one sense you could, but it would not be the most efficient thing to do. Simply comparing the overall means loses the fact that the means in centre 1, 12.4 and 5.8, are linked, as are those in each of the other centres. The two-way ANOVA procedure maintains that link, it simultaneously compares

each A mean with the corresponding B mean and comes up with an average, overall comparison. To be more precise, it takes as its signal, the average of the column of mean differences; although it is not the straight average, but a weighted average according to the size of the centre, with differences from the larger centres having more weight. Similarly the patient-to-patient variation is measured by the standard deviation and the key element in the noise is taken from a weighted combination of the individual standard deviations in each of the eight cells of the table.

This analysis assumes that the treatment effect is consistent across the centres. For the above data this seems a reasonable assumption, but we will return to a more formal evaluation of this assumption in the next section. The weighted average of the treatment differences that is the basis of the signal provides the best estimate of the overall treatment effect. In the above example this was 6.74 mmHg and we can construct confidence intervals around this value to allow an interpretation of the size of the (assumed common) true treatment effect.

5.3 Evaluating the homogeneity of the treatment effect

5.3.1 Treatment-by-centre interactions

The ICH Guideline is quite clear on the need to investigate the consistency of the treatment effect.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'If positive treatment effects are found in a trial with appreciable numbers of subjects per centre, there should generally be an exploration of the heterogeneity of treatments effects across centres, as this may affect the generalisability of the conclusions.'

The guideline then goes on to recommend ways in which this can be done.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Marked heterogeneity may be identified by graphical display of the results of individual centres or by analytical methods, such as a significance test of the treatment by centre interaction.'

We will firstly discuss a significance test for the treatment-by-centre interaction and then talk about possible graphical methods.

Consider the (hypothetical) data in Table 5.3 set out as before and displayed in Figure 5.2.

Table 5.3 Sample means in a multi-centre trial with a non-constant treatment difference

Centre	Treatment A	Treatment B	Difference (A − B)
1	12.4	5.8	6.6
2	9.7	6.0	3.7
3	14.6	7.5	7.1
4	5.0	3.9	1.1

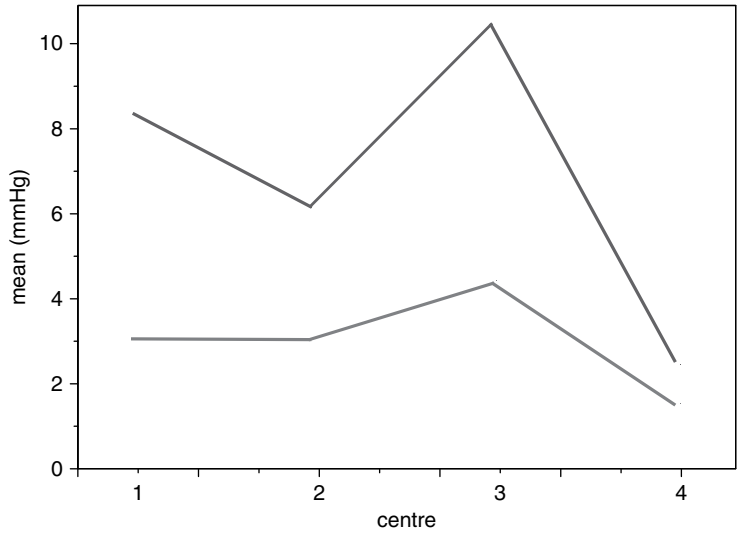


Figure 5.2 Multi-centre trial – heterogeneous treatment effects

In these data, treatment A is performing consistently better than treatment B as earlier, but now the extent of the difference is not consistent. There are large differences in centres 1 and 3 with smaller differences seen in centres 2 and 4.

To assess this consistency, two-way ANOVA additionally provides a p-value for the hypothesis:

$$H_0: \mu_{A1} - \mu_{B1} = \mu_{A2} - \mu_{B2} = \mu_{A3} - \mu_{B3} = \mu_{A4} - \mu_{B4}$$

against the general alternative hypothesis that the treatment differences are not all equal.

This null hypothesis is saying that the treatment difference/effect is consistent. If the *p*-value from this test is significant then we talk in terms of having a significant *treatment-by-centre* (or a *treatment* × *centre*) *interaction*. Power and sample size calculations (see later chapter on this topic) will have focused

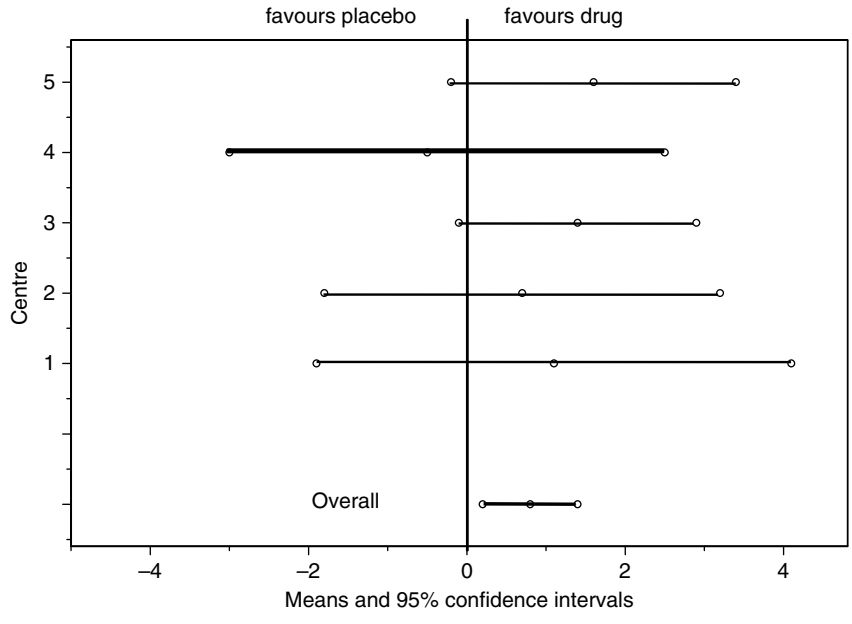


Figure 5.3 Displaying individual centre effects and the overall effect

on testing the main effect of treatment and not on the evaluation of the treatment-by-centre interaction. For this reason the test for this interaction will have low power and only pick up marked heterogeneity. To counter this to an extent, we would normally evaluate the significance of the interaction test at the 10 per cent level, so declare significant heterogeneity whenever $p \leq 0.10$ rather than when $p \leq 0.05$. In discussing the treatment-by-centre significance test, ICH states:

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'When using such a statistical significance test, it is important to recognise that this generally has low power in a trial designed to detect the main effect of treatment.'

The ICH Guideline also mentions the use of graphical methods and Figure 5.3 is the kind of plot they are looking for, displaying the treatment difference between the means, within each centre, together with a 95 per cent confidence interval. These graphs are very much like those used in Meta Analysis (see Chapter 15). A substantial amount of overlap of the confidence intervals across centres indicates no major concerns in relation to the homogeneity of treatment difference.

5.3.2 Quantitative and qualitative interactions

ICH makes the distinction between quantitative and qualitative interactions. A *quantitative interaction* refers to the situation where the treatment difference is consistently in one direction (for example A is always better than B), but there are differences in terms of magnitude. A *qualitative interaction* is where the treatment difference is in a different direction for some centres (for example A is better than B in some centres but B is better than A in other centres). The previous example is an example of a quantitative interaction. Had the A mean in centre 4 been 3.9 and the B mean 5.0, then the treatment difference (A − B) would equal −1.1 and we would have a qualitative interaction.

If heterogeneity of treatment effect is found then this could possibly undermine the generalisability of the results. For example, with a qualitative interaction, one treatment is performing better in some centres but worse in other centres and it will be difficult to draw a general conclusion of 'A is a better treatment than B'. Even a quantitative interaction will sometimes give problems in terms of estimating with confidence the magnitude of the treatment effect. ICH gives some guidance:

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'If heterogeneity of treatment is found, this should be interpreted with care and vigorous attempts should be made to find an explanation in terms of other features of trial management or subject characteristics . . . In the absence of an explanation, heterogeneity of treatment effect as evidenced, for example, by marked quantitative interactions implies that alternative estimates of the treatment effect may be required, giving different weights to the centres, in order to substantiate the robustness of the estimates of treatment effect. It is even more important to understand the basis of any heterogeneity characterised by marked qualitative, and failure to find an explanation may necessitate further clinical trials before the treatment effect can be reliably predicted.'

As ICH points out more work is needed to find an explanation, when interactions are seen. That explanation may come from looking, for example, at differential compliance within the centres or different characteristics of the patient sample recruited at the different centres. Inevitably a treatment-by-centre interaction is simply a surrogate for some underlying explanation. It must also be said that the explanation may be 'chance'! In a trial with a resonable number of centres, where A really is a better treatment than B, seeing one treatment reversal would not be unlikely. Senn (1997, Section 14.2.6) investigates this and shows, under some fairly standard conditions, that the probability of seeing at least one reversal goes above 50 per cent with six or more centres, so be cautious with over interpretation with regard to interactions.

5.4 Methods for binary, categorical and ordinal data

The *Cochran–Mantel–Haenszel (CMH) tests* are a collection of procedures that extend the simple chi-squared tests introduced in Chapter 4 to incorporate the multicentre setting. Landis, Heyman and Koch (1978) provide further details.

For binary data in multi-centre trials we will have a series of 2×2 tables, one for each of the centres. For categorical and ordinal data with c categories, we will have a series of $2 \times c$ tables. The CMH test in the first instance provides a single p-value for the main effect of treatment.

In terms of summary statistics for binary data, we usually work with the odds ratio, averaged over the different centres. It is also possible to obtain in a similar way 'average' values for the reduction in event rates relative risk (RR) and relative risk reduction (RRR) together with their confidence intervals. Generalising the number needed to treat (NNT) to allow this averaging to take place is more difficult and is generally not done. Should such a quantity be required then a combined 2×2 table for the data as a whole would be used. For ordinal data we can work with the common odds ratio averaged over the centres, but this is computationally more difficult and as yet not included in the standard software packages.

The same issues arise with the heterogeneity of the treatment effect as above with continuous data and indeed the ICH comments detailed there are relevant for all data types. For binary data there is a significance test, the Breslow–Day test (Breslow and Day (1994)), which provides a p-value for the homogeneity of the treatment effect over the centres. Again, graphical methods are also available and follow the approach seen earlier, plotting estimated odds ratios for each centre together with their corresponding 95 per cent confidence intervals.

5.5 Combining centres

The ideal situation in a multi-centre trial is to have a small number of large centres (or pre-defined pseudo-centres). This gives the necessary consistency and control yet still allows the evaluation of heterogeneity. In practice, however, we do not always end up in this situation and combining centres at the data analysis stage inevitably needs to be considered. From a statistical perspective adjusting for small centres in the analysis is problematic and leads to unreliable estimates of treatment effect so we generally have to combine.

There are no fixed rules for these combinations but several points should be noted:

- Combining centres just because they are small or combining centres to produce centres of similar size has no scientific justification (see CPMP (2003) 'Points to Consider on Adjustment for Baseline Covariates').

- Combinations should be based on similarity (by region, by country, by type, . . .). For example, in Depression a trial may be run across many centres with at most ten patients being recruited at each centre; some of the centres will be GP centres while others will be specialist psychiatric facilities. In this case combining by centre type (GP or psychiatrist) would make sense and this would allow the homogeneity of treatment effect between GP centres and specialist psychiatric centres to be investigated.
- Ideally, rules for combining centres should be detailed in the Statistical Analysis Plan.
- Any final decisions regarding combinations should be made at the Blind Review stage prior to breaking the blind.

We will discuss the decision-making process with regard to the Statistical Analysis Plan and the Blind Review in Section 16.3.

6

Adjusted analyses and analysis of covariance

6.1 Adjusting for baseline factors

We saw in the previous chapter how to account for centre in treatment comparisons using two-way ANOVA for continuous data and the CMH test for binary, categorical and ordinal data. These are examples of so-called *adjusted analyses*; we have adjusted for centre differences in the analysis.

There may be other factors that we wish to adjust, for example, age, sex, baseline risk and so on and there are several reasons why we might want to do this.

Firstly, if the randomisation has been stratified for baseline variables then from a theoretical statistical point of view these variables should be taken into account in the analysis. Secondly, the efficiency of the statistical analysis can be improved in several ways if baseline prognostic factors (factors which influence outcome) are included in the analysis. Finally, it provides a framework for the investigation of the consistency of the treatment effect according to different values for those factors.

As an example, suppose that the randomisation has been stratified on the basis of age and sex with four strata:

Males, < 50 yrs
Males, ≥ 50 yrs
Females, < 50 yrs
Females, ≥ 50 yrs

An *adjusted* (or *stratified*) analysis for a continuous outcome variable would be two-way ANOVA. The four strata would be handled in exactly the same way as if there were four centres in a multi-centre trial. The ANOVA approach will also

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

provide an (adjusted) estimate of the treatment effect and associated confidence intervals as in the previous chapter when looking at multi-centre trials.

If the outcome variables were binary, categorical or ordinal then the CMH test would be used with age and sex as above defining the four strata. The emphasis here, is still of course, to compare the treatments, but adjustment for these baseline factors has improved the efficiency of the analysis by comparing the treatments within each stratum (like with like) and then averaging those effects over the strata, as in the multi-centre setting. Again we will come out of this analysis with, for example in the binary case, an adjusted OR and a corresponding confidence interval.

If we had not stratified the randomisation for age and sex, but nonetheless recognised that these were important prognostic factors, it would still have been advantageous to analyse the data in the same way using this adjusted methodology.

In the previous chapter we also investigated the presence of treatment-by-centre interactions in order to explore the heterogeneity of the treatment effect. These methods extend directly to the present setting and we can in the same way look to see if the treatment effect is consistent across the strata.

As the number of baseline factors increases, however, this approach becomes a little unwieldy. The method of analysis of covariance is a more general methodology that can deal with this increase in complexity and which can also give improved ways of exploring interactions. We will develop this methodology later in the chapter.

6.2 Simple linear regression

Regression provides a collection of methods that allow the investigation of dependence; how an outcome variable depends upon something that we record/measure at baseline.

As an example, suppose in an oncology study we wish to explore whether time to disease recurrence from entry (months) into the study depends upon the size of the primary tumour measured at baseline (diameter in cm). The scatter plot in Figure 6.1 represents (artificial) data on 20 subjects.

A visual inspection of the plot would suggest that there is some dependence, but in many cases this will not be quite so clear cut. We explore the dependence from a statistical point of view by fitting a straight line to the data.

The equation of a straight line is:

$$y = a + bx$$

where a is the intercept (the value of y where the line crosses the y-axis) and b is the slope (the amount by which y increases when x increases by one unit).

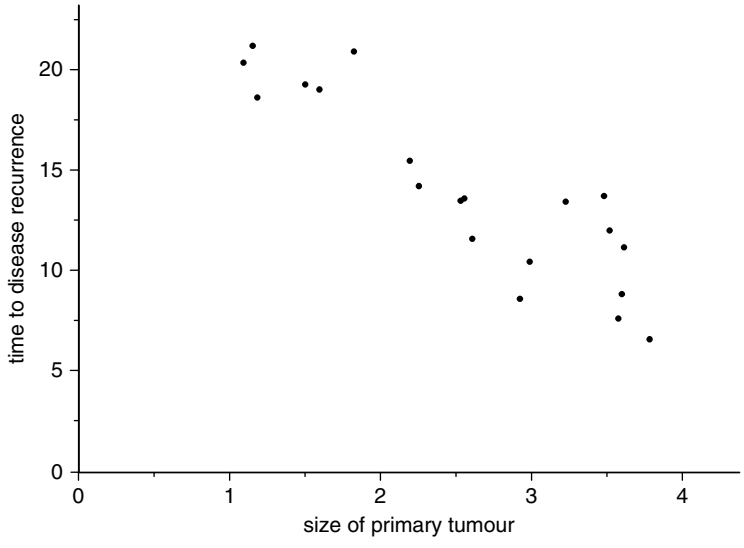


Figure 6.1 Scatter plot for dependence of time to disease recurrence on size of primary tumour

The value of b is of greatest importance. If b is positive then there is a positive dependence; as x increases then so does y. If b is negative then there is a negative dependence; as x increases then y decreases. Finally if $b = 0$ then there is no dependence; as x increases nothing happens to y.

The method that we use to fit the straight line so that it describes the data in the best possible way is called *least squares*. This involves measuring the vertical distance of each point from a line placed on the plot as shown in Figure 6.2, squaring each of those distances (this amongst other things gets rid of the sign) and choosing that line which makes the average of these squared distances as small as possible. In the above example, this *least-squares regression line* has the equation:

$$y = 25.5 - 4.48\,x$$

The value of the slope is -4.48 and this estimates the average increase in time to disease recurrence as the tumour size at baseline increases by 1 cm. The primary question of interest here is 'does time to disease recurrence depend upon tumour size at baseline?' To address this question we, as usual, formulate null and alternative hypotheses:

$$\mathrm{H}_0 :\ b = 0 \quad \mathrm{H}_1 :\ b \neq 0$$

and construct an appropriate test. This involves the signal, which is the estimate of b from the data, and a measure of noise. The measure of noise is the standard error of the estimate of b obtained from the data, and there is a formula for

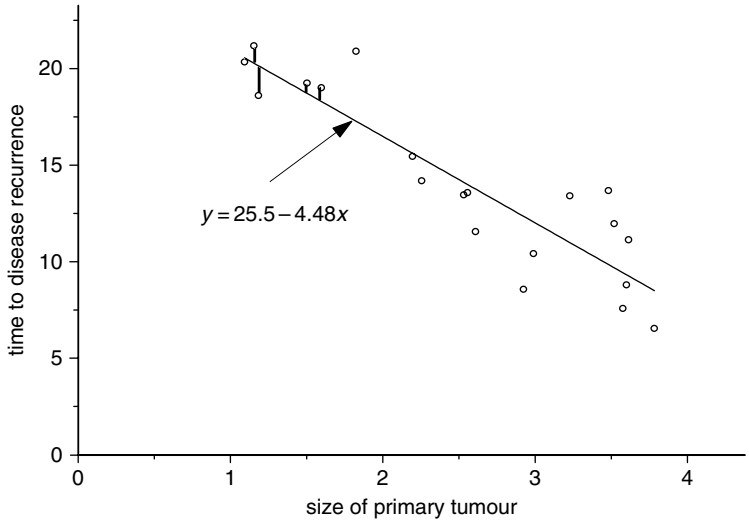


Figure 6.2 Least-squares regression line

this. Finally the test statistic is as before, the signal divided by the noise and the null distribution is t_{n-2}, where n is the number of subjects. A significant p-value ($p \leq 0.05$) from this test is telling us that b is significantly different from zero, indicating dependence. A non-significant p-value is telling us that there is insufficient evidence to conclude dependence. In fact for these data $p << 0.001$, a very highly significant dependence of time to disease recurrence on the size of the primary tumour at baseline. The slope of -4.48 indicates that on average for each 1 cm increase in the diameter of the primary tumour there is a 4.48 months decrease in the time to disease recurrence.

This technique of *simple linear regression* therefore provides a simple way to evaluate whether a particular baseline variable is predictive of outcome. We will extend these ideas in the next section to evaluate several baseline variables/factors simultaneously.

6.3 Multiple regression

In the previous section we saw how to study the dependence of an outcome variable on another variable measured at baseline. It could well be that there are several baseline variables which predict outcome and in this section we will see how to incorporate these variables simultaneously through a methodology termed *multiple (linear) regression*.

Taking up the example from the previous section it may be that time to disease progression depends potentially not just on size of the primary tumour, but also

on age and sex and we would like to explore the nature of that dependence. Clearly size of primary tumour and age are both numerical while sex is not; we incorporate qualitative variables of this kind by using so-called *indicator variables*. Generally these take the values zero and one according to the 'value' of the variable. It also does not matter which way round they are coded. Switching the codes would simply result in the coefficient of that variable changing sign.

$$\begin{aligned} \text{Let} \quad & x_1 = \text{size of primary tumour} \\ & x_2 = \text{age} \\ & x_3 = 0 \quad \text{male} \\ & \qquad 1 \quad \text{female} \end{aligned}$$

The extension of simple linear regression to deal with multiple baseline variables is somewhat difficult visually, but algebraically it is simply a matter of adding terms to the equation:

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Now b_1 measures the effect of size on time to disease recurrence while the coefficients b_2 and b_3 measure the effects of age and sex respectively. More specifically, b_1 and b_2 are the changes in the average time to disease recurrence as size and age each increase by one unit, respectively, while b_3 measures the sex effect; the average time to disease recurrence for females minus that for males. Each of these quantities, which we can estimate from the data again using the method of least squares, represents the contribution of each of the variables separately in the presence of the other variables.

The questions of interest revolve around the values of the b coefficients. Do age and size of primary tumour predict time to disease recurrence? Is there a sex difference? We address these questions by formulating hypotheses:

$$\begin{aligned} &H_{01}: \ b_1 = 0 \quad H_{11}: \ b_1 \neq 0 \\ &H_{02}: \ b_2 = 0 \quad H_{12}: \ b_2 \neq 0 \\ &H_{03}: \ b_3 = 0 \quad H_{13}: \ b_3 \neq 0 \end{aligned}$$

Each of these is evaluated by dividing the estimate of the corresponding b value by its standard error and comparing to the t_{n-4} distribution (the degrees of freedom for the appropriate t shape is number of subjects minus [1 + the number of x variables in the model]). Note that multiple regression with just a single variable reduces to simple linear regression.

Suppose that the fitted equation turns out to be:

$$y = 23.8 - 3.51x_1 - 1.74x_2 + 0.47x_3$$

Therefore for a 1 cm increase in the tumour size, time to disease recurrence reduces by on average an estimated 3.51 months; for each additional year in age there is on average an estimated 1.74 month reduction in time to disease recurrence and finally the time to disease recurrence is estimated to be slightly higher for females by, on average, 0.47 months. It is also fairly straightforward to construct confidence intervals around these estimates.

The p-values associated with each term in this model were:

$$\begin{aligned} H_{01}: \ b_1 &= 0 \quad p = 0.007 \\ H_{02}: \ b_2 &= 0 \quad p = 0.02 \\ H_{03}: \ b_3 &= 0 \quad p = 0.48 \end{aligned}$$

This suggests that size of primary tumour and age are important predictors of time to disease recurrence while sex looks to be unimportant.

Note that this approach is not the same as conducting three linear regression analyses on the baseline variables separately. In fact such an approach could give a confused picture if those variables were correlated. For example, suppose that age and size of primary tumour are correlated with older patients tending to present with larger tumours. Also suppose that it is size of primary tumour that is the driver in terms of time to disease recurrence and that age has no effect additional to that. The separate linear regressions would indicate that both size of primary tumour and age predicted outcome; age would be identified as a predictor of outcome only as a result of its correlation with size. Multiple regression, however, would give the correct interpretation. Size of primary tumour would be seen as a predictor of outcome, but once that effect is accounted for, age would add nothing to the prediction of outcome.

With a large number of potential baseline variables it may be of interest to select those variables that are impacting on outcome and methods (*stepwise regression*) are available for doing this. Using this methodology the unimportant variables are eliminated leaving a final equation containing just the important prognostic factors.

6.4 Logistic regression

Multiple regression as presented so far is for continuous outcome variables y. For binary, categorical and ordinal outcomes the corresponding technique is called *logistic regression*. Suppose that in our earlier example we defined success to be 'disease-free for five years' then we might be interested identifying those variables/factors at baseline that were predictive of the probability of success.

Define y now to take the value one for a success and zero for a failure. For mathematical reasons, rather than modelling y as we did for continuous outcome variables, we now model the probability that $y = 1$, written $\text{pr}(y = 1)$.

This probability, by definition, will lie between zero and one, so to avoid numerical problems we do not model $\text{pr}(y=1)$ itself but a transformation of $\text{pr}(y=1)$, the so-called *logit* or *logistic transform*:

$$\ell\text{n}\left\{\frac{\text{pr}(y=1)}{[1-\text{pr}(y=1)]}\right\} = a + b_1x_1 + b_2x_2 + b_3x_3$$

Computer packages such as SAS can fit these models, provide estimates of the values of the b coefficients together with standard errors, and give p-values associated with the hypothesis tests of interest. These hypotheses will be exactly as H_{01}, H_{02} and H_{03} in Section 6.3. Methods of stepwise regression are also available for the identification of a subset of the baseline variables/factors that are predictive of outcome.

The logistic regression model extends both to categorical data using the *polychotomous logistic model* and to ordinal data using the *ordinal logistic model*. For an example of the former see Marshall and Chisholm (1985) in the area of diagnosis.

6.5 Analysis of covariance for continuous data

6.5.1 Main effect of treatment

We will return to the example where we have just a single baseline variable, size of primary tumour, predicting the outcome, time to disease recurrence, but now in addition we have randomised the patients to one of two treatment groups, test treatment and placebo.

Figure 6.3 displays a possible pattern for the results. Note firstly that as before there appears to be a dependence of time to disease recurrence on the size of the primary tumour. In addition it also seems that the patients receiving the test treatment have longer times to disease recurrence compared to those receiving the control treatment, irrespective of the size of the primary tumour.

A formal comparison of the two treatments could be based on the unpaired t-test, comparing the mean time to disease recurrence in the test treatment group with the mean time to disease recurrence in the control group. While this is a valid test, it may not be particularly sensitive. The separation between the two groups is clear, but if we now simply read off the times to disease recurrence on the y-axis we will see considerable overlap between the groups; we will have lost some sensitivity by ignoring the size of the primary tumour variable.

Consider an alternative approach; applying simple linear regression to the data from these two groups of patients. The equations of these lines can be written:

$$\begin{aligned} y &= a_1 + bx \qquad \text{test treatment} \\ y &= a_2 + bx \qquad \text{placebo} \end{aligned}$$

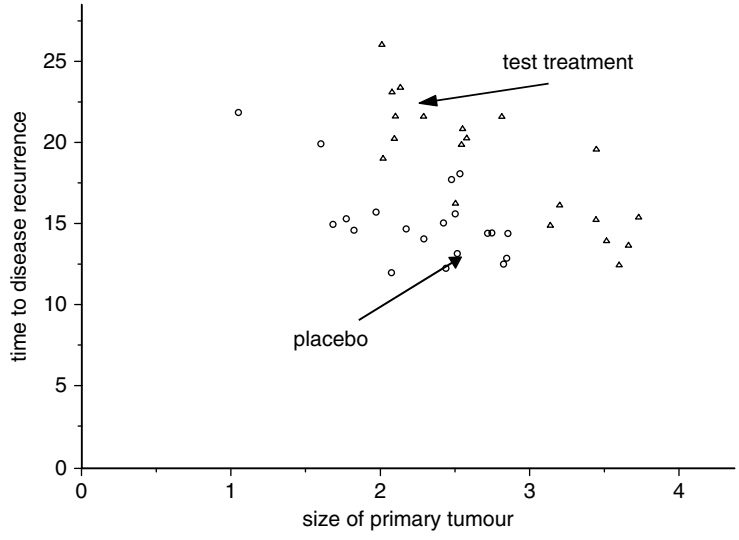


Figure 6.3 Scatter plot for two treatment groups

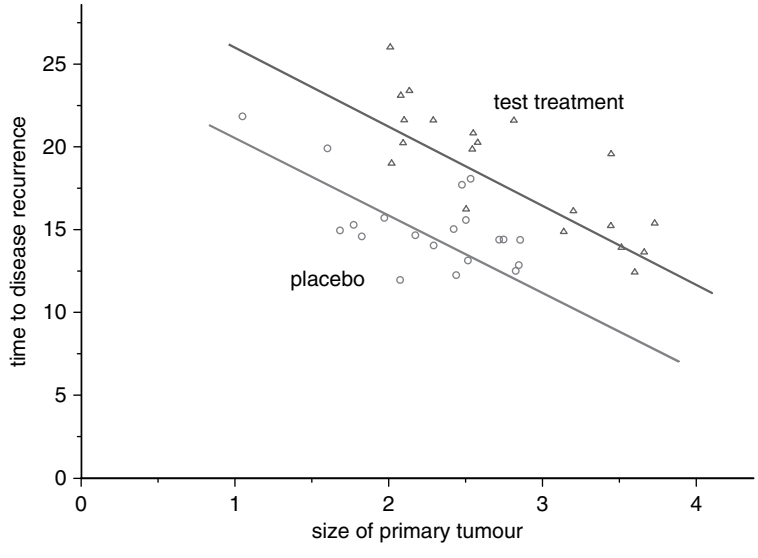


Figure 6.4 Scatter plot and fitted lines for two groups

Figure 6.4 shows these lines fitted to the data. Note that we have constrained the slopes of these lines to be the same; we will return to this point later. The intercepts, a_1 and a_2, are the points where the lines cross the y-axis.

Had the treatments been equally effective then the points in the placebo group would not have been, in general, below the points in the test treatment group.

The lines would be co-incidental with $a_1 = a_2$. Indeed the larger the treatment difference the bigger the difference between the two intercepts, a_1 and a_2. Our main interest is to compare the treatments and within this framework we compare the values of a_1 and a_2 through the null hypothesis $H_0: a_1 = a_2$ and the alternative hypothesis $H_1: a_1 \neq a_2$. The signal is provided by the estimate of $a_1 - a_2$ and the noise is the standard error of that estimate; we compare the signal-to-noise ratio to t_{n-3} to give the *p*-value.

The quantity $a_1 - a_2$ is the vertical distance between the lines and represents the (adjusted) difference in the mean time to recurrence in the test treatment group minus the mean time to recurrence in the control group; the treatment effect. It is straightforward also to obtain a confidence interval around this adjusted treatment effect to capture the true difference.

This technique is called *analysis of covariance (ANCOVA)* and size of the primary tumour is termed the *covariate*. Taking account of the covariate here has led to a much more powerful analysis than that provided by the simple unpaired t-test. Of course the main reason why we are seeing such an improvement in sensitivity is that the covariate is such a strong predictor of outcome. These improvements will not be quite so great with weaker predictors.

It is also possible to include more than one covariate in the analysis in cases where several are thought to be influential for the outcome by simply adding on terms to the above equations as with multiple regression. That is:

$$\begin{aligned} y &= a_1 + b_1x_1 + b_2x_2 + b_3x_3 \qquad \text{test treatment} \\ y &= a_2 + b_1x_1 + b_2x_2 + b_3x_3 \qquad \text{placebo} \end{aligned}$$

Again we test the hypothesis $H_0: a_1 = a_2$. The estimate of $a_1 - a_2$ is used as the signal and its standard error as the noise and now this ratio is compared to t_{n-q}, where n is the number of subjects and q is the number of covariates $+1$ to give a *p*-value.

We call these equations (or models), *main effects models*. In the next subsection we will be adding to the main effects (of treatment and the covariates), treatment-by-covariate interaction terms.

6.5.2 Treatment-by-covariate interactions

Returning to the case with a single covariate we have assumed that the two lines are parallel. This may not be the case. Figure 6.5 shows a situation where it would not be appropriate to assume parallel lines. Here patients presenting with small tumours do much better in the test treatment group compared to the placebo group, but there are virtually no differences between the treatments for those patients presenting with large tumours.

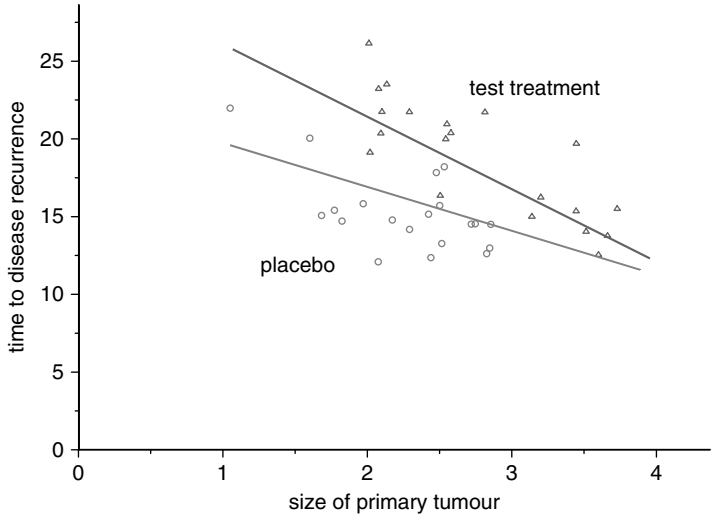


Figure 6.5 Scatter plot showing treatment-by-covariate interaction

We have previously discussed treatment-by-centre interactions, where the treatment effect is seen to be different in the different centres. We now have a treatment-by-covariate interaction, where the treatment effect depends on the value of the covariate. We can investigate this by considering a more general model where we allow the lines to have different slopes as well as different intercepts:

$$y = a_1 + b_1 x \qquad \text{test treatment}$$
$$y = a_2 + b_2 x \qquad \text{placebo}$$

This now gives us the opportunity to assess whether there are any interactions by fitting these two lines to the data and formulating a test of the hypothesis $H_0: \ b_1 = b_2$ against the alternative $H_1: \ b_1 \neq b_2$. A significant p-value, and as with treatment-by-centre interactions we would looking at $p < 0.10$ as 'significant', would indicate the presence of an interaction. In this case talking in terms of the 'treatment effect' makes little sense as a consistent treatment effect does not exist. A non-significant p-value would suggest that it is safe to assume that the treatment effect is fairly constant across the levels of the covariate, and the model, with a common slope b, provides an adequate description of the data.

If a significant treatment-by-covariate interaction is found then it would be useful to divide the patients into subgroups in terms of the size of the primary tumour, say small, medium and large and look at the treatment difference within those subgroups to try to better understand the nature of the treatment-by-covariate interaction.

In the presence of several covariates there will be a series of b coefficients, two for each covariate as follows:

$$\begin{aligned} y &= a_1 + b_{11}x_1 + b_{12}x_2 + b_{13}x_3 \quad \text{test treatment} \\ y &= a_2 + b_{21}x_1 + b_{22}x_2 + b_{23}x_3 \quad \text{placebo} \end{aligned}$$

Assessing the treatment-by-covariate interactions is then based on comparing the bs: b_{11} with b_{21}, b_{12} with b_{22} and b_{13} with b_{23} in separate hypothesis tests.

6.5.3 A single model

These models can be written in a more precise form by defining a binary indicator to denote treatment. Let $z = 0$ for patients randomised to the placebo group and let $z = 1$ for patients randomised to the test treatment. The model with a single covariate and assuming a common slope can then be written:

$$y = a + c\,z + b\,x$$

Now, when $z = 0$ (placebo group), $y = a + bx$ and when $z = 1$ (test treatment group), $y = (a + c) + bx$.

The b in this model is as previously, but now $a = a_1$ and $c = a_2 - a_1$. We refer to this in mathematics as a re-parameterisation, don't be put off by it! The hypothesis $H_0: \ a_1 = a_2$ is now replaced by the hypothesis $H_1: \ c = 0$. None of this changes the analysis in any sense, it is just a more convenient way to write down the model and will be useful later when we bring together the ideas of ANOVA and ANCOVA.

Although conceptually it is useful to think of fitting straight lines to each of the treatment groups separately, this is not in practice how it is done. We simply fit the single equation above. This also allows us to use the information on the noise from the two groups combined to obtain standard errors. Finally, we build in the interaction terms by adding on to this common equation a *cross-product term*, $z \times x$ and using another re-parameterisation:

$$y = a + c\,z + b\,x + d\,zx$$

so when $z = 0$ (control treatment group), $y = a + bx$ and when $z = 1$ (test treatment group), $y = (a + c) + (b + d)x$.

b_1 in the previous model is now b in this common model while $d = b_2 - b_1$. Assessing the presence of a treatment-by-covariate interaction (common slope) is then done through the hypothesis that $d = 0$.

For several covariates we simply introduce a cross-product term for each covariate with corresponding coefficients d_1, d_2 and d_3. The presence of treatment-by-covariate interactions can then be investigated through these coefficients.

6.5.4 Connection with adjusted analyses

Analysis of covariance is a form of adjusted analysis; we are providing an adjusted treatment effect in the presence of covariates. This is very much like the adjusted analysis we presented earlier on in this chapter. For the single covariate example, had we defined strata according to the size of the primary tumour, say small, medium and large, and then undertaken two-way ANOVA to compare the treatments, then we would have got very similar results to those seen here through ANCOVA. This applies to the *p*-values for the assessment of treatment difference, the estimated (adjusted) treatment difference, the associated confidence intervals and the *p*-values for studying the homogeneity of the treatment effect, which is simply looking for treatment-by-covariate interactions.

6.5.5 Advantages of analysis of covariance

Analysis of covariance offers a number of advantages over simple two treatment group comparisons:

- Produces improvements in efficiency (smaller standard errors, narrower confidence intervals, increased power).
- Corrects for baseline imbalances. Randomisation will, on average, produce groups that are comparable in terms of baseline characteristics. It is inevitable, however, that small differences will still exist and if these are differences in important prognostic factors then they could have an impact on the treatment comparisons. By chance there will be occasions also where substantial imbalance exists. Figure 6.6 illustrates such a case and it can be seen how such imbalances could cause bias in our evaluation of a treatment effect. Here, purely by chance, we have ended up with a predominance of patients with large tumours in the test treatment group while in the control treatment group there are many more patients with small and moderate size tumours. A simple unpaired t-test would possibly fail to detect a treatment difference or even conclude a difference in the wrong direction as a result of this baseline imbalance. ANCOVA helps correct for those baseline imbalances; once the two regression lines are estimated then ANCOVA 'ignores' the data and simply works with the distance between the lines for the treatment effect.
- Allows assessment of prognostic factors. Fitting the ANCOVA model provides coefficients for the covariates and although this is not the primary focus of the analysis, these coefficients and associated confidence intervals provide information on the effect of the baseline covariates on outcome.

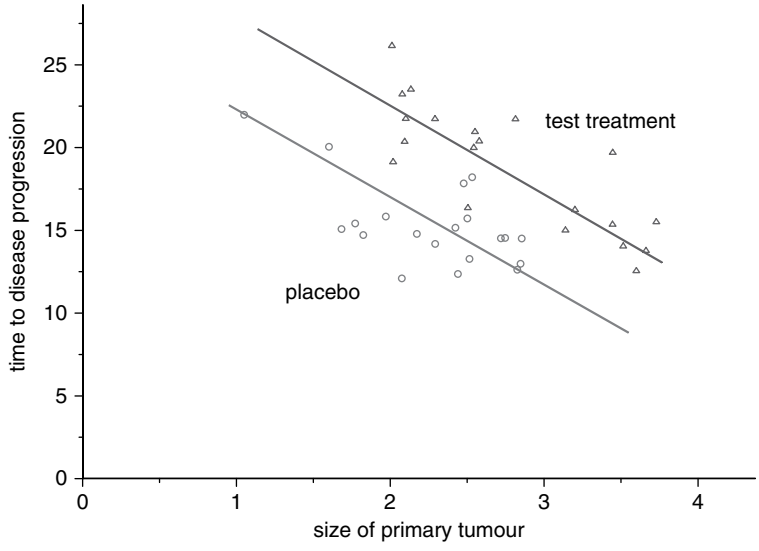


Figure 6.6 Effect of baseline imbalances

This information can be useful for trial planning, for example choosing factors on which to stratify the randomisation, for the future.

- Provides a convenient framework for the evaluation of treatment-by-covariate interactions; in some cases such interactions are anticipated, in other cases such analyses are exploratory.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'The treatment effect itself may also vary with subgroup or covariate – for example, the effect may decrease with age or may be larger in a particular diagnostic category of subjects. In some cases such interactions are anticipated or are of particular prior interest (e.g. geriatrics), and hence a subgroup analysis, or a statistical model including interactions, is part of the planned confirmatory analysis. In most parts, however, subgroup or interaction analyses are exploratory and should be clearly identified as such; they should explore the uniformity of any treatment effects found overall.'

Adjusted analyses presented earlier in this chapter also share some of these advantages and provide improvements in efficiency, can also account for baseline imbalances and allow the evaluation of the homogeneity of the treatment effect. On this final point, however, adjusted analyses are less able to identify the nature of those interactions. With ANCOVA it is possible to say which particular covariates are causing such interactions. A further point to note here and, as mentioned

earlier, is that adjusted analyses become more difficult as the number of covariates increases, although some would argue that including more than a small number of covariates is not needed (see Section 6.7).

Should treatment-by-covariate interactions be found, either through a test of homogeneity in an adjusted analysis or through ANCOVA, then analysis usually proceeds by looking at treatment differences within subgroups. Plots of treatment effects with associated confidence intervals within these subgroups are useful in this regard.

One disadvantage of ANCOVA is that the modelling does involve a number of assumptions and if those assumptions are not valid then the approach could mislead. For example, it is assumed (usually) that the covariates affect outcome in a linear way; there is invariably too little information in the data to be able to assess this assumption in any effective way. In contrast, with an adjusted analysis, assumptions about the way in which covariates affect outcome are not made and in that sense it can be seen as a more robust approach. In some regulatory circles adjusted analyses are preferred to ANCOVA for these reasons.

6.6 Binary, categorical and ordinal data

Analysis of covariance for these data types is based on logistic regression. With continuous data, although we developed the concepts by looking at separate lines for the different treatment groups, we ended up writing down a single equation. This is really the start point with binary, categorical and ordinal regression and logistic regression. We will focus this discussion on the most common setting, the binary case; extensions have been developed for ordinal data.

Again let $z = 0$ for patients in the control group and $z = 1$ for patients in the test treatment group and assume that we have several covariates, say x_1, x_2 and x_3. The main effects model looks at the dependence of $\text{pr}(y = 1)$ on treatment and the covariates:

$$\ell\text{n}\left\{\frac{\text{pr}(y=1)}{[1-\text{pr}(y=1)]}\right\} = a + c\,z + b_1x_1 + b_2x_2 + b_3x_3$$

The coefficient c measures the impact that treatment has on $\text{pr}(y = 1)$. If $c = 0$ then $\text{pr}(y = 1)$ is unaffected by which treatment group the patients are in; there is no treatment effect. Having fitted this model to the data and in particular obtained an estimate of c and its standard error then we can test the hypothesis $\text{H}_0 : \ c = 0$ in the usual way through the signal-to-noise ratio.

The quantity c is very closely related to the odds ratio; in fact c is the log of the OR, adjusted for the covariates. The anti-log of c (given by e^c) gives the adjusted OR. Confidence intervals in relation to this OR can be constructed initially by obtaining a confidence interval for c itself and then taking the anti-log of the lower and upper confidence limits for c.

Example 6.1: Effect of betamethasone on incidence of neonatal respiratory distress

This randomised trial (Stutchfield *et al.* (2005)) investigated the effect of betamethasone on the incidence of neonatal respiratory distress after elective caesarean section. Of the 503 women randomised to the active treatment, 11 babies were subsequently admitted to the special baby unit with respiratory distress compared to 24 babies out of 495 women randomised to the control group.

The odds ratio, for the binary outcome (baby admitted to the special baby unit for respiratory distress) is then 11/492 divided by 24/471, giving a value 0.439. The chi-square test comparing the treatments with regard to the rates of admission gave $p = 0.02$.

A logistic regression analysis was undertaken with:

$z = 0$ if mother randomised to control
$z = 1$ if mother randomised to betamethasone

The analysis was adjusted for gestational age and took account of 'two' covariates; the standard gestational age was 39 weeks and:

$x_1 = 1$ if gest. age, 37 weeks
0 otherwise
$x_2 = 1$ if gest. age, 38 weeks
0 otherwise

The coefficient of the treatment indicator z was -0.840 giving an (adjusted) odds ratio for the treatment effect of 0.432 ($e^{-0.840} = 0.432$). The coefficient b_1 of x_1 was 2.139 and the coefficient b_2 of x_2 was 1.472. So a value of $x_1 = 1$ (gestational age $= 37$ weeks) gives an increase of 2.139 on the log odds scale, or equivalently, an increase of 8.5 ($e^{2.139} = 8.5$) on the odds of being admitted to the special baby unit compared to the standard 39 weeks gestational age. For 38 weeks gestational age the increase to the odds of being admitted is 4.4 ($e^{1.472} = 4.4$) compared to a gestational age of 39 weeks.

We can also investigate the presence of treatment-by-covariate interactions by including cross-product terms:

$$\ell n \left\{ \frac{\text{pr}(y=1)}{[1-\text{pr}(y=1)]} \right\} = a + c\,z + b_1 x_1 + b_2 x_2 + d_1\,z x_1 + d_2\,z x_2$$

Questions relating to those interaction terms are addressed through the *d* coefficient as before. In the above example, looking for treatment-by-covariate interactions would be asking whether the treatment benefit, in terms of a reduction in the likelihood of the baby suffering respiratory distress, was the same for babies delivered at 37, 38 and 39 weeks.

Logistic regression offers similar advantages as ANCOVA for continuous data; correcting for baseline imbalances, allowing the evaluation of the effects of the covariates and providing a convenient framework for the identification of treatment-by-covariate interactions. With regard to efficiency the issues are slightly different. It is important with logistic regression to identify and include those covariates that are predictive of outcome in the modelling; otherwise the treatment effects could be biased. See Ford *et al.* (1995) for further discussion on this point.

6.7 Regulatory aspects of the use of covariates

ICH E9 and the CPMP (2003) 'Points to Consider on Adjustment for Baseline Covariates' make a number of useful points:

- Pre-planning. It is important that the covariates to be included are decided in advance.

  CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'

  Covariates to be included in the analysis must be pre-specified in the protocol or in the statistical analysis plan.'

  If new knowledge becomes available regarding important covariates after completion of the statistical analysis plan then 'modify' the plan at the blind review stage.

- Baseline imbalances

  CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'

  'Baseline imbalance in itself should not be considered an appropriate reason to include a baseline measure as a covariate.'

  In the final section of this chapter we will say more about baseline imbalances and how to deal with them.

- Covariates affected by treatment allocation. Variables measured after randomisation (e.g. compliance, duration of treatment) should not be used as covariates in a model for evaluation of the treatment effect as these may be influenced by the treatment received. A similar issue concerns 'late baselines', that is covariate measures that are based on data captured after randomisation. The term *time-dependent covariate* is sometimes used in relation to each of the examples above.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'It is not advisable to adjust the main analyses for covariates measured after randomisation because they may be affected by the treatments.'

CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'

'When a covariate is affected by the treatment either through direct causation or through association with another factor, the adjustment may hide or exaggerate the treatment effect. It therefore makes the treatment effect difficult to interpret.'

- It is often good practice for continuous outcome variables recording change from baseline to adjust the analysis, or include in the analysis as a covariate, the baseline value of the outcome variable. When this is done, including the outcome variable itself or the change from baseline in that variable makes no difference mathematically to the analysis; the same p-values and estimates of treatment effect will be obtained, so the choice is one of interpretability. If the baseline value is not included in this way then there could be problems with regression towards the mean.

  Regression towards the mean is a phenomenon that frequently occurs with data and in a wide variety of situations. For example, in a chronic condition like asthma, patients will often enter a trial because they have recently suffered a number of exacerbations and are having a particularly bad time with their condition. Almost irrespective of treatment they will probably improve because asthma severity is cyclical and intervention has occurred at a low point. If change from baseline was used as the variable to measure the effectiveness of a treatment then the mean change from baseline in each of the two treatment groups would undoubtedly over-estimate the benefit due to treatment in both groups; part of those improvements would be due to *regression towards the mean*. Further, in a randomised comparison, if one treatment group, by chance, contained patients with poorer baseline values, then comparing the mean change from baseline in one group with the mean change from baseline in the other group

could give a biased conclusion, the bias being caused by regression towards the mean. Of course randomisation should protect against this, but in particular cases imbalances can occur and including the baseline value as a covariate in ANCOVA or adjusting through ANOVA will correct for regression towards the mean.

CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'

'When the analysis is based on a continuous outcome there is commonly the choice of whether to use the raw outcome variable or the change from baseline as the primary endpoint. Whichever of these endpoints is chosen, the baseline value should be included as a covariate in the primary analysis. The use of change from baseline without adjusting for baseline does not generally constitute an appropriate covariate adjustment. Note that when the baseline is included as a covariate in the model, the estimated treatment effects are identical for both 'change from baseline' and the 'raw outcome' analysis.'

- How many covariates? It is usually not appropriate to include lots of covariates in an analysis.

CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'

'No more than a few covariates should be included in the primary analysis.'

Remember however that variables used to stratify the randomisation should be included. It is also not usually appropriate to select covariates within ANCOVA models using stepwise (or indeed any other) techniques. The main purpose of the analysis is to compare the treatment groups not to select covariates.

CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'

'Methods that select covariates by choosing those that are most strongly associated with the primary outcome (often called 'variable selection methods') should be avoided. The clinical and statistical relevance of a covariate should be assessed and justified from a source other than the current dataset.'

Including covariates that are highly correlated adds little to the analysis and should be avoided. Knowledge of the nature of the correlation should help to prevent this happening.

6.8 Connection between ANOVA and ANCOVA

It probably comes as no surprise to learn that there are mathematical connections between ANOVA and ANCOVA for continuous data.

Suppose that we have just two centres (and two treatment groups). If we define binary indicators, say z and x, to denote treatment group and centre, respectively, then including z and x in an ANCOVA is identical mathematically to the corresponding two-way ANOVA. This connection is true more generally. If we were now to add more centres (say four in total) to the ANOVA then defining binary indicators to uniquely define these centres; $x_1 = 1$ for a patient in centre 1, $x_2 = 1$ for a patient in centre 2, $x_3 = 1$ for a patient in centre 3 with 0 values otherwise, then ANCOVA with terms z, x_1, x_2 and x_3 would be mathematically the same as ANOVA. We would obtain the same p-values, (adjusted) estimates of treatment effect, confidence intervals etc.

We can see, therefore, that ANCOVA is a very powerful technique that mathematically incorporates ANOVA.

There are also connections between the Cochran–Mantel–Haenszel procedures and logistic regression for binary and ordinal data, but these issues are beyond the scope of this text.

6.9 Baseline testing

It is generally accepted amongst statisticians that baseline testing, producing p-values for comparisons between the treatment groups at baseline, is of little value. If randomisation has been performed correctly, then 5 per cent of significance test comparisons at baseline will give significant results; any imbalances seen at baseline must be due to chance. The only value to such testing is to evaluate whether the randomisation has been performed correctly, for example, in the detection of fraud. Altman (1991), Section 15.4, provides an extensive discussion on the issue of baseline testing.

CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'

'Statistical testing for baseline imbalance has no role in a trial where the handling of randomisation and blinding has been fully satisfactory.'

It is nonetheless appropriate to produce baseline tables of summary statistics for each of the treatment groups. These should be looked at from a clinical perspective and imbalances in variables that are potentially prognostic noted. Good practice hopefully will have ensured that the randomisation has been stratified for important baseline prognostic factors and/or the important prognostic factors

have been included in some kind of adjusted analysis, for example ANCOVA. If this is not the case then sensitivity analyses should be undertaken again, through ANCOVA, to make sure that those imbalances are not the cause of, say, an observed positive (or negative) treatment difference.

CPMP (2003): 'Points to Consider on Adjustment for Baseline Covariates'

'When there is some imbalance between the treatment groups in a baseline covariate that is solely due to chance then adjusted treatment effects may account for this observed imbalance when unadjusted analyses may not. If the imbalance is such that the experimental group has a better prognosis than the control group, then adjusting for the imbalance is particularly important. Sensitivity analyses should be provided to demonstrate that any observed positive treatment effect is not solely explained by imbalances at baseline in any of the covariates.'

7

Intention-to-treat and analysis sets

7.1 The principle of intention-to-treat

When we analyse data there are inevitably questions that arise regarding how we deal with patients who withdraw, with patients who have violated the protocol, with patients who have taken a banned concomitant medication and so on. The principle of intention-to-treat helps guide our actions. We will firstly explain the principle through two examples and then discuss various aspects of its interpretation before addressing the application of the principle in practical situations.

Example 7.1: Surgery compared to radiotherapy in operable lung cancer

Consider a (hypothetical) trial comparing surgery and radiotherapy in the treatment of operable lung cancer. Assume that a total of 200 patients were randomised to one of these two groups as shown in Figure 7.1.

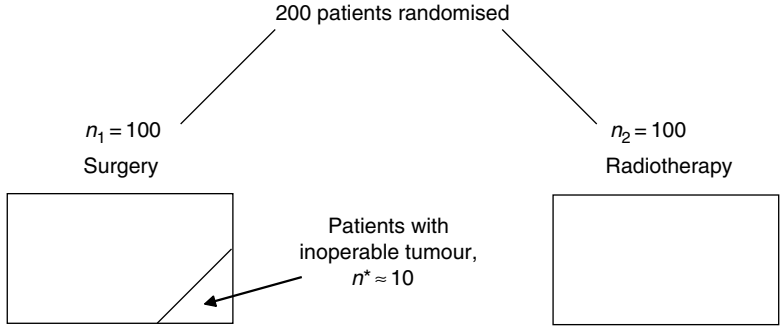


Figure 7.1 Randomised trial comparing surgery and radiotherapy

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

Unfortunately some of the patients randomised to the surgery group were found on the operating table not to have operable tumours. The intended operative procedure could not be undertaken for these patients and they were simply closed up and in fact given the radio-therapy regimen defined for the radio-therapy group. Assume that there were ten such patients. The primary endpoint was survival time and all 200 patients were followed up until death. At the analysis stage we have to decide how to deal with the ten patients who were assigned to surgery, but who did not receive the intended surgery. Several possibilities exist for the analysis and we will consider the following three:

Option 1: compare the mean survival time of the 90 who received the intended surgery with the mean survival time of the 100 who were assigned to radio-therapy.

Remember that all 200 patients provide data on the primary endpoint so this analysis ignores the data on ten of those patients; those who were assigned to surgery, but did not receive the intended operative procedure. Perhaps we can think of ways of including those data. Options 2 and 3 include these patients.

Option 2: compare the mean survival time of the 90 patients who received the intended surgery with the mean survival time for the $100 + 10$ patients who ended up getting radio-therapy.

The argument here is the ten patients who 'switched' from surgery to radio-therapy are most likely to behave like the 100 patients initially assigned to the radio-therapy group.

Option 3: compare the groups according to the randomisation scheme; that is the mean survival time of the 100 patients initially assigned to surgery with the mean survival time of the 100 patients initially assigned to radio-therapy.

These three options differ merely in the way the ten patients who were assigned to surgery, but did not receive the intended surgery are handled. Option 1 ignores them, option 2 puts them in the radio-therapy 'group', while option 3 leaves them in the surgery 'group'.

Two things to note immediately. Firstly, the ten patients in the surgery group who do not have operable tumours are likely to have very poor prognoses. They are the patients who have advanced tumours, maybe multiple tumours, which cannot easily be excised. Secondly, randomisation will have given us balanced groups so equally there will be approximately ten patients in the radiotherapy group who also do not have operable tumours and similarly very poor prognosis, but these patients remain unseen and unidentified.

Now consider the implications arising out of each of these options.

Option 1: this option removes ten very poor prognosis patients from the surgery group, but leaves a similar group of very poor prognosis patients in the radiotherapy group. This clearly introduces bias as the comparison is not comparing like with like. The radiotherapy group will, because of the omission of ten very specific patients in the surgery group, be on average, a better prognosis group than the surgery group to which they are being compared.

Option 2: this option is twice as bad! The ten very poor prognosis patients have been transferred to the radiotherapy group to give a total of 20 very poor prognosis patients in that group. The bias in the resulting analysis is likely to be even bigger than the bias in option 1.

Option 3: this is the only valid option. It is the only option that compares groups which are alike in terms of the mix of patients. This is the intention-to-treat option which compares the groups as randomised.

A number of issues arise out of these considerations. While option 1 makes sense from a statistical perspective, does it make sense from a clinical standpoint? Remember the purpose of this trial is to compare surgery and radiotherapy in operable lung cancer yet we are including ten patients in the surgery group who did not receive the planned surgery. Well, in fact, the comparison in option 3 is not comparing surgery with radiotherapy, it is comparing two treatment strategies. Strategy 1 is to give the patient surgery, however, if it is found that the tumour is not operable then close the patient up and give them radiotherapy, while strategy 2 is to give radiotherapy.

Although you may agree that this provides a comparison of these two treatment strategies you may say that this is not of interest to you clinically and you are looking for a comparison of pure surgery with pure radiotherapy. On that basis you may therefore prefer to go for option 1 which seems to provide exactly what you want as it gives a comparison of pure surgery with pure radiotherapy! This view, however, would be both naive and incorrect. Option 1 does not provide a *valid* comparison of pure surgery and pure radiotherapy because the groups are not alike; it is subject to bias because the radiotherapy group in this comparison is on average a better prognosis group that the surgery 'group'.

A question which sometimes arises here is; can't we just remove ten patients from the radiotherapy group and then compare the 90 who received surgery with the resulting radiotherapy group? No, although this would equalise the numbers in the two groups being compared, it would not necessarily equalise the mix of patients in the two groups. Remember the ten patients who did not get surgery are not just any ten patients; they are a selected group of very poor prognosis patients. Well, as an alternative could we not remove the ten patients in the radiotherapy group who turn out to have the worst survival experience? The answer again is

no, this would not necessarily make the resulting comparison valid as it is based on a very strong assumption, which could never be verified, that the ten patients who did not get the intended surgery are indeed the ten worst patients. Unfortunately the comparison of pure surgery and pure radiotherapy is not possible in this trial. The only question that can be answered relates to the evaluation of the two treatment strategies as mentioned above.

Example 7.2: Clofibrate in the reduction of mortality after myocardial infarction

This is a well-known placebo-controlled trial which evaluated clofibrate in terms of reducing mortality in patients suffering a myocardial infarct was reported by Coronary Drug Research Group (1980).

The groups were compared overall and the five year death rate amongst the 1103 patients randomised to clofibrate was 20.0 per cent compared to a five year death rate amongst the 2789 placebo patients of 20.9 per cent. These differences were not statistically significant with $p = 0.55$.

The trialists then investigated the impact of compliance on these results. Patients were defined as good compliers if they took at least 80 per cent of the prescribed dose during the treatment period. Poor compliers were patients who took less than 80 per cent. In the clofibrate group the good compliers were seen to have only a 15.0 per cent five year death rate while the poor compliers had a 24.6 per cent five year death rate, a clear difference both clinically and statistically with $p = 0.001$. So, in fact, the active medication does work, it is simply a matter of taking the medication. Patients who take the medication do well, those who fail to take the medication do not do well.

This same comparison however was then undertaken amongst the placebo patients. The good compliers on placebo only had a 15.1 per cent five year death rate while the poor compliers had a 28.3 per cent five year death rate with $p = 00000000000000047$! These placebo tablets are remarkable!

A moment's thought will suffice to realise that the conclusions we are drawing from these analyses are nonsense. The bottom line is that the active medication is having no effect and the initial overall comparison is telling us that. However, compliance is linked to other things that are potentially having an effect, such as giving up smoking, starting to take regular exercise, modifying one's diet to reduce the amount of fat consumed and so on. The patients who do these things are the ones who do everything that their doctors tell them to do, including taking the medication! So, the taking of the medication is correlated with some other things that are beneficial and causing an apparent treatment effect related to compliance.

These two examples should give a clear indication of the dangers of compromising the randomisation at the analysis stage. Even small departures in terms of excluding patients from the analysis could have a major impact on the validity of the conclusions.

The *principle of intention-to-treat (ITT)* tells us to compare the patients according to the treatments to which they were randomised. Randomisation gives us comparable groups, removing patients at the analysis stage destroys the randomisation and introduces bias. Randomisation also underpins the validity of the statistical comparisons. If we depart from the randomisation scheme then the statistical properties of our tests are compromised.

The FDA guideline on anti-microbial drugs captures the issues well.

FDA (1998): 'Developing Anti-microbial Drugs – General Considerations for Clinical Trials'

'The intent-to-treat principle suggests that eligible, randomized patients should be evaluated with respect to outcome based on original treatment assignment regardless of modifications to treatment occurring after randomisation. The statistical analysis seeks to establish if the particular assignment received is predictive of outcome, and the study can be interpreted as a strategy trial where the initial assignment is only the beginning of the treatment strategy. However, many researchers seek to glean results from the clinical trial that would have been observed if all patients had been able to remain on their initial assignment. This leads to analysis of subsets that exclude patients with imperfect compliance or follow-up data. However, the validity of these analyses rests on the assumption that the two treatment groups, after excluding such patients, differ only by the treatment received. This assumption could be violated in many subtle ways. For example, differential toxicity related to severity of illness could lead to selection bias. Similarly, the subjects unable to comply with medication may be those most at risk of a negative outcome and their exclusion may bias the treatment comparison.'

7.2 The practice of intention-to-treat

7.2.1 Full analysis set

The previous section clearly indicates the need to conform to the principle of intention-to-treat to ensure that the statistical comparison of the treatment groups remains valid. In practice compliance with this principle is a little more difficult and the regulators, recognising these difficulties, allow a compromise. This involves the definition in particular trials of the *full analysis set* which gets us as close as we possibly can get to the intention-to-treat ideal.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'The intention-to-treat principle implies that the primary analysis should include all randomised subjects. Compliance with this principle would necessitate complete follow-up of all randomised subjects for study outcomes. In practice this ideal may be difficult to achieve, for reasons to be described. In this document the term 'full analysis set' is used to describe the analysis set which is as complete as possible and as close as possible to the intention-to-treat ideal of including all randomised subjects.'

The regulators are telling us therefore, to get as close as possible and they go on in the ICH E9 guideline to outline circumstances where it will usually be acceptable to omit subjects without causing bias.

These potential exclusions are:

- Subjects who violate the inclusion/exclusion criteria
- Subjects who fail to take at least one dose of study medication
- Subjects who do not provide any post-baseline data

These omissions will not cause bias only under some circumstances. In particular, subjects in each of the treatment groups should receive equal scrutiny for protocol violations and all such violators should be excluded, in relation to the first point. For the second and third points, the fact that patients do not take study medication or do not provide any post-baseline data should be unrelated to the treatments to which such subjects were assigned. Any potential bias arising from these exclusions should be fully investigated.

The term full analysis set was introduced in order to separate the practice of intention-to-treat from the principle, but practitioners still frequently use the term *intention-to-treat population* when referring to this set. The term *modified intention-to-treat population* is also in common use within particular companies and also by regulators in some settings where exclusions from strict intention-to-treat are considered.

CPMP (1997): 'Note for Guidance on Evaluation of New Anti-Bacterial Medicinal Products'

'The modified ITT, where unqualified patients are excluded, and patients with clinically and or microbiologically documented infections (as stated in the protocol), who have received at least one dose of the investigated drug thus addressed, is particularly valid for regulatory purposes.'

This quote also makes a further important point relating to anti-infective trials. It is not uncommon in such trials to be including patients on clinical grounds

that do not end up having the target organism. It is only when the results of the laboratory analysis are known that it becomes entirely clear that the patient should not have been included. These non-eligible patients are invariably excluded from the analysis of efficacy and this does not cause bias providing all patients, irrespective of treatment group, are receiving equal scrutiny for this violation. This permits the analysis to be based on a clear definition of the target population.

In superiority trials the full analysis set is invariably the basis for the primary analysis. The regulatory preference for this stems in part because the full analysis set also tends to give a conservative view of the treatment difference, as a result of including in the analysis subjects who have not conformed entirely with the protocol. The regulators can be assured that if the analysis based on this set gives a statistically significant result, then the treatment in question is effective. This preference, however, only applies when considering superiority trials. In equivalence and non-inferiority this analysis set tends to be anti-conservative. This issue will be discussed later, in the chapter on equivalence and non-inferiority testing.

7.2.2 Per-protocol set

The *per-protocol set* is described as follows by the regulators.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'The 'per-protocol' set of subjects, sometimes described as the 'valid cases', the 'efficacy' sample or the 'evaluable subjects' sample, defines a subset of the subjects in the full analysis set who are more compliant with the protocol . . . '

The definition of a per-protocol set of subjects allows us to get closer to the scientific question by including only those patients who comply with the protocol to a defined extent. The per-protocol set, like the full analysis set, must be pre-specified in the protocol and then defined at the patient level at the blind review, following database lock, but before breaking the blind. It must be noted, however, that the per-protocol set is subject to bias and further, tends to overestimate the treatment effect. For this reason it is usually used only as a secondary analysis, supportive hopefully of the findings based on the full analysis set.

7.2.3 Sensitivity

It is good statistical practice to evaluate the sensitivity of the conclusions to different choices of the analysis sets.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'In general, it is advantageous to demonstrate a lack of sensitivity of the principle trials results to alternative choices of the set of subjects analysed. In confirmatory trials it is usually appropriate to plan to conduct both an analysis of the full analysis set and a per-protocol analysis, so that any differences between them can be the subject of explicit discussion and interpretation'.

This regulatory statement is not saying that the analyses based on the full analysis set and the per-protocol set are in any sense co-primary. The full analysis set will provide the primary analysis and usually this analysis must give $p \leq 0.05$ for a positive result. The per-protocol set, however, does not need to give $p \leq 0.05$, but should provide results which are qualitatively similar in terms of the direction of the treatment effect and with effect size not too dissimilar from that seen for the full analysis set.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'When the full analysis set and the per-protocol set lead to essentially the same conclusions, confidence in the trial results is increased . . . '

For further discussion on the definition of analysis sets and additional practical advice see Gillings and Koch (1991).

7.3 Missing data

7.3.1 Introduction

The discussion in the previous section regarding the practical application of the principle of intention-to-treat does not, however, give the full picture. While this principle plus consideration of the per-protocol set may clearly define the sets of subjects to be analysed, we still have to decide how to deal with the missing data caused by failure to complete the study entirely in line with the protocol.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Missing values represent a potential source of bias in a clinical trial. Hence, every effort should be undertaken to fulfil all the requirements of the protocol concerning the collection

and management of data. In reality, however, there will almost always be some missing data. A trial may be regarded as valid, nonetheless, provided the methods of dealing with missing values are sensible, and particularly if those methods are pre-defined in the protocol.'

It is worth remembering that there is no single, perfect way of handling missing data. The only truly effective way of dealing with missing data is not to have any in the first place! In practice we always need to consider the sensitivity of our trial results to the methods employed to handle the missing data. This is particularly true if the amount of such data is large.

There are a number of alternative approaches to dealing with missing data in common practice. Amongst these are the following:

- Complete cases analysis
- Last observation carried forward (LOCF)
- Success/failure classification
- Worst case/best case imputation

7.3.2 Complete cases analysis

One very simplistic way of handling missing data is to remove those patients with missing data from the analysis in a *complete cases analysis* or *completers analysis*. By definition this will be a per-protocol analysis which will omit all patients who do not provide a measure on the primary endpoint and will of course be subject to bias. Such an analysis may well be acceptable in an exploratory setting where we may be looking to get some idea of the treatment effect if every subject were to follow the protocol perfectly, but it would not be acceptable in a confirmatory setting as a primary analysis.

7.3.3 Last observation carried forward (LOCF)

This analysis takes the final observation for each patient and uses it as that patient's endpoint in the analysis. For example, in a 12 month trial in acute schizophrenia, a patient who withdraws at month 7 due to side effects will have their month 7 value included in the analysis of the data.

In one sense this approach has clinical appeal. The final value provided by the patient who withdrew at month 7 is a valid measure of how successful we have been in treating this patient with the assigned treatment and so should be part of the overall evaluation of that treatment. In some circumstances, however,

this argument breaks down. For example, if there is an underlying worsening trend in disease severity then patients who withdraw early will tend to provide better outcomes than those who withdraw later on in the treatment period. If one treatment has more early dropouts than the other, possibly because of side effects say, then there will be bias caused by the use of LOCF. Multiple sclerosis and Alzheimer's disease are settings where this could apply. The opposite will of course be true in cases where the underlying trend is one of improvement; depression would be one such therapeutic area. These scenarios emphasise the earlier point that there is no universally valid way to deal with missing data.

7.3.4 Success/failure classification

One particularly simple way of dealing with missing data is to use a binary outcome as the endpoint with dropouts being classified as treatment success or failure or depending on the outcome at the time of dropping out. For example, in a hypertension trial success may be defined as diastolic blood pressure below 90 mmHg at the end of the treatment period or at the time of treatment withdrawal with all other outcomes being classified as failures. Reducing the outcome to a dichotony in this way will, however, lead to a loss of power and this loss of power needs to be considered outcomes when calculating sample size at the planning stage.

A success/failure approach will be particularly effective if the endpoint is already binary, for example cured/not cured in a trial of an anti-infective. In the earlier discussion of modified ITT the CPMP (1997) 'Note for Guidance on the Evaluation of New Anti-bacterial Medicinal Products goes on to say: *'Patients who have no measurements after baseline are included as failures in the analysis.'*

7.3.5 Worst case/best case imputation

This method gives those subjects who withdraw for positive reasons the best possible outcome value for the endpoint and those who withdraw for negative reasons, the worst value. This may seem a little extreme and a lesser position would be to look at the distribution of the endpoint for the completers and to use say the upper quartile (the value that cuts off the best 25 per cent of values) for those subjects withdrawing for positive reasons and the lower quartile (the value that cuts off the worst 25 per cent of values) for those subjects who withdraw for negative reasons.

Alternative applications of these rules would be to use the best case value (or the upper quartile) for subjects in the control group and the worst case value (or the lower quartile) for subjects in the test treatment group. Certainly if the treatment comparison is preserved under such a harsh scheme then we can be very confident of a true benefit for the test treatment!

7.3.6 Sensitivity

In all cases and particularly where the extent of missing data is substantial, several analyses will usually be undertaken to assess the sensitivity of the conclusions to the method used to handle missing data. If the conclusions are fairly consistent across these different analyses then we are in a good position. If, however, our conclusions are seen to change, or to depend heavily on the method used for dealing with missing data, then the validity of those conclusions will be drawn into question.

Our discussion so far regarding sensitivity has focused on using several different approaches to both the definition of the analysis set and the handling of missing data. One of the main goals of a trial is to estimate the magnitude of the treatment benefit and these sensitivity evaluations will give a series of estimates. For estimation purposes, and in particular here we are thinking in terms of confidence intervals, we should choose a method of imputation that makes sense clinically rather than go for extremes which, while providing a conservative view of the treatment effect, do not give a sensible, realistic estimate of the clinical benefit.

7.3.7 Avoidance of missing data

The CPMP guideline on missing data (CPMP (2001) 'Points to Consider on Missing Data') includes several key points on the avoidance of missing data. As we can see from the above discussion, missing data causes problems and we should avoid it wherever possible.

CPMP (2001): 'Points to Consider on Missing Data'

'Several major difficulties arise as a result of the presence of missing values and these are aggravated as the number of missing values increases. Thus, it is extremely important to avoid the presence of unobserved measurements as much as possible, by favouring designs that minimise this problem, as well as strengthening data collection regardless of the patient's adhesion to the protocol and encouraging the retrieval of data after the patient's drop-out.'

Allowing dose reductions or drug 'holidays' will possibly keep patients in a trial and avoid them dropping out, in addition to providing a closer model to what will actually happen in practice in real life. Continued follow-up for patients who withdraw from medication should also be considered (sometimes referred to as 'partial' withdrawal) and this will give much more flexibility when it comes to analysing the data once the trial is complete.

Example 7.3: Bosentan therapy in pulmonary arterial hypertension

This was a placebo-controlled trial (Rubin *et al.* (2002)) using change from baseline to week 16 in exercise capacity (distance in metres walked in six minutes) as the primary endpoint and change from baseline to week 16 in the Borg dyspnea index and in the WHO functional class. Missing data at week 16 were handled as follows:

'For patients who discontinued the study medication because of clinical worsening, the values recorded at the time of discontinuation were used; patients for whom no value was recorded (including patients who died) were assigned the worst possible value (0 m). For all other patients without a week 16 assessment, the last six-minute walking distance, score on the Borg dyspnea index, and WHO functional class were used as week 16 values.'

So for patients without week 16 values, LOCF was used, unless the patient withdrew because of clinical worsening, in which case either the data at the time of withdrawal was used or a worst case imputed value (zero metres for the walk test, Borg index = 0 or WHO = IV) if there were no data at withdrawal. This set of rules, which were pre-defined, are clinically appropriate and should minimise bias.

7.4 Intention-to-treat and time-to-event data

In order to illustrate the kinds of arguments and considerations which are needed in relation to intention-to-treat, the discussion in this section will consider a set of applications where problems frequently arise. In Chapter 13 we will cover methods for the analysis of time-to-event or so-called survival data, but for the moment I would like to focus on endpoints within these areas that do not use the time-point at which randomisation occurs as the start point for the time-to-event measure. Examples include the time from rash healing to complete cessation of pain in Herpes Zoster, the time from six weeks after start of treatment to first seizure in epilepsy and time from eight weeks to relapse amongst responders at week 8 in severe depression.

In each of these situations there are clinical arguments that support the use of the particular endpoint concerned. From a statistical point of view, however, each of these endpoints gives an analysis that is subject to bias in a clear violation of the principle of intention-to-treat. We will look at each of the settings in turn.

In the case of a randomised trial in Herpes Zoster, patients have the potential to cease pain prior to rash healing and these patients would not enter the analysis of time to cessation of pain from rash healing. Invariably the likelihood that pain will cease early in this way will depend upon the treatment received. As a consequence the sets of patients in each of the two treatment groups entering the analysis of time to cessation of pain from rash healing will not necessarily be alike. This selection phenomenon will result in a clear violation of intention-to-treat and the resultant analysis will be biased. See Kay (1995) for further discussion on this point. There have been some attempts (see for example Arani *et al.* (2001)) to justify such an analysis using complex statistical modelling but this approach has been shown by Kay (2006) to be flawed and the problem of violation of intention-to-treat remains. In Herpes Zoster, the pain that remains following rash healing is known as post-hepatic neuralgia (PHN) and there is strong interest in evaluating the relative effects of treatments on PHN. Looking at time from rash healing to cessation of pain is an attempt to focus on this. Unfortunately it is not possible, even in a randomised trial, to analyse this endpoint in an unbiased way. The only way to identify the relative effect of the two treatments with regard to PHN is to compare the proportion of all patients in the two treatment group still with pain at specified points through time.

In newly diagnosed epilepsy, it is standard practice to use the time from six weeks (or sometimes three months) following the start of treatment to the first seizure, as the primary or certainly important secondary endpoint. Again, however, we hit problems with selection effects and intention-to-treat. Brodie *et al.* (1995) evaluate lamotrigine compared to carbamazepine in a randomised trial in patients with newly diagnosed epilepsy. In a secondary analysis of time (from randomisation) to withdrawal it is clear that by six weeks, approximately 18 per cent of the patients have withdrawn in the lamotrigine group, while 27 per cent of patients have withdrawn in the carbamazepine group. The analysis of time from six weeks to first seizure excludes these patients. This is a very long way from being a randomised comparison and is potentially subject to substantial bias. Even if the withdrawal rates had been the same, the potential for bias would remain. It is not so much that the numbers of patients withdrawing are different, it is that the comparability in terms of the mix of patients has been compromised, the differential withdrawal rates just makes things worse. From a clinical point of view excluding the first six weeks of treatment does make sense as it is recognised that it takes some time to stabilise the dose, but again, unfortunately, the endpoint that apparently captures this, time from six weeks to first seizure, cannot be evaluated in an unbiased way. An alternative and appropriate approach to look at the effectiveness of treatment following dose stabilisation has been suggested by Brodie and Whitehead (2006). These authors (using three months as the stabilisation period rather than six weeks) consider the following endpoint; time from randomisation to withdrawal, whenever this occurs, or to first seizure from three months onwards.

This endpoint combines both tolerability (withdrawal) and efficacy (first seizure), but does not penalise a treatment in terms of seizures during the stabilisation phase (the first three months) providing tolerability is not a problem over that period. From a pragmatic perspective this alternative endpoint makes a lot of sense, from the patients' point of view the important issue is longer term stabilisation, free of seizures and this endpoint captures precisely that.

Finally, in severe depression, many trials are designed to investigate treatment relapse in patients who have responded following treatment. Response could be defined, for example, by a reduction in the score on the 17 point Hamilton Depression Scale (HAMD-17) to below 15 with relapse defined as an increase to 16 or above. Typically response is assessed following eight weeks of treatment and the endpoint of interest in evaluating relapse is time from eight weeks to relapse. Patients who have not responded by week 8 are usually withdrawn for lack of efficacy. These extension studies in responders are not randomised comparisons and the analysis is based solely on those patients who are responders at eight weeks. Storosum *et al.* (2001) recognise the potential for bias in this analysis. In common with the previous two settings there is a violation of the principle of intention-to-treat. An alternative analysis, which looks at time to treatment failure with treatment failure defined as withdrawn from treatment for lack of efficacy up to and including week 8, or HAMD-17 ≥ 16 beyond week 8 would maintain all patients in the treatment comparison. This comparison takes account of possible differential effects of treatment up to and including week 8 in terms of achieving a response and beyond week 8 is looking at the proportion of patients whose response is maintained.

As a general rule, time-to-event endpoints that do not use the point of randomisation as the start point should be avoided as there is always the potential for patient selection to take place between the point of randomisation and when the clock starts ticking for the proposed endpoint.

7.5 General questions and considerations

One question that is frequently asked is; what do you do with patients who get given the wrong treatment by mistake? It must be said that this does not happen very frequently, but when it does it is necessary to dig a little and try to find out why this has happened. If it is an isolated case and is clearly an administrative error then it seems most reasonable to include that patient in the group according to treatment received. If, however, it is not an isolated case, maybe there are several such mistakes in the same centre, then this draws into question the validity of what is happening at that centre and one starts to think in terms of fraud, has the investigator correctly followed the randomisation scheme? In such situations

there may be consideration, ultimately, of removing all of the data from that centre from the analysis.

The considerations so far in this chapter have been on the evaluation of efficacy. For safety we usually define the s*afety set* as the set of subjects who receive at least one dose of study medication. Usually the safety set will coincide with the full analysis set, but not always. There may well be a patient who started on medication, but withdrew immediately because of a side effect. This patient is unlikely to have provided post baseline efficacy data and so could be excluded from the full analysis set.

In cross-over trials, considerations of analysis sets and missing data is somewhat different. In these trials each subject provides a response on each of the treatments. The analysis of such data focuses on the treatment difference within each subject. When a subject drops out during the second period and therefore fails to give a response for the treatment given in period 2 then it is not possible to calculate a treatment difference and so this patient would not be included in the analysis. So, in cross-over trials we are usually forced to exclude the dropouts. Does this compromise the validity of the treatment comparison? In terms of bias the answer is, not usually, since exclusions of this kind will deplete each of the treatment 'groups' equally as the same subject is being omitted from both 'groups', although the potential for bias should always be considered. In terms, however, of extrapolating the conclusions from the trial to the general population there could be problems. If particular kinds of patients are being omitted from the analysis essentially because they are prone to side effects, possibly from one of the two treatments being compared, then the trial population may not be representative of the population defined by the inclusion/exclusion criteria. However, in phase I studies with healthy volunteers, these aspects are unlikely to be an issue and it is common practice to 'replace' dropouts with other subjects in order to achieve the required sample size.

A key aspect of the definition of analysis sets and the way that missing data is to be handled is pre-specification. Usually these points will be covered in the protocol, if not, in the statistical analysis plan. If methods are not pre-specified then there will be problems as the way that these issues are dealt with could then be data driven, or at least there may be suspicion of that. This is, of course, not unique to analysis sets and missing data, but is true more generally in relation to the main methods of statistical analysis.

To conclude this discussion it is worth covering just a few misconceptions:

- Does having equal numbers of subjects in the treatment groups at the statistical analysis stage protect against bias?
  This corresponds to similar dropout rates across the treatment groups. The answer to the question is no! It is the mix of patients that is the basis of a valid comparison, not the numbers of patients. It is almost inevitable that if

two treatments are truly different then different kinds of subjects will drop out from the two groups. For example, in a placebo-controlled trial, those withdrawing from the active treatment group could well be withdrawing for side effects while the dropouts in the placebo group could be withdrawing because of lack of effect.

- Does basing the sample size calculation on the per-protocol set and then increasing the sample size to allow for dropouts ensure that the per-protocol set will not be subject to bias?
  No! It often makes sense to power for the per-protocol set and then factor upwards to allow for dropouts as this will also ensure that there is enough power for the full analysis set providing any extra patient-to-patient variation in the full analysis set does not counterbalance the increase in sample size, but the analysis based on the per-protocol set is still subject to bias. See Section 8.5.2 for further discussion on this point.

- Does pre-specifying in the protocol that the analysis based on the per-protocol set will be the primary analysis protect against bias?
  As mentioned elsewhere, it is good scientific practice to pre-specify the main methods of statistical analysis in the protocol, but just because something is specified in the protocol it does not mean that it is correct. So again the answer is no.

8

Power and sample size

8.1 Type I and type II errors

The statistical test procedures that we use unfortunately are not perfect and from time to time we will be fooled by the data and draw incorrect conclusions. For example, we know that 17 heads and 3 tails can (and will) occur with 20 flips of a fair coin (the probability from Chapter 3 is 0.0011); however, that outcome would give a significant *p*-value and we would conclude incorrectly that the coin was not fair. Conversely we could construct a coin that was biased 60 per cent/40 per cent in favour of heads and in 20 flips see say 13 heads and 7 tails. That outcome would lead to a non-significant *p*-value ($p = 0.224$) and we would fail to pick up the bias. These two potential mistakes are termed type I and type II errors.

To explain in a little more detail, consider a parallel group trial in which we are comparing two treatment means using the unpaired t-test. The null hypothesis H_0: $\mu_1 = \mu_2$ that the treatment means are equal is either true or not true; God knows, we don't! We mere mortals have to make do with data and on the basis of data we will see either a significant *p*-value ($p \leq 0.05$) or a non-significant *p*-value ($p = \text{NS}^*$). The various possibilities are contained in Table 8.1.

Suppose the truth is that $\mu_1 = \mu_2$, the treatments are the same. We would hope that the data would give a non-significant *p*-value and our conclusion would be correct, we are unable to conclude that differences exist. Unfortunately that does not always occur and on some occasions we will be hoodwinked by the data and get $p \leq 0.05$. On that basis we will declare statistical significance and draw the conclusion that the treatments are different. This mistake is called the *type I error*. It is the *false positive* and is sometimes referred to as the *α error*.

* $p = \text{NS}$ is shorthand to say that *p* is not statistically significant at the 5 per cent level. Its use in reporting trial results is not recommended; exact *p*-values should be used

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

Table 8.1 Type I and type II errors

	H_0 true, $\mu_1 = \mu_2$	H_0 not true, $\mu_1 \neq \mu_2$
Data gives p = NS (cannot conclude $\mu_1 \neq \mu_2$)	√	x
Data gives $p \leq 0.05$ (conclude $\mu_1 \neq \mu_2$)	x	√

Conversely suppose that in reality $\mu_1 \neq \mu_2$, the treatments are different. In this case we would hope that $p \leq 0.05$, in which case our conclusion will be the correct one; treatment differences. Again this will not always happen and there will be occasions when, under these circumstances, we get $p =$ NS. On this basis we will conclude 'no differences'. This second potential mistake is called the *type II error*. This is the *false negative* or the β *error*, the treatments really are different, but we have missed it!

There is a well-known theorem in statistics, called the Neyman–Pearson Lemma, which shows that, for a given sample size, it is simply not possible to eliminate these two mistakes; we must always trade them off against each another.

Usually the type I error is fixed at 0.05 (5 per cent). This is because we use 5 per cent as the significance level; the cut-off between significance ($p \leq 0.05$) and non-significance ($p > 0.05$). The null distribution tells us precisely what will happen when the null hypothesis is true; we *will* get extreme values in the tails of that distribution, even when $\mu_1 = \mu_2$. However, when we do see a value in the extreme outer 5 per cent, we declare significant differences and by definition this will occur 5 per cent of the time when H_0 is true.

The type II error is a little more difficult to pin down. It is related to another quantity called power. If type II error is 10 per cent then power is 90 per cent; *power* is 100 minus type II error. Type II error is missing a real difference, power is capturing a real difference; if there is a 10 per cent chance of missing the bus, there is a 90 per cent chance of catching the bus and they are opposites in this sense! We control type II error by controlling power; for example we may design our trial to have 80 per cent power, in which case the type II error is controlled at 20 per cent.

8.2 Power

As seen in the previous section, power measures our ability to detect treatment differences. A convenient mathematical way of thinking about power is:

$$\text{power} = \text{probability}(p \leq 0.05)$$

When we say that a trial has 80 per cent power to detect a certain level of effect, for example 4 mmHg, what we mean is that if we conduct the trial and the true difference really is 4 mmHg then there is an 80 per cent chance of coming out of the trial with a significant *p*-value, and declaring differences.

We can in fact calculate power in advance of running the trial by speculating about what may happen. Assume in a parallel group cholesterol lowering study comparing a test treatment with placebo that there are 50 patients per group. The unpaired t-test will be used to compare the mean reduction in cholesterol level between the groups at the conventional two-sided significance level of 0.05. Assume also that the standard deviation for the reduction in cholesterol is 1.1 mmol/l. For various values for the treatment difference, the calculated power is given in Table 8.2.

So, for example, if the true difference between the treatments means was 0.50 mmol/l then this trial would have a 62.3 per cent chance of coming out with

Table 8.2 Power for various treatment differences, n = 50 per group

Treatment difference, $\mu_1 - \mu_2$	Power
0.25	0.206
0.50	0.623
0.75	0.926
1.00	0.995

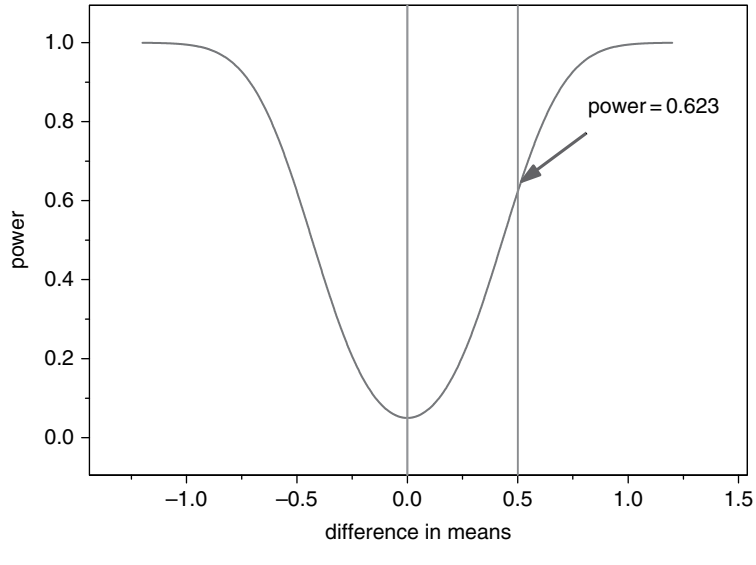


Figure 8.1 Power curve for $n = 50$ per group

a significant p-value ($p \leq 0.05$). Similarly if the true difference were 0.75 mmol/l then the chance of getting a significant result would be 92.6 per cent.

Figure 8.1 plots the values for power against the true difference in the treatment means. Certain patterns emerge. Power increases with the magnitude of the treatment difference; large differences give high values for power and the value for power approaches one as the treatment difference increases in either a negative or a positive direction. The implication here is that large differences are easy to detect, small differences are more difficult to detect. The power curve is symmetric about zero and this is because our test is a two-tailed test; a difference of +1 mmol/l has just the same power as a difference of −1 mmol/l.

Suppose now that the trial in the example were a trial in which a difference of 0.5 mmol/l was viewed as an important difference. Maybe this reflects the clinical relevance of such a difference or perhaps from a commercial standpoint it would be a worthwhile difference to have. Under such circumstances only having 62.3 per cent power to detect such a difference would be unacceptable; this corresponds to a 37.7 per cent type II error, an almost 40 per cent chance of failing to declare significant differences. Well, there is only one thing you can do, and that is to increase the sample size. The recalculated values for power are given in Table 8.3 with a doubling of the sample size to 100 patients per group.

The power now to detect a difference of 0.5 mmol/l is 89.5 per cent, clearly a substantial improvement on 62.3 per cent. Figure 8.2 shows the power curve for a sample size of 100 patients per group and it can be seen that the values for power have increased across all the potential values for the true treatment difference. These arguments form the basis of the sample size calculation; we think in terms of what level of effect it is important to detect, either from a clinical, regulatory or commercial perspective, and choose our sample size to give high power for detecting such effects. In our example if we had said that we require 80 per cent power to detect a difference of 0.50 mmol/l then a sample size of 76 per group would have given us exactly that. For 90 per cent power we would need 102 patients per group.

Before moving on to discuss sample size calculations in more detail, it is worth noticing that the power curve does not come down to zero at a difference of 0.0,

Table 8.3 Power for n = 100 per group

Treatment difference, $\mu_1 - \mu_2$	Power
0.25	0.362
0.50	0.895
0.75	0.998
1.00	1.000

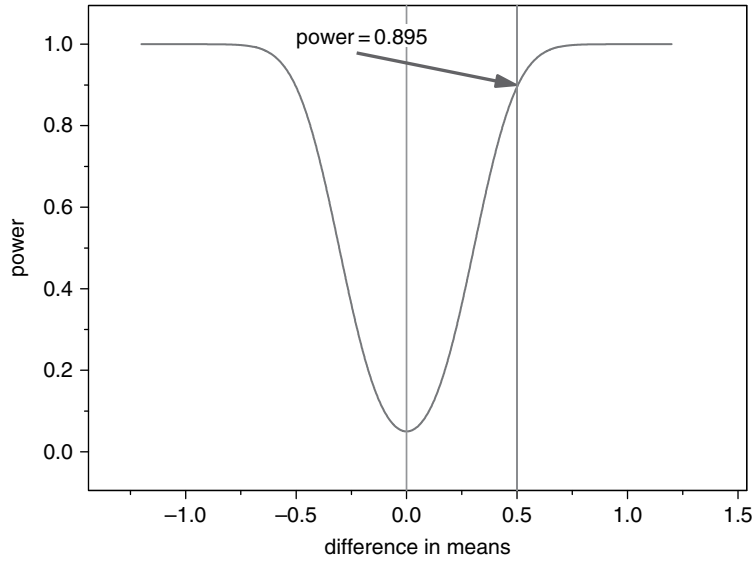


Figure 8.2 Power curve for $n = 100$ per group

the curve actually crosses the y-axis at the significance level, 0.05. Recall that power can be thought of as probability ($p \leq 0.05$). Even when the treatments are identical (difference = 0.0) there is still a 0.05 chance of getting a significant p-value and this is the type I error, and the reason why the power curve cuts at this point. This issue tells as what happens if we want to change the significance level, say for example, from 0.05 to 0.01. (We sometimes do this when dealing with multiplicity and we will look in more detail at this issue in Chapter 10.) Reducing the significance level will pull down the power curve so that it crosses at 0.01 and the effect of this will be to reduce all of the power values. Even when there are true treatment differences, achieving $p \leq 0.01$ is much more difficult than achieving $p \leq 0.05$ and so the power comes down. In practice of course we may need to consider increasing the sample size to compensate for this to recover the required power.

8.3 Calculating sample size

Once the requirements of a trial have been specified then calculating sample size is fairly straightforward; formulas exist for all of the commonly occurring situations.

In all cases we need to specify the required values of the type I error and the power. Usually we set the type I error at 5 per cent and the recommended minimum value for power is 80 per cent, although for important trials 80 per cent is not enough and 90 per cent at least is recommended.

The remaining quantities that need to be considered when calculating sample size depend upon the particular statistical test to be used:

- For the unpaired t-test we need to specify the standard deviation, σ, for the primary endpoint, and the level of effect, d, we are looking to detect with say 90 per cent power.

  There is usually an implicit assumption in this calculation that the standard deviations are the same in each of the treatment groups. Generally speaking this assumption is a reasonable one to make as the effect of treatment will be to change the mean with no effect on the variability. We will say a little more about dealing at the analysis stage with situations where this is not the case in a later section. The sample size calculation, however, is also easily modified, if needed, to allow unequal standard deviations.

- For the paired t-test, the standard deviation of the within-patient differences for the primary endpoint needs to be specified and again, the level of effect to be detected.

- For the χ^2 test we need to know the success/event rate in the control group and as usual some measure of the treatment difference we are looking to detect.

We commonly refer to the level of effect to be detected as the *clinically relevant difference (crd)*; what level of effect is an important effect from a clinical standpoint. Note also that crd stands for 'commercially relevant difference'; it could well be that the decision is based on commercial interests. Finally crd stands for 'cynically relevant difference'! It does happen from time to time that a statistician is asked to 'do a sample size calculation, oh and by the way, we want 200 patients!' The issue here of course is budget and the question really is; what level of effect are we able to detect with a sample size of 200?

The standard deviations referred to above often provide the biggest challenge. The information for this will come from previous data; for that same endpoint, from a similar population/sample of patients, treated for the same period of time etc., similarly for the success/event rate in the control group for binary data. We should try and match as closely as possible the conditions of the historical data with those pertaining to the trial being planned.

Example 8.1: Unpaired t-test

In a placebo-controlled hypertension trial, the primary endpoint is the fall in diastolic blood pressure. It is required to detect a clinically relevant difference of 8 mmHg in a 5 per cent level test. Historical data suggests that $\sigma = 10$ mmHg. Table 8.4 provides sample sizes for various levels of power and differences around 8 mmHg; the sample sizes are per group.

Table 8.4 Sample sizes per group

crd	Power		
	80%	85%	90%
6 mmHg	44	50	59
8 mmHg	24	29	33
10 mmHg	16	18	22

So for 90 per cent power, 33 patients per group are required to detect a difference of 8 mmHg. Smaller differences are more difficult to detect and 59 patients per group are needed to have 90 per cent power to detect a difference of 6 mmHg. Lowering the power from 90 per cent to 80 per cent reduces the sample size requirement by just over 25 per cent.

Example 8.2: χ^2 test

In a parallel group, placebo-controlled trial in acute stroke the primary endpoint is success on the Barthel index at month 3. Previous data suggests that the success rate on placebo will be 35 per cent and it is required to detect an improvement in the active treatment group to 50 per cent. How many patients are needed for 90 per cent power?

For 90 per cent power, 227 patients per group are needed. For 80 per cent power, the sample size reduces to 170 patients per group. If the success rate in the placebo group, however, were to be 40 per cent and not 35 per cent then the sample size requirements per group would increase to 519 for 90 per cent power and 388 for 80 per cent power to detect an improvement to 50 per cent in the active group.

Machin et al. (1997) provide extensive tables in relation to sample size calculations and include in their book the formulas and many examples. In addition there are several software packages specifically designed to perform power and sample size calculations, namely nQuery (www.statsol.ie) and PASS (www.ncss.com). The general statistics package SPLUS (www.insightful.com) also contains some of the simpler calculations.

It is generally true that sample size calculations are undertaken based on simple test procedures, such as the unpaired t-test or the χ^2 test. In dealing with both continuous and binary data it is likely that the primary analysis will ultimately be based on adjusting for important baseline prognostic factors. Usually such analyses will give higher power than the simple alternatives. These more

complex methods of analysis, however, are not taken into account in the sample size calculation for two reasons. Firstly, it would be very complicated to do so and would involve specifying the precise nature of the dependence of the primary endpoint on the factors to be adjusted for and knowledge regarding how those baseline factors will be distributed within the target population. Secondly, using the simple approach is the safe approach as generally speaking the more complex methods of analysis that we end up using will lead to an increase in power.

Finally note that in our considerations we have worked with groups of equal size. It is straightforward to adapt the calculations for unequal randomisation schemes and the computer packages mentioned earlier can deal with these. Altman (1991), Section 15.3 provides a simple method for adapting the standard sample size calculation to unequal group sizes as follows. If N is the calculated sample size based in an equal randomisation and k represents the ratio of the number of patients in one group compared to the other group, then the required number of patients for a k to 1 randomisation is:

$$N' = N\frac{(1+k)^2}{4k}$$

So for example, if a 2 to 1 randomisation is required and 200 patients would have been needed for an equal allocation then the revised sample size is:

$$N' = 200 \times \frac{9}{8} = 225$$

a fairly modest increase. In general a 2 to 1 randomisation will lead to a 12.5 per cent increase in sample size compared to 1 to 1; a 3 to 1 randomisation would lead to a 33.3 per cent increase.

8.4 Impact of changing the parameters

8.4.1 Standard deviation

It is interesting to see the impact of a change in the standard deviation on the required sample size. Consider the example from the previous section where we were looking to detect a treatment effect of 8 mmHg with a standard deviation of 10 mmHg. For 90 per cent power the total sample size requirement was 66 patients. If the standard deviation was not 10 mmHg but 20 mmHg then the required sample size would be 264. A doubling of the standard deviation has led to a four-fold increase in the sample size. The formula for sample size contains not the standard deviation by itself but the variance (= standard deviation squared)

and this is what drives this increase. Even a modest increase in the standard deviation, say from 10 mmHg to 12 mmHg, would require 96 patients in total compared to 66.

There are several implications of this sensitivity of sample size on the standard deviation:

- Good information is needed for the standard deviation, if you get it slightly wrong you could be severely underpowered. Be realistic and if anything conservative.
- Work hard to control the patient-to-patient variability, which not only depends on in-built patient differences, but also on extraneous variability caused by an inconsistent measurement technique, data recording and sloppy methodology. Tightening up on these things will over time bring σ down and help to keep sample sizes lower than they would otherwise be.

8.4.2 Event rate in the control group

Again referring to an example in the previous section, where the event was success on the Barthel index at month 3, we had an event rate in the control group of 35 per cent and we were looking to detect an improvement of 15 per cent in absolute terms to 50 per cent. A sample size of 227 per group gave 90 per cent power. Figure 8.3 illustrates how this sample size depends upon the success rate in the control group; note we are looking in each case for an absolute 15 per cent improvement.

The curve is symmetric around a rate of 0.425 since we are undertaking two-tailed tests and changing the labels for success and failure will simply re-package the same calculation. So, for example, comparing 30 per cent to 45 per cent produces the same sample size as comparing 55 per cent to 70 per cent. The sample sizes are much reduced as the success rates move either down towards 0 per cent or up towards 100 per cent. In those regions, of course, the relative changes in either the success rate or the failure rate are large and it is this that controls the calculation.

8.4.3 Clinically relevant difference

For continuous data, the sample size is inversely proportional to the square of the clinically relevant difference. So if the crd is reduced by a factor of two then the sample size is increased by a factor of four, if the crd is increased by a factor of two then the sample size is reduced by a factor of four. In our earlier example the sample size requirement to detect a difference of 8 mmHg was 33

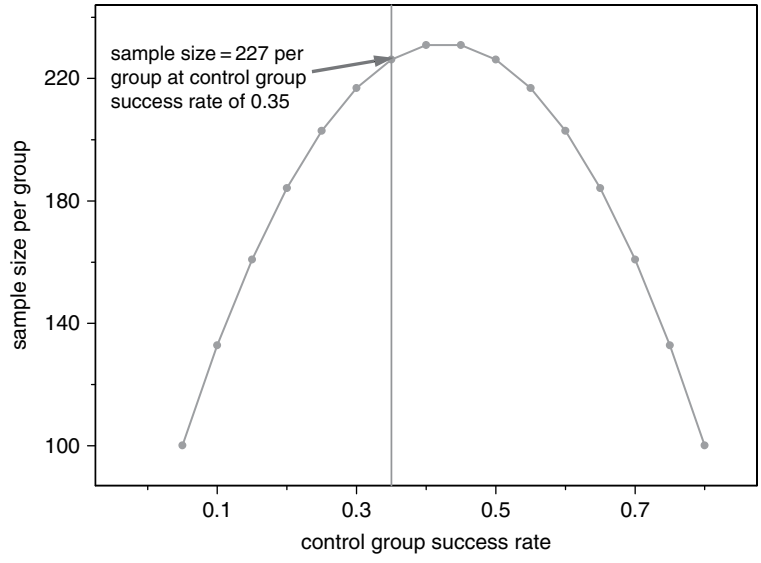


Figure 8.3 Sample size for detecting absolute improvement of 0.15 in success rates

patients per group. To detect a difference of 4 mmHg we require 132 patients per group.

For binary data this same relationship between the crd, in terms of the absolute difference in success rates, and the sample size is only approximately true. In the example we were looking to detect an improvement in the success rate from 35 per cent to 50 per cent, an absolute difference of 15 per cent and we needed a sample size of 227 patients per group. If we were to halve that difference and look for an improvement from 35 per cent to 42.5 per cent then the sample size requirement would be 885 per group, an increase in the sample size by a factor of 3.9.

8.5 Regulatory aspects

8.5.1 Power > 80 per cent

The recommendation for at least 80 per cent power comes from the ICH E9 guideline:

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed . . . The probability of type II error is conventionally set

at 10 per cent to 20 per cent; it is in the sponsor's interest to keep this figure as low as feasible especially in the case of trials that are difficult or impossible to repeat.'

The guideline stresses that it is 'in the sponsor's interest' to have power as high as possible. Too often researchers see the power calculation as merely something that has to go into the protocol to satisfy ethics committees and regulators and go for the 'minimum' requirement. Also it is tempting to choose ambitious values for the key parameters, such as the standard deviation of the primary endpoint or the event rate in the control group or the crd, to produce a sample size that is comfortable from a budgeting or practical point of view, only to be disappointed once the data appears. Be realistic in the choice of these quantities and recognise that 80 per cent power is 20 per cent type II error, a one in five chance of failing to achieve statistical significance even if everything runs perfectly.

8.5.2 Powering on the per-protocol set

Generally speaking we power based on the per-protocol set and then increase the sample size requirement to give the number required in the full analysis set. Under some circumstances, for example in anti-infective trials, we factor up further to take into account the patients that are recruited, but are not eligible for the full analysis set.

So if we need 250 patients for the required power and we expect 10 per cent of patients to be excluded from the full analysis set for the per-protocol analysis then we need to recruit 278 patients (90 per cent of 278 gives 250). The argument here is that because the full analysis set will be larger than the per-protocol set then having enough power for the latter will automatically give enough power for both. Generally this will be true, but sometimes it is not quite so simple. The variability in the outcome measure across the full analysis set may well be larger than that in the per-protocol set and so simply factoring up the sample size to account for the non-evaluable patients may not sufficiently counterbalance the increase in the standard deviation.

In a similar way it may be that the crd seen in the analysis based on the per-protocol set is larger than that seen in the full analysis set and this anticipated difference may also need to be factored in.

It must be noted, however, that even if the sample size calculation gives enough power for the per-protocol analysis the potential for bias in that analysis still remains.

8.5.3 Sample size adjustment

It will sometimes be the case that there are gaps in our knowledge and it will not be possible to give values for the standard deviation or for the event rate in the

control group with any degree of confidence. In these circumstances it is possible to revisit the sample size calculation once the trial is underway.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'In long term trials there will usually be an opportunity to check the assumptions which underlay the original design and sample size calculations. This may be particularly important if the trial specifications have been made on preliminary and/or uncertain information. An interim check conducted on the blinded data may reveal that overall response variances, event rates or survival experience are not as anticipated. A revised sample size may then be calculated using suitably modified assumptions . . . '

Note that this calculation must be undertaken on the blinded data to avoid any formal or informal treatment comparison. If such a comparison were to be made then there would be a price to pay in terms of the type I error. We will say much more about this in a later section dealing with interim analysis where the goal is to formally compare the treatment arms as the data accumulates.

Using all of the data as a single group to calculate the standard deviation in the continuous case, however, will give an overestimate of the within-group standard deviation, particularly if the treatment effect is large. It must be accepted therefore that this could lead to some overpowering as a consequence, although in practice experience suggests that this in minor. For binary data we will end up with a combined event rate and this must be unpicked to enable a sample size recalculation to be done. For example, if the original calculation is based upon detecting an increase in the success rate from 35 per cent to 50 per cent, then we are expecting an overall success rate of 42.5 per cent. If at the interim check this overall rate turns out to be 45 per cent then we are looking for an increase from 37.5 per cent to 42.5 per cent if we want to retain the ability of the trial to have enough power to detect an absolute 15 per cent improvement.

8.6 Reporting the sample size calculation

A detailed statement of the basis of the sample size calculation should be included in the protocol and in the final report. This statement should contain the following:

- Significance level to be used; this will usually be 5 per cent and relate to a two-tailed test or < 5 per cent depending on issues of multiplicity.
- Required power (≥ 80 per cent).

- The primary endpoint on which the calculation is based together with the statistical test procedure.
- Estimates of the basic quantities needed for the calculation such as the standard deviation or the event rate in the control group and the sources of those estimates.
- The clinically/commercially relevant difference (crd). If the expected difference is larger than this then it could be worth considering powering for the expected effect, the sample size will be lower.
- The withdrawal rate to enable the trial to be powered based on those patients who do not withdraw, and the resulting recruitment target.

The CONSORT statement (Moher *et al.* (2001)) sets down standards for the reporting of clinical trials and their recommendations in relation to the sample size calculation are in line with these points.

There may, of course, be cases, especially in the early exploratory phase, where the sample size has been chosen on purely practical or feasibility grounds. This is perfectly acceptable in that context and the sample size section in the protocol should clearly state that this is the case.

The following example is taken from a published clinical trial.

Example 8.3: Xamoterol in Severe Heart Failure

Below is the sample size statement from The Xamoterol in Severe Heart Failure Study Group (1990).

'It was estimated that 228 patients would have to complete the study to give a 90 per cent chance of detecting a 30-second difference in exercise duration between placebo and xamoterol at the 5 per cent level of significance. The aim was therefore to recruit at least 255 patients to allow for withdrawals. A blinded re-evaluation of the variance of the exercise data after the first 63 patients had completed the study, and a higher drop-out rate (15 per cent) than expected (10 per cent) caused the steering committee (in agreement with the safety committee) to revise the recruitment figure to at least 450.'

Consider each of the elements in the calculation and reporting of that calculation in turn:

- The primary endpoint on which the calculation is based is the exercise duration.
- The required power was set at 90 per cent, a type II error of 10 per cent.

Example 8.3: (Continued)

- The type I error was set at 5 per cent. Note in general the alternative phrases for the type I error; significance level, α error, false positive rate.
- Which statistical test was to be used for the comparison of the treatment groups in terms of the primary endpoint do you think? This is a comparison between two independent groups in a parallel group trial and the primary endpoint is continuous so the sample size calculation will undoubtedly have been based on the two-sample t-test (although this is not specified).
- The trial appears to have been powered in terms of the per-protocol set. Recruitment was set at 255 patients to allow for dropouts.
- There were two reasons for increasing the sample size: a larger than expected standard deviation (variance) for the primary endpoint and a higher drop-out rate (15 per cent compared to 10 per cent).

Most of the elements are contained within the sample size section according to the requirements set down in the CONSORT statement; the only omissions seem to be specification of the statistical test on which the sample size calculation was based, the assumed standard deviation of the primary endpoint and the basis of that assumption.

9

Statistical significance and clinical importance

9.1 Link between *p*-values and confidence intervals

In Chapter 3 we developed the concepts of both the confidence interval (CI) and the *p*-value. At that stage these ideas were kept separate. There is in fact a close link between the two and in this section we will develop that link.

Consider an application of the unpaired t-test with:

$$H_0: \ \mu_1 = \mu_2 \qquad H_1: \ \mu_1 \neq \mu_2$$

Note that the null hypothesis could be re-written $H_0: \mu_1 - \mu_2 = 0$.

The 95 per cent confidence interval (a, b) for the difference in the treatment means, $\mu_1 - \mu_2$, provides a range of plausible values for the true treatment difference. With 95 per cent confidence we can say that $\mu_1 - \mu_2$ is somewhere within the range from a to b.

Consider the results presented in Table 9.1 for two separate trials. In trial 1 the *p*-value is indicating that the treatment means are different. The confidence interval is also supporting treatment differences, with the magnitude of that difference lying between 2.1 and 5.7 units with 95 per cent confidence. So at least informally in this case the *p*-value and the confidence interval are telling us similar things; the treatment means are different. In trial 2, the *p*-value is not supporting differences while the 95 per cent confidence interval tells us that μ_1 could be bigger than μ_2 by as much as 4.6 units, but also that μ_2 could be bigger than μ_1 by as much as 1.3 units and everything in between is also possible so certainly 0 is a plausible value for $\mu_1 - \mu_2$. In this case again the *p*-value and the confidence interval are saying similar things and neither is able to discount the equality of the treatment means.

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

Table 9.1 p-values and confidence intervals for 2 trials (hypothetical)

	p-value	95% CI for $\mu_1 - \mu_2$
Trial 1	$p < 0.05$	(2.1, 5.7)
Trial 2	$p =$ NS	(−1.3, 4.6)

In fact the link between the p-value and the confidence interval is not just operating at this informal level, it is much stronger and there is a mathematical connection between the two which goes as follows:

- If $p < 0.05$ then the 95 per cent CI for $\mu_1 - \mu_2$ will exclude zero (and vice versa).
- If $p =$ NS then the 95 per cent CI for $\mu_1 - \mu_2$ will include zero (and vice versa).

Note that if p is exactly equal to 0.05 then one end of the confidence interval will be equal to zero, this is the boundary between the two conditions above.

One element that makes the link work is the correspondence between the significance level (5 per cent) and the confidence coefficient (95 per cent). If we were to use 1 per cent as the cut-off for significance then the same link would apply but now with the 99 per cent confidence interval.

There is a misunderstanding regarding a similar potential link between the p-value and the confidence intervals for the individual means. A significant p-value does not necessarily correspond to non-overlapping confidence intervals for the individual means. See Julious (2004) for further discussion on this issue.

This link applies also to the p-value from the unpaired t-test and the confidence interval for μ, the mean difference between the treatments, and in addition extends to adjusted analyses including ANOVA and ANCOVA and similarly for regression. For example, if the test for the slope b of the regression line gives a significant p-value (at the 5 per cent level) then the 95 per cent confidence interval for the slope will not contain zero and vice versa.

When dealing with binary data a similar link applies, but now with the confidence interval for the odds ratio and the p-value for the χ^2 test, with one important difference; it is the value one (and not zero) that is excluded or included from the confidence interval when p is either significant or non-significant respectively. Recall that for the odds ratio it is the value one that corresponds to equal treatments. The link for binary data is not in fact exact in the strict mathematical sense, but in practice this correspondence can be assumed to apply pretty much all of the time except right on the boundary of 0.05 for the p-value where from time to time one end of the confidence interval may not quite fall on the appropriate side of one. Similar comments to these apply to the relative risk.

9.2 Confidence intervals for clinical importance

Example 9.1: A series of trials in hypertension (hypothetical)

In a collection of four placebo-controlled trials in hypertension a difference of 4 mmHg in terms of mean fall in diastolic bp is to be considered of clinical importance; anything less is unimportant. The results, are given in Table 9.2, where μ_1 and μ_2 are the mean reductions in diastolic bp in the active and placebo groups respectively.

Table 9.2 p-values and confidence intervals for 4 trials

	p-value	95% CI for $\mu_1 - \mu_2$
Trial 1	$p < 0.05$	(3.4 mmHg, 12.8 mmHg)
Trial 2	$p < 0.05$	(1.2 mmHg, 2.9 mmHg)
Trial 3	$p =$ NS	(–3.5 mmHg, 3.1 mmHg)
Trial 4	$p =$ NS	(–2.6 mmHg, 14.3 mmHg)

Note the mathematical connection again with the first two trials giving significant p-values and the second two trials giving non-significant p-values.

Consider now the interpretation of the trials in Example 9.1

- Trial 1 has given statistical significance and has detected something of clinical importance.
- Trial 2 has also given statistical significance, but the difference detected is clinically unimportant.

  In comparing the results from trials 1 and 2 it is clear that the p-value does not tell the whole story. In terms of p-values they are indistinguishable but the first trial has demonstrated a clinically important difference while trial 2 has detected something that is clinically irrelevant.
- Trial 3 has given non-significance statistically and inspecting the confidence interval tells us that there is nothing in terms of clinical importance either; at most with 95 per cent confidence the benefit of the active treatment is only 3.1 mmHg.
- Trial 4 is different, however. We do not have statistical significance, but the confidence interval suggests that there could still be something of clinical importance with potential differences of 5 mmHg, 10 mmHg, even

14 mmHg. This is classically the trial that is too small with low power to detect even large differences.

Again the *p*-value is not giving the whole story. There is clearly nothing of any clinical importance in trial 3 but in trial 4 there could be something worthwhile, it is just that the trial is too small.

It should be clear from the development in this example that statistical significance and clinical importance are somewhat different things. The *p*-value tells us nothing about clinical importance. Just because we have statistical significance it does not mean, necessarily, that we have detected a clinically important effect. Vice versa, just because we have a non-significant result does not necessarily mean the absence of something of clinical importance. The most appropriate way to provide information on clinical benefit is by presenting observed treatment differences together with confidence intervals.

Gardner and Altman (1989) capture the essence of this argument:

'Presenting p-values alone can lead to them being given more merit that they deserve. In particular, there is a tendency to equate statistical significance with medical importance or biological relevance. But small differences of no real interest can be statistically significant with large sample sizes, whereas clinically important effects may be statistically non-significant only because the number of subjects studied was small.'

The regulators are not only interested in statistical significance but also in clinical importance. This allows them, and others, to appropriately balance benefit and risk. It is good practice therefore to present both *p*-values and confidence intervals and indeed this is a requirement within a submission. Most journals nowadays also require results to be presented in the form of confidence intervals in addition to *p*-values.

9.3 Misinterpretation of the *p*-value

9.3.1 Conclusions of similarity

In Section 3.3.1 we defined the *p*-value and briefly mentioned a common incorrect definition. We will return now to discuss why this leads to considerable misinterpretation. In the example of Section 3.3.1 we had observed a treatment difference of 5.4 mmHg with a *p*-value of 4.2 per cent and the proposed definition (incorrect) was that 'there is a 4.2 per cent probability that $\mu_1 = \mu_2$'.

The problem with this definition is the misinterpretation when the *p*-value is large. As an extreme case suppose that we ran a trial in hypertension with two patients per group and also suppose that even though in truth the true treatment

means could be very different, the patients in the active group had bp reductions of 7 mmHg and 5 mmHg respectively while in the placebo group the reductions were 2 mmHg and 10 mmHg. These data give a mean reduction in the active group of 6 mmHg and a mean reduction in the placebo group of 6 mmHg. In a two-sample t-test the resulting *p*-value would be one (signal $= \bar{x}_1 - \bar{x}_2 = 0$). A *p*-value of one or 100 per cent corresponds to certainty and taking the above definition of the *p*-value is telling us on the basis of the observed data it is certain that the true treatment means are identical! I hope we would all agree that this conclusion based on two patients per group would be entirely inappropriate.

This example, of course, is purely hypothetical, but in practice we do see large *p*-values, say of the order of 0.70 or 0.80, which have come from situations where, in truth, the treatments could be very different but we have ended up with a large *p*-value merely as a result of a small sample size or a large amount of patient-to-patient variation (or both) and as a consequence we have a large amount of noise, a small signal-to-noise ratio and a large *p*-value. A *p*-value of this order of magnitude, under the above (incorrect) definition, is giving a probability of something close to certainty that the treatment means are identical.

It is all too common to see a conclusion that treatments are the same (or similar) simply on the back of a large *p*-value; this is not necessarily the correct conclusion. Presentation of the 95 per cent confidence interval will provide a statement about the possible magnitude of the treatment difference. This can be inspected and only then can a conclusion of similarity be made if this interval is seen to exclude clinically important differences. We will return to a more formal approach to this in Chapter 12 where we will discuss equivalence and non-inferiority.

9.3.2 The problem with 0.05

A further aspect of the *p*-value that causes some problems of interpretation is the cut-off for significance at 0.05. This issue was briefly raised in Section 3.3.5 where it was pointed out that 0.05 is a completely arbitrary cut-off for statistical significance and that *p*-values close to 0.05, but sitting on either side of 0.05 should not really lead to different conclusions.

Too often a *p*-value less than 0.05 is seen as definitive proof that the treatments are different while a *p*-value above 0.05 is seen as no proof at all. The *p*-value is a measure of the compatibility of the data with equal treatments, the smaller the *p*-value the stronger the evidence against the null hypothesis. The *p*-value is a measure of evidence in relation to the null hypothesis; treating $p \leq 0.05$, $p > 0.05$ in a binary way as proof/no proof is a gross over-simplification and we must never lose sight of that.

10
Multiple testing

10.1 Inflation of the type I error

Whenever we undertake a statistical test in a situation where the two treatments being compared are the same, there is a 5 per cent probability of getting a statistically significant result purely by chance; this is the type I error. If we were to conduct several tests in this same setting then the probability of seeing one or more significant *p*-values purely by chance will start to mount up. For example, if we were to conduct five tests on independent sets of data, say on five distinct subgroups, then the probability of getting at least one false positive result is 22.6 per cent*. For 50 tests this probability becomes 92.3 per cent, virtual certainty. This should come as no surprise, the 1 in 20 probability of the false positive on each occasion will eventually happen by chance. Certainly with 50 tests the most surprising thing would be if you did not see the false positive on at least one occasion.

The problem with this so-called *multiplicity* or *multiple testing* arises when we make a claim on the basis of a positive result which has been 'generated' simply because we have undertake lots of comparisons. Inflation of the type I error rate in this way is of great concern to the regulatory authorities; they do not want to be registering treatments that do not work. It is necessary therefore to control this inflation. The majority of this chapter is concerned with ways in which the potential problem can be controlled, but firstly we will explore ways in which it can arise.

* The probability of no significant results in five tests is $0.95 \times 0.95 \times 0.95 \times 0.95 \times 0.95 = 0.774$, so the probability of one or more significant results is $1 - 0.774 = 0.226$

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

10.2 How does multiplicity arise

There are a number of settings which can result in multiplicity:

- Multiple endpoints
- Multiple pairwise comparisons in multi-arm trials
- Comparing treatments at multiple time-points
- Comparing treatments within many subgroups
- Interim analyses
- Using different statistical tests on the same data
- Using different analysis sets or different algorithms for missing data

This list is not exhaustive, but these represent the main areas of concern.

We will explore each of these in turn, but before this it is worth making some preliminary points. Firstly, not all multiple testing is a bad thing. For example, it is good practice to analyse several different analysis sets (the final bullet point) to gauge the sensitivity of the results to the choice of analysis set. The problem arises when the results of these comparisons are 'cherry picked' with only those analyses that have given significant results being then used to make a confirmatory claim and those giving non-significant results just ignored or pushed to the background. Secondly, if this process of cherry picking is to be in any sense allowed then there will be a price to pay in terms of reducing the level at which statistical significance can be declared. We will say more about specific methods for making this reduction later, but basically the idea is to divide up the 5 per cent allowable false positive rate across the numerous tests that are going to be the basis of any confirmatory claims. For example, if there are five tests that make up the confirmatory analysis and a claim is going to be made on any of these tests that yield a significant result then the level at which statistical significance can be declared will reduce from 5 per cent to 1 per cent; the argument here is that five lots of 1 per cent make up 5 per cent so the overall type I error rate remains controlled at 5 per cent. For ten tests the *adjusted significance level* (sometimes denoted by α') would be 0.5 per cent. This is the simplest form of adjustment and is known as the *Bonferroni correction*.

10.3 Regulatory view

The regulatory position with regard to multiplicity is well expressed in ICH E9.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'When multiplicity is present, the usual frequentist approach to the analysis of clinical trial data may necessitate an adjustment to the type I error. Multiplicity may arise for, example,

from multiple primary variables, multiple comparisons of treatments, repeated evaluation over time and/or interim analyses. Methods to avoid or reduce multiplicity are sometimes preferable when available, such as the identification of the key primary variable (multiple variables), the choice of a critical treatment contrast (multiple comparisons), the use of a summary measure such as 'area under the curve' (repeated measures). In confirmatory analyses, any aspects of multiplicity which remain after steps of this kind have been taken should be identified in the protocol; adjustment should always be considered and the details of any adjustment procedure or an explanation of why adjustment is not thought to be necessary should be set out in the analysis plan.'

Note that these recommendations relate to confirmatory claims and statements, for post hoc exploratory investigations there are no restrictions in this multiplicity sense. Any findings arising from such analyses, however, cannot be viewed as confirmatory unless the finding can be clearly replicated in an independent setting.

These comments are directed primarily at efficacy and do not tend to be applied to safety, unless a safety claim (e.g. drug A is less toxic than drug B) is to be made. With the routine evaluation of safety, where *p*-values are being used as a flag for potential concerns, we tend to be conservative and not worry about inflating the type I error. It is missing a real effect, the type II error, that concerns us more.

10.4 Multiple primary endpoints

10.4.1 Avoiding adjustment

As mentioned in the previous section, multiplicity can lead to adjustment of the significance level. There are, however, some situations when adjustment is not needed although these situations tend to have restrictions in other ways. We will focus this discussion in relation to multiple primary endpoints and in subsequent sections use similar arguments to deal with other aspects of multiple testing.

As ICH E9 points out: '*There should generally be only one primary variable,*' and when this is the case there is clearly no need for adjustment. However there may well be good scientific and commercial reasons for including more than one primary variable, for example to cover the different potential effects of the new treatment.

10.4.2 Significance needed on all endpoints

In some therapeutic settings the regulators require us to demonstrate effects in terms of two or more endpoints. For example, in asthma, we look for effects both in terms of lung function and symptoms.

CPMP (2002): 'Note for Guidance on the Clinical Investigation of Medicinal Products in the Treatment of Asthma'

'A significant benefit for both primary endpoints, lung function and the symptom based clinical endpoint, should be demonstrated so that no multiplicity adjustment to significance levels is indicated.'

Under such circumstances no adjustment to the significance level is needed; we have to show significance for both endpoints.

10.4.3 Composite endpoints

Another way of avoiding adjustment is to combine the multiple measurements into a single *composite variable*. Examples would be disease-free survival in oncology, where the variable is the time to either disease recurrence or death, whichever occurs first, or a composite of death, non-fatal stroke, MI and heart failure, a binary outcome in a cardiovascular setting. This approach does not require adjustment of the significance level; we are back to having a single primary endpoint.

There are some additional requirements, however, when using composite variables and these relate to the individual components, which should all be supportive of a difference in a single direction. A large positive effect with one of the components could potentially be masking a negative effect in a different component and this would be unacceptable. One way of investigating this would be to consider treatment effects in terms of the components singly. Alternatively, clinically appropriate combinations can be considered. For example, with the binary outcome, death, non-fatal stroke, MI and heart failure, one approach could be to break down the binary outcome into three separate outcome variables at the first level:

- Death and non-fatal stroke
- Death and MI
- Death and heart failure

Each of these variables should be showing an effect in a consistent direction. At the second level, death on its own should be looked at and an effect confirmed. Providing treatment effects are seen to be consistently in the positive direction, these do not necessarily need to be statistically significant, a claim can be made in relation to the composite.

10.4.4 Variables ranked according to clinical importance

It may be possible with several primary endpoints to rank these in terms of their clinical importance and with this structure adopt a testing strategy which

Table 10.1 Hierarchical testing

	Case 1	Case 2	Case 3
Endpoint 1	$p \leq 0.05$	$p \leq 0.05$	$p = \text{NS}$
Endpoint 2	$p \leq 0.05$	$p = \text{NS}$	
Endpoint 3	$p = \text{NS}$		

avoids having to adjust the significance level. This ranking, which of course should be pre-specified in the protocol, then determines the order in which the statistical testing is done. No adjustment to the significance level is required, but claims cannot be made beyond the first non-significant result in the hierarchy. Consider the setting with three primary endpoints, ranked according to their clinical relevance (Table 10.1). In case 1, claims can be made on endpoints 1 and 2. In case 2 a claim can be made on endpoint 1 only, because as endpoint 2 is non-significant then we are not allowed to look at endpoint 3. In case 3 no claims can be made. The CPMP (2002) 'Points to Consider on Multiplicity Issues in Clinical Trials' specifically mentions some examples of this strategy:

'Typical examples are: (i) acute effects in depressive disorders followed by prevention of progression, (ii) reduction of mortality in acute myocardial infarction followed by prevention of other serious events.'

Clearly it is very important that we get the hierarchy correct. Generally this would be determined by the clinical relevance of the endpoints, although under some circumstances it could be determined, in part, by the likelihood of seeing statistical significance with the easier hits towards the top of the hierarchy.

These ideas can also be considered as a way of dealing with secondary endpoints which might be considered for inclusion in a claim. In many cases secondary endpoints are simply primary endpoints lower down in the hierarchy.

There is also the possibility of mixing hierarchical considerations with adjustment. For example, in the case of a single primary endpoint and two secondary endpoints, of equal importance to each other, the primary endpoint would be evaluated at $\alpha = 0.05$ while each of the secondary endpoints would use $\alpha = 0.025$. Claims could only be considered for the secondary endpoints if the primary endpoint gave $p \leq 0.05$, but then additional claims could be made on whichever of the secondary endpoints gives $p \leq 0.025$. In theory the use of both a hierarchy and Bonferroni-type adjustments could move beyond a second level, all that is needed is that 0.05 is assigned to each level of the hierarchy. For example, there could be a single endpoint at the first level, two endpoints at the second level (with a Bonferroni adjustment) and finally a single endpoint at the third level.

Testing at the second level only occurs if the level 1 endpoint is significant at $\alpha = 0.05$, while testing at the third level can take place providing either of the p-values at the second level is statistically significant at $\alpha = 0.025$.

10.5 Methods for adjustment

The Bonferroni method of adjustment has been mentioned earlier in this chapter as a method of preserving the overall 5 per cent type I error rate. In general, if there m confirmatory comparisons and claims to be made on whichever are statistically significant then the Bonferroni correction requires that each comparison be evaluated at level $\alpha' = \alpha/m$. In a strict statistical sense this is the correct adjustment only for tests based on independent sets of data. For example, if there are four non-overlapping (independent) subgroups of patients, for example, 'males aged under 65', 'males aged 65 or over', 'females aged under 65' and 'females aged 65 or over', then an adjustment which uses the 0.0125 level of significance for each of the subgroups will have an overall type I error rate of 5 per cent. In most cases when we use this adjustment, the tests that make up the set of comparisons will not be independent in this sense. With multiple primary endpoints there will possibly be correlation between those endpoints, with multiple treatment comparisons of, say, several dose levels with placebo, the placebo group will be common to those comparisons and hence there will be a connection across the tests and so on. Where this is the case, the Bonferroni correction provides a conservative procedure; in other words the effective overall type I error rate will be less than 5 per cent. As an extreme example of this conservativeness, suppose that two primary endpoints were perfectly correlated. The Bonferroni adjustment would require each of the endpoints be evaluated at the 2.5 per cent level of significance, but because of the perfect correlation the overall type I error rate would also be 2.5 per cent, considerably less than the 5 per cent requirement.

The considerations so far are based on the presumption that the type I error rate is divided equally across all of the comparisons. This does not always make sense and indeed it is not a requirement that it be done in this way. For example, with two comparisons there would be nothing to prevent having a 4 per cent type I error rate for one of the comparisons and a 1 per cent type I error rate for the other, providing this methodology is clearly set down in the protocol. We will see a setting below, that of interim analysis, where it is usually advantageous to divide up the error rate unequally. Outside of interim analysis, however, it is rare to see anything other than an equal subdivision.

Interim analyses arise when we want to look at the data as it accumulates with the possibility of stopping the trial at the interim stage if the data suggests, for example, overwhelming efficacy of the test treatment compared to placebo. If we were to introduce, for example, two interim looks in additional to the final

analysis at the end of the trial then we have an overall testing strategy which consists of three tests and some account of this multiplicity is required. There has been a considerable amount of theory developed in this area and the resulting procedures not only preserve the 5 per cent type error rate, but also do not pay as big a price as Bonferroni. Remember that Bonferroni only strictly applies to independent tests. In the context of interim analysis the data sets that are being analysed are overlapping in a very structured way. With a sample size of 600 and three looks, the first interim analysis after 200 patients provides precisely half of the data on which the second interim analysis, based on 400 patients, is to be undertaken, while these 400 patients provide two-thirds of the data on which the final analysis is to be conducted.

Pocock (1977) developed a procedure which divides the type I error rate of 5 per cent equally across the various analyses. In the example above with two interim looks and a final analysis, Bonferroni would suggest using an adjusted significance level of 0.017 ($= 0.05 \div 3$). The Pocock method however gives us the correct adjusted significance level as 0.022 and this exactly preserves the overall 5 per cent type I error rate.

While this equal division of the type I error may work for some settings, it is more likely that we would want firstly to preserve most of the 5 per cent for the final, and most important analysis and secondly, would only want to stop a trial in the case of strong evidence of overwhelming efficacy. The methods of O'Brien and Fleming (1979) divide up the type I error rate unequally, with very stringent levels at the early interims, becoming less stringent at subsequent analyses, and leaving most of the 5 per cent over for the final analysis. In the case of two interim looks and a final analysis, the adjusted significance levels are 0.00052, 0.014 and 0.045. As can be seen, these adjusted significance levels are very stringent early on and most of the 0.05 has been left over for the final analysis.

The methods as presented here assume that the analyses are equally spaced in terms of the numbers of patients involved at each stage. It is possible to deviate from this in a planned way using so-called *alpha-spending functions*.

It is also possible to stop trials for reasons other than overwhelming efficacy, for example for futility, where at an interim stage it is clear that if the trial were to continue it would have little chance of giving a positive result. We will say more about interim analysis in a later chapter and in particular consider the practical application of these methods.

10.6 Multiple comparisons

In the case of multiple treatment groups it is important to recognise the objectives of the trial. For example, in a three-arm trial with test treatment, active comparator and placebo, the primary objective may well be to demonstrate the effectiveness

of the test treatment and this will be the basis of the claim, while a secondary objective will be to demonstrate the non-inferiority, or perhaps superiority, of the test treatment compared to the active control. This secondary objective may, for example, be driven by market positioning. In this case we have a hierarchy with the primary objective based on a test undertaken at the 5 per cent level of significance, with the test treatment versus active control comparison relegated to a second level in the hierarchy and again this would be conducted with $\alpha = 0.05$. Of course this second comparison cannot be undertaken if the primary objective is not achieved; this makes sense because it would have little value in this scenario if we were unable to demonstrate that the test treatment works.

As a second example consider a trial with four treatment arms: placebo, and low, medium and high doses of drug A. If we wanted to come out of this trial with a confirmatory statement concerning the effectiveness of drug A at a particular dose level then one strategy would be to undertake three tests, each dose level against placebo, and make a claim based on whichever of these is significant. An adjustment would be required and Bonferroni would give an adjusted significance level of 0.017. Alternatively it may be worth considering a hierarchy in the order: high dose versus placebo, medium dose versus placebo, low dose versus placebo, with no adjustment of the 5 per cent significance level. The constraint here, of course, is that you can only make claims up to the first non-significant result. This strategy would get you to the *minimum effective dose* providing things are well behaved and there is an underlying monotonic dose-response relationship (the higher the dose, the bigger the effect).

10.7 Repeated evaluation over time

It is not uncommon to measure variables of interest at several time-points during follow-up. Undertaking statistical testing at each of those time-points is inappropriate and leads to inflation of the type I error. By far the best way to deal with this problem is to reduce the multiple measurements for each subject to a single measure. Several possibilities exist, such as the average of all the measurements, the average of the measurements over the final three months, the achievement of a pre-defined percentage fall from baseline (a binary outcome) and so on. The ICH E9 guideline as quoted earlier mentions area under the curve (AUC); this is simply a sophisticated form of averaging. The chosen measure should be that measure that provides the clearest clinical interpretation.

There are a fairly complex set of statistical techniques, which go under the heading of *repeated measures ANOVA,* that do not summarise the serial measurements for each subject as mentioned above, but leave them separated as they are. These methods then provide *p*-values relating to a comparison of the set of

complete profiles for subjects in treatment group A with the set of profiles for subjects in treatment group B. Other hypotheses, for example, relating to possible differential treatment differences over time can also be evaluated. While these methods offer a statistical solution they are somewhat divorced from methods based on subject level outcomes which can offer a clear clinical interpretation. As a consequence they are only used to a limited extent in practice and the approach based upon summary measures is generally preferred by both regulators and practitioners. Matthews *et al.* (1990) give a general discussion regarding the analysis of repeated measures, but in particular state, in relation to repeated measures ANOVA and similar approaches, that: *'None of these methods provides results that are as easy to understand as the method of summary measures, nor are they as easy to use.'*

The position of the regulatory authorities is best illustrated by a recent Marketing Authorisation Application by Eli Lilly (www.emea.europa.eu/humandocs/PDFs/EPAR/cymbalta/19256704en6.pdf) for their anti-depressant duloxetine. The phase III protocols specified a repeated measures ANOVA as the primary method of analysis and the regulatory submission was based on such analyses. The CPMP, however, asked for an additional simpler analysis based on change from baseline as the outcome measure and using LOCF.

10.8 Subgroup testing

Subgroup testing through a post hoc evaluation of treatment effects within those subgroups can never be used to recover a 'failed' study. If a claim is to be considered for a specific subgroup then this would need to form part of the pre-planned confirmatory strategy. As a very simple example, suppose that a trial is to recruit both males and females, where a claim is to be considered for either males or females (or both) depending on the results of significance tests conducted separately within these two subgroups. The appropriate strategy would then be to use an adjusted significance level of 0.025 (Bonferroni) for each of the subgroups.

Usually evaluation of treatment effects within subgroups is undertaken as part of the assessment of the homogeneity of treatment effect. In large trials it is common to display treatment differences within subgroups defined by important prognostic factors, certainly, for example, those that were used as stratification factors, in the form of point estimates of the treatment effect and 95 per cent confidence intervals for the true treatment difference together, possibly, with a p-value for an assessment of the treatment-by-covariate interaction (see Section 6.5.2). In addition if there were indications that the treatment effect is not homogeneous, then such a display would be of value in explaining the nature of the interaction.

Example 10.1: Pravastatin in preventing cardiovascular disease

Figure 10.1 is taken from Nakamura *et al.* (2006) who reported a large placebo-controlled randomised trial evaluating the effect of pravastatin in preventing cardiovascular disease. The overall treatment effect was positive, with a hazard ratio of 0.67 ($p = 0.01$). We will cover hazard ratios and their use in survival analysis in Chapter 13; for the moment simply note that, like the odds ratio and the relative risk, a value of one corresponds to equal treatments. The homogeneity of the treatment effect was assessed by looking at the p-value for the treatment-by-covariate interaction and also by calculating the hazard ratio separately in various subgroups defined by baseline factors of interest as seen in Figure 10.1.

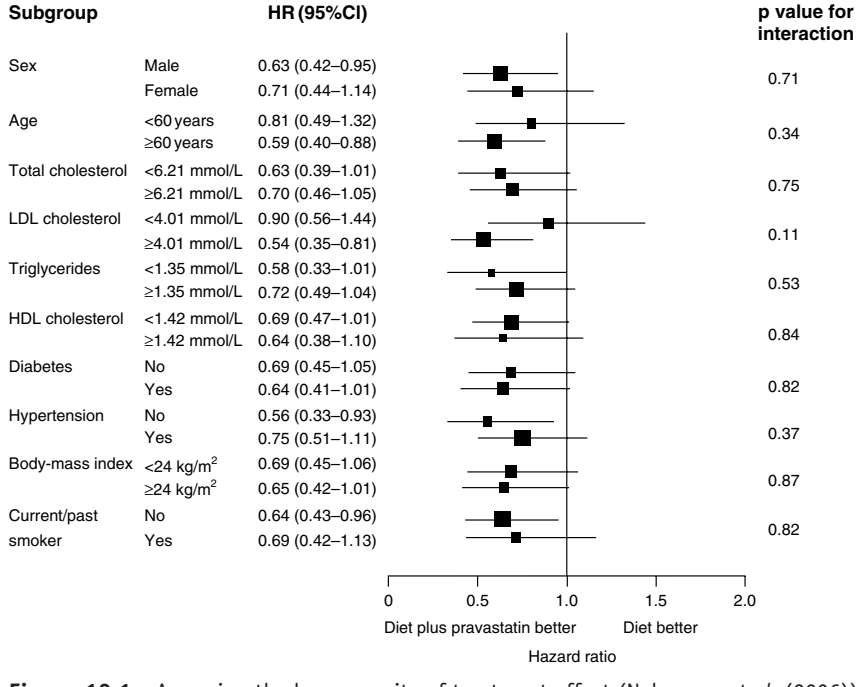


Figure 10.1 Assessing the homogeneity of treatment effect (Nakamura *et al.* (2006)). Reproduced with kind permission from *The Lancet*

Recall from Section 6.5.2 that, when assessing interactions, we use a significance level of 0.10 rather than 0.05 due to a lack of power. In Figure 10.1 most of the interactions, except that involving baseline LDL cholesterol, give p-values well above 0.10 and so there is no evidence of treatment-by-covariate interactions for

most of the baseline factors. This homogeneity of treatment effect can also be seen visually by inspecting the plot and observing similar values for the hazard ratios in the various subgroups, for example, males and females, and 95 per cent confidence intervals that are almost completely overlapping. The *p*-value for the treatment-by-LDL cholesterol is, by contrast, almost significant when compared with the 0.10 cut-off and this provides some evidence of a differential treatment effect.

Note however that the hazard ratio in both high and low baseline LDL cholesterol subgroups is below 1.00 indicating a benefit of pravastatin (a quantitative interaction), although this benefit is marginal in those patients presenting with LDL cholesterol < 4.0 mmol/l.

The regulators are particularly interested in seeing the homogeneity of treatment effect and looking in subgroups will be important to give assurance that a positive treatment effect can be assumed across the population as a whole. If there is evidence that this is not the case, especially if the effect in one subgroup is seen to be negative, then specific subgroups could be excluded from the label.

10.9 Other areas for multiplicity

10.9.1 Using different statistical tests

Using several different statistical methods, for example, an unpaired t-test, an analysis adjusted for centre effects, ANCOVA adjusting for centre and including baseline risk as a covariate, etc., and choosing that method which produces the smallest *p*-value is another form of multiplicity and is inappropriate.

It is good practice to pre-specify in the protocol, or certainly in the statistical analysis plan, the statistical method to be used for analysis for each of the endpoints within the confirmatory part of the trial. This avoids the potential for bias at the analysis stage, which could arise if a method were chosen, for example, which maximised the treatment difference. As a consequence changing the method of analysis following unblinding of the study in an unplanned way, even if there seem sound statistical reasons for doing so, is problematic. Such a switch could only be supported if there was a clear algorithm contained within the statistical analysis plan which specified the rules for the switch. An example of this would be as follows:

> 'The treatment means will be compared using the unpaired t-test. If, however, the group standard deviations are significantly different according to the F-test, then the comparison of the means will be based on Welch's form of the unpaired t-test.'

The blind review does offer an opportunity to make some final changes to the planned statistical methods and this opportunity should not be missed but remember this is based on blinded data.

10.9.2 Different analysis sets

In a superiority trial the primary analysis will be based on the full analysis set with the per-protocol set being used as the basis for a supportive secondary analysis, and in this sense there will be no multiplicity issues. The form of the analysis, however, depends in addition on the methods to be used to account for missing data and these should clearly be pre-specified. It is also good practice to explore the robustness of the conclusions to both the choice of the per-protocol set and the methods to be used for missing data. These analyses again will be supportive (or not) of the main conclusions and no multiplicity aspects arise.

In equivalence and non-inferiority trials (see Chapter 12), the full analysis set and the per-protocol set have equal status and are treated as co-primary. The requirement, therefore, is to show 'significance' for each of these analyses. This is another case where 'significance is needed on all endpoints' with both analyses being conducted at the usual 5 per cent significance level.

11

Non-parametric and related methods

11.1 Assumptions underlying the t-tests and their extensions

The t-tests and their extensions ANOVA, ANCOVA and regression all make assumptions about the distribution of the data in the background populations. If these assumptions are not appropriate then strictly speaking the *p*-values coming out of those tests together with the associated confidence intervals are not valid.

The assumptions are essentially of two kinds: *homogeneity of variance* and *normality*. Consider, to begin with, the unpaired t-test. This test assumes that the two population distributions from which the data are drawn firstly have the same standard deviation (homogeneity of variance) and secondly have the normal distribution shape. In contrast, the paired t-test makes only one assumption and that is that the population of differences at the patient level (for example, response on A − response on B) are normally distributed. For the extensions, ANOVA, ANCOVA and regression, there are both homogeneity of variance and normality assumptions underpinning the methods. We will focus primarily on the simple settings in exploring the issues associated with these assumptions and in presenting other methods that are available if these assumptions do not hold.

There is, in fact, one additional assumption that the above procedures make and that is independence; the way a particular patient responds is not linked to the way another patient responds. In a randomised clinical trial this assumption is unlikely to be violated and we will not discuss this issue further here.

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

11.2 Homogeneity of variance

We will focus in our development in this section on the unpaired t-test. The constant variance assumption can be assessed by undertaking a test (the so-called *F-test*) relating to the hypotheses:

$$H_0 : \sigma_1 = \sigma_2 \quad H_1 : \sigma_1 \neq \sigma_2$$

Here σ_1 and σ_2 are the true standard deviations within treatment groups 1 and 2 respectively. A significant *p*-value from this test would indicate that constant variance cannot be assumed and therefore that the *p*-value coming out of the unpaired t-test is not correct. Should this happen then we would need to use an alternative form of the unpaired t-test that allows for non-constant variance. This form of the unpaired t-test is known as *Welch's approximation*. It involves a slightly different formula for the standard error of the difference $\bar{x}_1 - \bar{x}_2$ and a different calculation of the degrees of freedom for the t-distribution on which the *p*-value calculation is based. These details need not concern us here, suffice it to say that the issue of non-constant variance is fairly straightforward to deal with. In addition, our experience in clinical trials tells us that non-constant variance does not seem to occur particularly often in practice and when it does it tends to be associated additionally with violation of the normality assumption. Conveniently, taking care of the normality assumption by transforming the data (see Section 11.4) often also takes care of the non-constant variance.

11.3 The assumption of normality

While constancy of variance does not seem to be too much of a concern in our clinical trials, it is not uncommon for the assumption of normality to be violated. Many laboratory variables do not usually display the normal distribution shape, while in pharmacokinetics several of the quantities that we routinely calculate, such as AUC and C_{max}, frequently have distributions which appear as in Figure 11.1. We talk in terms the data being *positively skewed*. In part, the reason why we often see these positively skewed distributions is that there is a physical boundary at zero; it is not possible to observe a negative value, although it is possible to see larger values for some patients well away from the bulk of the data.

Checking the assumption of normality can be undertaken in one of two ways. Firstly we have graphical methods, such as a *quantile–quantile plot* (also known as a *normal probability plot*), where normal data displays itself as a straight line. Departures from a straight line plot are indicative of non-normality. Figure 11.2 is a quantile–quantile plot to assess the normality of 100 observations simulated from the distribution displayed in Figure 11.1 while Figure 11.3 shows the histogram

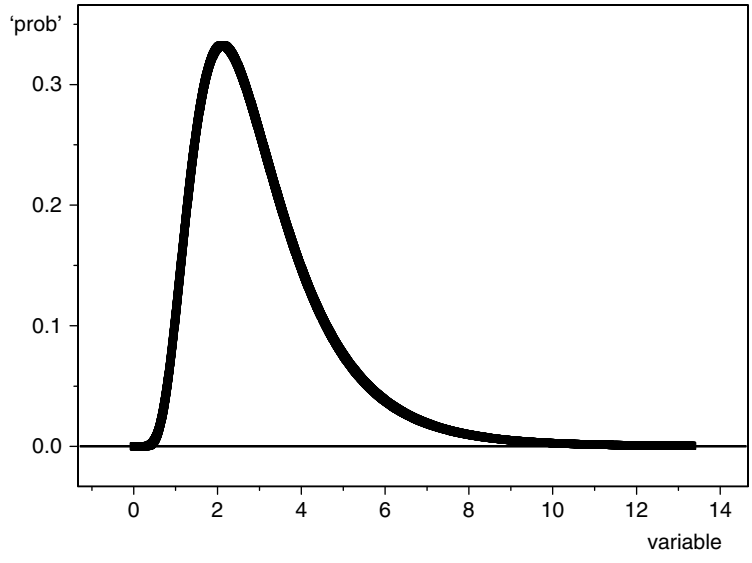


Figure 11.1 Positively skewed (non-normal) distribution

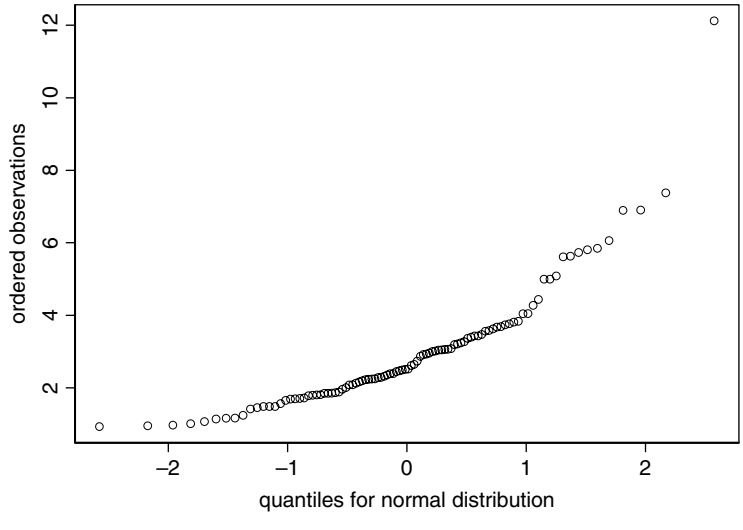


Figure 11.2 Quantile–quantile plot to assess normality

for these same data. The quantile–quantile plot clearly does not conform well to a straight line and the histogram reflects the positive skew nature of these data.

This visual approach based on inspecting the normal probability plot may seem fairly crude. However, most of the test procedures, such as the unpaired t-test, are what we call *robust* against departures from normality. In other words, the

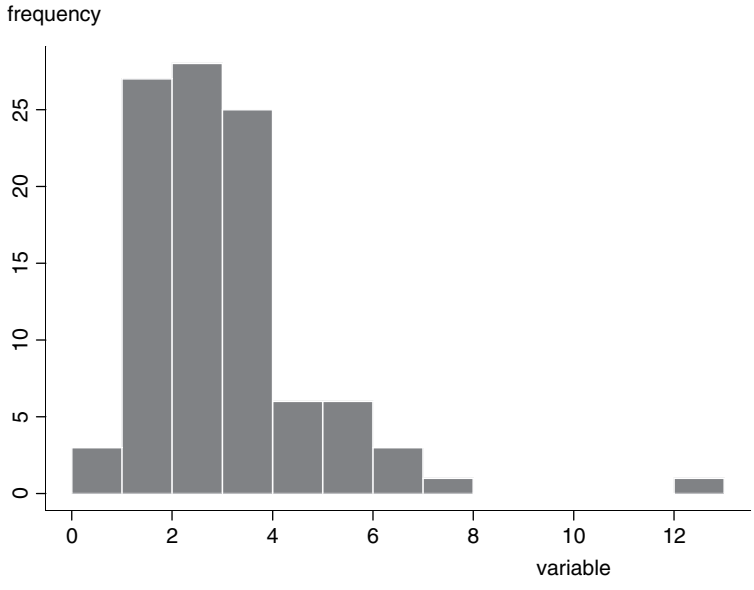


Figure 11.3 Histogram for simulated data

p-values resulting from these tests remain approximately correct unless we depart substantially from normality, particularly with large sample sizes. The normal probability plot is sensitive enough to detect such substantial departures.

Secondly we have a statistical test, the *Shapiro–Wilks test*, which gives a *p*-value for the following setting:

$$H_0: \text{ normal} \quad H_1: \text{ non-normal}$$

A significant *p*-value is indicating that the data is not normally distributed and leads to the rejection of H_0; a non-significant *p*-value tells us that there is no evidence for non-normality and in practice it will then be safe to assume that the data is at least approximately normally distributed.

Our discussions here suggest that we look for normality in each of the treatment groups separately. In practice this is not quite what we do; we look in fact at the two groups combined and evaluate the normality of the so-called *residuals*. This approach 'standardises' the data values in each of the two groups to have mean zero by subtracting the observed mean in group A from the observations in group A and the observed mean in group B from the observations in group B. This standardisation gives the combined group of adjusted observations the same mean of zero and allows all of the data to be considered as a single group for the purpose of evaluating the assumption of normality. In a similar way this approach is used to deal with ANOVA where there are several 'groups' (A and

B observations in each of several centres) and also with more complex structures which form the basis of ANCOVA and regression. For example, in regression the assumption of normality applies to the vertical differences between each patient's observation y and the value of y on the underlying straight line that describes the relationship between x and y. We therefore look for the normality of the residuals: the vertical differences between each observation and the corresponding value on the fitted line.

11.4 Transformations

We will concentrate in this section on the parallel group case with two treatments where for normally distributed data we would be undertaking the unpaired t-test. If the data are clearly non-normal then the first approach in analysing the data would be to consider transforming the data to recover normality. As we have mentioned in the previous section it is not uncommon to have data which are positively skewed with values which cannot be negative. A transformation which often successfully transforms the data to be normal is the log transformation. It does not matter to which base we take these logs; the usual choices would be to base e (natural logarithms) or to base 10, either of these is equal to a constant times the other. Table 11.1 shows the effect of taking logs, to base 10, of various values.

The effect of the log transformation on the values 1, 10, 100 and 1000 is to effectively bring them closer together. On the original scale these numbers are getting progressively further apart whereas on the log scale they become equally spaced. Also for values on the original scale between zero and one the log transformation gives a negative value. The log transformation 'brings in' the large positive values and 'throws out' the values below one to become negative and this has the effect of making the positively skewed distribution look more symmetric. If this transformation is successful in recovering normality then we simply analyse the data on the log scale using the unpaired t-test. The resulting p-value provides a valid comparison of the two treatments in terms of the means on the log scale.

Table 11.1 The log transformation

x	$\log_{10}x$
0	$-\infty$
1	0
10	1
100	2
1000	3

Similarly we can calculate 95 per cent confidence intervals for the difference in the means on the log scale.

While the *p*-value allows us the ability to judge statistical significance, the clinical relevance of the finding is difficult to evaluate from the calculated confidence interval because this is now on the log scale. It is usual to 'back-transform' the lower and upper confidence limits, together with the difference in the means on the log scale, to give us something on the original data scale which is more readily interpretable. The back-transform for the log transformation is the anti-log.

Mean values are usually calculated as *arithmetic means*. However, there is another kind of mean value, the *geometric mean*. The arithmetic mean is $(x_1 + x_2 + \ldots + x_n)/n$ while the geometric mean is defined as $\sqrt[n]{x_1 \times x_2 \times \ldots x_n}$: the *n*th root of all the data values multiplied together. When the anti-log is applied to the difference in the means on the log scale then the result is numerically equal to the ratio of the geometric means for the original data. This is why in pharmacokinetics, where it is standard practice to log transform $C_{\max}$ and AUC prior to analysis, it is the ratio of geometric means together with confidence intervals for that ratio that are quoted. More generally, we often quote geometric means where we are using the log transformation in the analysis of data.

In the paired t-test setting it is the normality of the differences (response on A − response on B) that is required for the validity of the test. The log transformation on the original data can sometimes be effective in this case in recovering normality for these differences. In other settings, such as ANOVA, ANCOVA and regression, log transforming the outcome variable is always worth trying, where this is a strictly positive quantity, as an initial attempt to recover normality.

It is worth noting finally that log transforming positively skewed data often kills two pigeons with one pebble; recovering both normality and constant variance. With skewed data the group with the larger outcome values will tend to have more spread and the log transformation will then generally bring the spread of the data in each of the groups into line.

The log transformation is by far the most common transformation, but there are several other transformations that are from time to time used in recovering normality. The *square root transformation*, $\sqrt{x}$, is sometimes used with count data while the *logit transformation*, $\log(x/1-x)$, can be used where the patient provides a measure which is a proportion, such as the proportion of days symptom-free in a 14 day period. One slight problem with the logit transformation is that it is not defined when the value of x is either zero or one. To cope with this in practice, we tend to add 1/2 (or some other chosen value) to x and $(1-x)$ as a 'fudge factor' before taking the log of the ratio.

Figure 11.4 is the quantile–quantile plot for the log transformed data from Figure 11.3 while Figure 11.5 is the histogram of these same data. The quantile–quantile plot is approximately linear indicating that the log transformation has

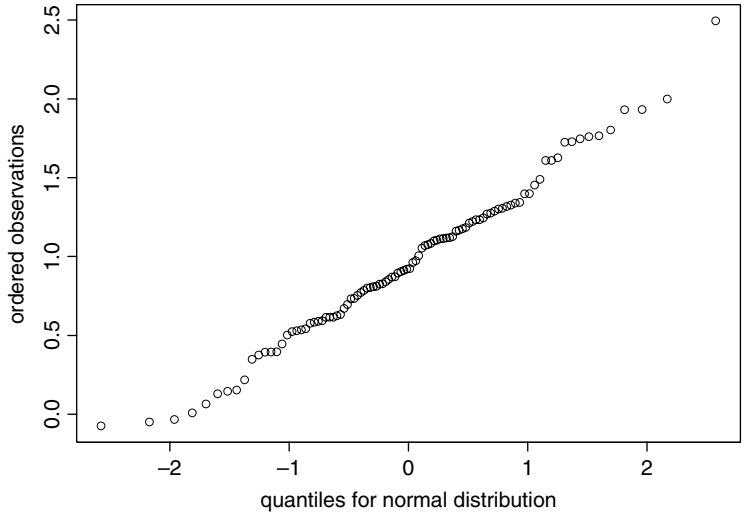


Figure 11.4 Quantile–quantile plot for log transformed data

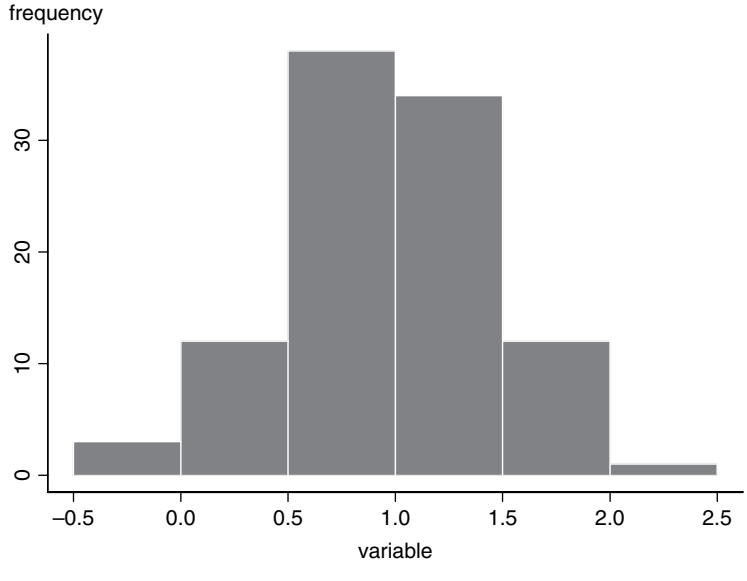


Figure 11.5 Histogram for log transformed data

recovered normality for these data and the histogram clearly conforms more closely to the normal distribution shape. An assumption of normality would now be an entirely reasonable assumption to make.

11.5 Non-parametric tests

11.5.1 The Mann–Whitney U-test

The following hypothetical data in Table 11.2 are simulated from two distinct non-normal distributions. The population, from which the observations in group A are taken, has a mean of 1 while the population from which the B group observations were taken has mean equal to 1.5.

The Mann–Whitney U-test is equivalent to an alternative test called the *Wilcoxon rank sum test*. These tests were developed independently, but subsequently shown to be mathematically the same. We will develop the test using the Wilcoxon rank sum methodology.

The first step in undertaking this test is to order (or *rank*) the observations in both groups combined, from smallest to largest. This ranking is seen in Table 11.2. There are 31 observations in total and the average of the numbers 1 through to 31 is 16 $(16 = (1 + 31)/2)$. If there were no differences on average between the two populations then the average rank in each of the two groups would be close

Table 11.2 Hypothetical non-normal data for two groups

Group A ($n = 15$)	Rank	Group B ($n = 16$)	Rank
1.45	1	3.54	18
2.55	11	3.04	15
4.35	21	5.33	24
1.93	5	3.16	17
3.95	19	3.10	16
7.90	29	2.69	12
1.78	3	2.95	14
4.60	22	5.89	26
1.84	4	5.17	23
2.54	10	2.13	7
5.47	25	4.30	20
2.29	8	8.48	30
2.91	13	12.88	31
1.46	2	6.43	27
2.30	9	2.01	6
		7.03	28

to 16. The signal for the Wilcoxon rank sum test is the difference between the observed average rank in the smaller of the two groups (group A in this case) minus the expected average rank (equal to 16) under the assumption that the treatments on average are the same. The average of the ranks attached to the 15 observations in group A is equal to 12.13 (= 182/15). This average is below 16 and is an indication that in group A the observations look to be smaller than the observations in group B and the value of the signal is $-3.87 (= 12.13 - 16)$. The noise attached to this signal is the associated standard error and this is given by:

$$se = \sqrt{\frac{n_2 (n_1 + n_2 + 1)}{12 n_1}} = \sqrt{\frac{16 \times (15 + 16 + 1))}{12 \times 15}} = 1.69$$

The signal-to-noise ratio for these data is then -2.29. We obtain the *p*-value by comparing the value of this test statistic with what we would expect to see if the treatments were the same. This null distribution is a special form of the normal distribution, called the *standard normal*; the normal distribution with mean zero and standard deviation equal to one. The *p*-value turns out to be 0.022, a statistically significant result indicating treatment differences. Inspecting the data it is clear that it is group A that is, on average, giving smaller values.

This methodology assumes that all of the observations are distinct so that a unique ranking can be defined. In many cases, however, this will not be true and we will see tied values in the data. In these situations we assign average ranks to the tied values. For example had the observations 2.54 and 2.55 both been equal to 2.55 then we would have attached the rank 10.5 (the average of the ranks 10 and 11 corresponding to these two observations) to both values; the ranking then proceeds 8, 9, 10.5, 10.5, 12, 13 and so on. With three values all equal, say in positions 14, 15 and 16 then the average rank 15 is attached to each of the observations. Providing the number of tied values is not too large then the same formulas as above can be used to calculate the value of the test statistic. For more frequent ties then the formula for the standard error needs to be modified. More details can be found in van Belle *et al.* (2004), Section 8.6.3.

Example 11.1: Natalizumab in the treatment of relapsing multiple sclerosis

Miller *et al.* (2003) report a trial comparing two dose levels of natalizumab (3 mg per kg and 6 mg per km) with placebo. The primary endpoint was the number of new brain lesions during the six month treatment period. Table 11.3 presents the data.

Example 11.1: (Continued)

Table 11.3 Number of new enhancing lesions

	Placebo ($n = 71$)	Natalizumab (3 mg, $n = 68$)	Natalizumab (6 mg, $n = 74$)
No lesions	23 (32%)	51 (75%)	48 (65%)
1–3 lesions	18 (25%)	14 (21%)	20 (27%)
4–6 lesions	13 (18%)	1 (1%)	5 (7%)
7–9 lesions	0	0	0
10–12 lesions	3 (4%)	1 (1%)	0
> 12 lesions	14 (20%)	1 (1%)	1 (1%)

The distribution of number of new lesions (count data) is clearly not normal within each of the treatment groups. There is a peak at zero in each of the groups with then fewer and fewer patients as the number of lesions increases. A log transformation would not work here because of the presence of the zero values for the endpoint. The authors used the Mann–Whitney U-test to compare each of the natalizumab dose groups with placebo obtaining $p < 0.001$ in each case. Each dose level is significantly better than placebo in reducing the number of new enhancing lesions.

11.5.2 The Wilcoxon signed rank test

This test is the non-parametric equivalent of the paired t-test. Recall from Section 11.3 that the paired t-test assumes that the population of differences for each patient follows the normal shape. If this assumption is violated then the paired t-test does not apply although, as with the unpaired t-test, the paired t-test is fairly robust against modest departures from normality.

The test is again based on a ranking procedure. Under the assumption that the treatments being compared, A and B, are the same then the number of positive A − B differences should be equal to the number of negative A − B differences. For example, suppose there are 12 patients, then under the assumption that the treatments are the same we should see approximately six positive A − B differences and six negative A − B differences. Further, the magnitude of the positive differences and the negative differences should look similar. Having calculated the A − B differences, the first step is to assign ranks to all of the patients according to the magnitude of those differences (ignoring the sign). Secondly we add up the ranks attached to those differences that were positive. The average rank of the positive differences should be equal to the average rank of the negative differences

and both should be equal to 6.5 (the average of the numbers 1 to 12) under the null hypothesis of no treatment differences. The signal for the test statistic is formed from the observed average rank for the positive differences minus the expected average rank for those positive differences; this latter quantity is 6.5 in the example. The standard error associated with the observed average rank for the positive differences then makes up the noise and we compare the signal-to-noise ratio with the standard normal distribution to give us the *p*-value.

Had we chosen to calculate the differences B − A, then the signal would have exactly the same as the signal based on the A − B differences, but with the opposite sign, and the two-sided *p*-value would be completely unchanged.

11.5.3 General comments

Non-parametric tests, as seen in the two procedures outlined earlier in Section 11.5, are based on some form of ranking of the data. Once the data are ranked then the test is based entirely on those ranks; the original data play no further part. It is therefore the behaviour of ranks that determines the properties of these tests and it is this element that gives them their robustness. Whatever the original data looks like, once the rank transformation is performed then the data become well-behaved.

It may seem strange to see the normal distribution play a part in the *p*-value calculations in Section 11.5.1 and 11.5.2. The appearance of this distribution is in no sense related to the underlying distribution of the data. For the Mann–Whitney U-test for example it relates to the behaviour of the average of the ranks within each of the individual groups under the assumption of equal treatments where the ranks in those groups of sizes n_1 and n_2 are simply a random split of the numbers 1 through to $n_1 + n_2$.

In terms of summary statistics, means are less relevant because of the inevitable skewness of the original data (otherwise we would not be using non-parametric tests). This skewness frequently produces extremes, which then tend to dominate the calculation of the mean. Medians are usually a better, more stable, description of the 'average'.

Extending non-parametric tests to more complex settings, such as regression, ANOVA and ANCOVA is not straightforward and this is one aspect of these methods that limits their usefulness.

11.6 Advantages and disadvantages of non-parametric methods

It is fair to say that statisticians tend to disagree somewhat regarding the value of non-parametric methods. Some statisticians view them very favourably while others are reluctant to use them unless there is no other alternative.

Clearly the main advantage of a non-parametric method is that it makes essentially no assumptions about the underlying distribution of the data. In contrast, the corresponding parametric method makes specific assumptions, for example, that the data are normally distributed. Does this matter? Well, as mentioned earlier, the t-tests, even though in a strict sense they assume normality, are quite robust against departures from normality. In other words you have to be some way off normality for the *p*-values and associated confidence intervals to be become invalid, especially with the kinds of moderate to large sample sizes that we see in our trials. Most of the time in clinical studies, we are within those boundaries, particularly when we are also able to transform data to conform more closely to normality.

Further, there are a number of disadvantages of non-parametric methods:

- With parametric methods confidence intervals can be calculated which link directly with the *p*-values; recall the discussion in Section 9.1. With non-parametric methods the *p*-values are based directly on the calculated ranks and it is not easy to obtain a confidence interval in relation to parameters that have a clinical meaning that link with this. This compromises our ability to provide an assessment of clinical benefit.
- Non-parametric methods reduce power. Therefore if the data are normally distributed, either on the original scale or following a transformation, the non-parametric test will be less able to detect differences should they exist.
- Non-parametric procedures tend to be simple two group comparisons. In particular, a general non-parametric version of analysis of covariance does not exist. So the advantages of ANCOVA, correcting for baseline imbalances, increasing precision, looking for treatment-by-covariate interactions, are essentially lost within a non-parametric framework.

For these reasons non-parametric methods are used infrequently within the context of clinical trials and they tend only to be considered if it is clear that a corresponding parametric approach, either directly or following a data transformation, is unsuitable.

11.7 Outliers

An *outlier* is an unusual data point well away from most of the data. Usually the outlier in question will not have been anticipated and the identification of these points and appropriate action should be decided at the blind review.

The appropriate method for dealing with an outlier will depend somewhat on the setting, but one or two general points can be made. The first thing that should

be done is to check that the value is both possible from a medical perspective and correct. For example, a negative survival time is not possible and this could well have been as a result of an incorrect date at randomisation being recorded. Hopefully these problems will have been picked up by data management, but sometimes things slip through. Clearly, if the data point is incorrect then it should be corrected before analysis.

An extreme, large positive, value may sometimes be a manifestation of an underlying distribution of data that is heavily skewed. Transforming the data to be more symmetric may then be something to consider.

Analysing the data with and without the outliers may ultimately be the appropriate approach, just to ensure that the conclusions are unaffected by their presence. ICH E9 provides some clear guidance on this point.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'If no procedure for dealing with outliers was foreseen in the trial protocol, one analysis with the actual values and at least one other analysis eliminating or reducing the outlier effect should be performed and differences between their results discussed.'

12
Equivalence and non-inferiority

12.1 Demonstrating similarity

In this chapter we will move away from superiority trials to look at methods for the evaluation of equivalence and non-inferiority. The setting in all cases here is the comparison of a new treatment to an active control where we are looking to demonstrate similarity (in some defined sense) between the two treatments.

It should be clear from our earlier development, especially the discussion in Sections 9.2 and 9.3, that obtaining a non-significant p-value in a superiority evaluation does not demonstrate that the two treatments are the same; a non-significant p-value may simply be the result of a small trial, with low power even to detect large differences. ICH E9 makes a clear statement in this regard.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Concluding equivalence or non-inferiority based on observing a non-significant test result of the null hypothesis that there is no difference between the investigational product and the active comparator is inappropriate.'

Unfortunately this issue is not well understood within the clinical trial community and the misinterpretation of non-significant p-values is all too common. See Jones *et al.* (1996) for further discussion on these points.

Equivalence trials are, of course, routinely used in the evaluation of bioequivalence and the methodology there is well established, both European and FDA guidelines exist. More recently we have seen the need to establish therapeutic equivalence and Ebbutt and Frith (1998) provide a detailed case study in the development of an alternative propellant for the asthma inhaler. More usually,

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

however, in a therapeutic setting we will use a non-inferiority design, where we are looking to establish that our new treatment is 'at least as good' or 'no worse than' an existing treatment. We will of course need to define 'at least as good' or 'no worse than' in an operational sense for this to be unambiguous.

One question that often arises is; shouldn't such evaluations always be based on non-inferiority rather than equivalence, because surely having a new treatment which is potentially more efficacious than an existing treatment is a positive outcome that we would want to allow? Well usually this would be the case, but not always. The Ebbutt and Frith (1998) case study is one such situation. The standard metered dose inhaler has a CFC gas as the propellant and this causes environmental damage. An alternative propellant was sought to reduce the environmental burden. The primary endpoint for the studies covered by Ebbutt and Frith was PEF, peak expiratory flow rate, a measure of lung function. The trials developing the alternative propellant were all equivalence trials making comparisons between the two devices, the existing inhaler, and the new inhaler with the alternative propellant. It was necessary to show that the new inhaler provided an increase if PEF that was on average the same as that provided by the existing inhaler. The requirement was to match the effectiveness of the two inhalers as eventually the new inhaler would be used as a substitute for the existing inhaler. Should, for example, the new inhaler result in a substantially greater increase in PEF then this could mean that the new device was delivering a higher dosage, an unsatisfactory situation which could be associated with safety issues. More recently we have seen guidance from the CHMP (CHMP (2006) 'Guideline on Similar Biological Medicinal Products Containing Biotechnology-Derived Proteins as Active Substance: Non-Clinical and Clinical Issues') with regard to establishing similarity and equivalence methodology is also required in this case.

As mentioned in Section 1.10, there are essentially two areas where we would want to conduct non-inferiority trials; firstly where inclusion of a placebo for either practical or ethical issues is not possible and we are therefore looking to demonstrate the efficacy of the new treatment indirectly by showing similarity to an established active treatment and secondly where it is necessary to show that there is no important loss of efficacy for a new treatment compared to an existing treatment.

Finally, before we move on to look at statistical methods, it is worth mentioning that many people feel uncomfortable with the term non-inferiority. In a strict sense, any reduction in the mean response is saying that the new treatment is not as good as the existing treatment and so is inferior. We, however, are using the term non-inferiority to denote a non-zero, but clinically irrelevant reduction in efficacy, which we need to define in an appropriate way. Some practitioners use the term *one-sided equivalence* as an alternative to non-inferiority.

A good overview of various aspects of non-inferiority trials is provided by Kaul and Diamond (2006).

12.2 Confidence intervals for equivalence

We will start by looking at equivalence and then move on to consider non-inferiority. The first step in establishing equivalence is to define equivalence. Following Ebbutt and Frith (1998), suppose we are looking to establish the equivalence of a new asthma inhaler device with an existing inhaler device in a trial setting and further suppose that our clinical definition of equivalence is 15 l/min. In other words, if the difference in the mean increase in PEF following four weeks of treatment is less than 15 l/min then we will conclude that the two devices provide a clinically equivalent benefit. We may want to argue over whether 15 l/min is the appropriate value, but whatever we do we must choose a value. The ± 15 l/min values are termed the *equivalence margins* and the interval -15 l/min to $+15$ l/min is also called the *equivalence region* (see Figure 12.1).

The next step is to undertake the trial and calculate the 95 per cent confidence interval for the difference in the means (mean increase in PEF on new inhaler (μ_1) – mean increase in PEF on existing inhaler (μ_2)). As a first example, suppose that this confidence interval is (-7 l/min, 12 l/min). In other words, we can be 95 per cent confident that the true difference, $\mu_1 - \mu_2$, is between 7 l/min in favour of the existing inhaler and 12 l/min in favour of the new inhaler.

As seen in Figure 12.1, this confidence interval is completely contained between the equivalence margins -15 l/min to 15 l/min and all of the values for the treatment difference supported by the confidence interval are compatible with the definition of clinical equivalence; we have established equivalence as defined.

In contrast, suppose that the 95 per cent confidence interval had turned out to be (-17 l/min, 12 l/min). This interval is not entirely within the equivalence margins and the data are supporting potential treatment differences below the lower equivalence margin. In this case we have not established equivalence.

Note that there are no conventional *p*-values here. Such *p*-values have no role in the evaluation of equivalence; establishing equivalence is based entirely on the use of confidence intervals.

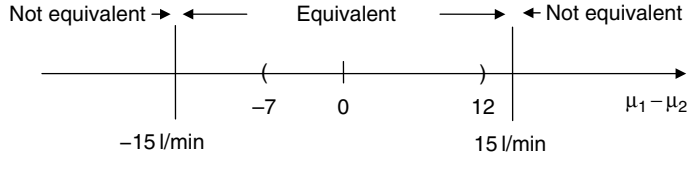


Figure 12.1 Establishing equivalence

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Statistical analysis is generally based on the use of confidence intervals. For equivalence trials, two-sided confidence intervals should be used. Equivalence is inferred when the entire confidence interval falls within the equivalence margins.'

The confidence intervals we have used to date are all two-sided. We will talk later about one-sided confidence intervals.

12.3 Confidence intervals for non-inferiority

For non-inferiority, the first step involves defining a non-inferiority margin. Suppose that we are developing a new treatment for hypertension and potentially the reason why the new treatment is better is that it has fewer side effects, although we are not anticipating any improvement in terms of efficacy. Indeed, suppose that we are prepared to pay a small price for a reduction in the side effects profile; say up to 2 mmHg in the mean reduction in diastolic blood pressure.

In Figure 12.2, μ_1 and μ_2 are the mean reductions in diastolic blood pressure in the test treatment and active control groups respectively. If the difference in the means is above zero then the test treatment is superior to the active control, if the difference is zero then they are identical. If the difference falls below zero the test treatment is not as good as the active control. This, however, is a price we are prepared to pay, but only up to a mean reduction in efficacy of 2 mmHg; beyond that, the price is too great. The non-inferiority margin is therefore set at -2 mmHg.

Step 2 is then to run the trial and compute the 95 per cent confidence interval for the difference, $\mu_1 - \mu_2$, in the mean reductions in diastolic blood pressure. In the above example suppose that this 95 per cent confidence interval turns out to be (-1.5 mmHg, 1.8 mmHg). As seen in Figure 12.2, all of the values within this interval are compatible with our definition of non-inferiority; the non-inferiority of the test treatment has been established. In contrast, had the 95 per cent confidence interval been, say, (-2.3 mmHg,

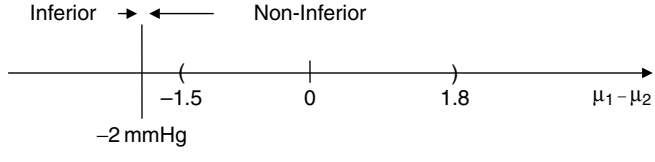


Figure 12.2 Establishing non-inferiority

1.8 mmHg) then non-inferiority would not have been established since the lower end of that confidence interval falls below −2 mmHg. Note again that there is no mention of conventional *p*-values, they have no part to play in non-inferiority.

In order to demonstrate non-inferiority, it is only one end of the confidence interval that matters; in our example it is simply the lower end that needs to be above −2 mmHg. It is therefore not really necessary to calculate the upper end of the interval and sometimes we leave this unspecified. The resulting confidence interval with just the lower end is called a *one-sided 97.5 per cent confidence interval*; the two-sided 95 per cent confidence interval cuts off 2.5 per cent at each of the lower and upper ends, having the upper end undefined leaves just 2.5 per cent cut off at the lower end. The whole of this confidence interval must be entirely to the right of the non-inferiority margin for non-inferiority to be established.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'For non-inferiority a one-sided confidence interval should be used.'

Example 12.1: Fluconazole compared to amphotericin B in preventing relapse in cryptococcal meningitis

The aim of this study reported by Powderly *et al.* (1992) was to establish the non-inferiority of a test treatment, fluconazole, compared to an established treatment, amphotericin B, in preventing the relapse of cryptococcal meningitis in HIV-infected patients. It was thought that fluconazole would be less effective than amphotericin B, but would offer other advantages in terms of reduced toxicity and ease of administration; fluconazole was an oral treatment while amphotericin B was given intravenously. The non-inferiority margin was set at −15 per cent in terms of relapse rates.

$$\text{Let } \theta_F = \text{relapse rate on fluconazole}$$

$$\text{Let } \theta_A = \text{relapse rate on amphotericin B}$$

In Figure 12.3, a positive difference for $\theta_A - \theta_F$ is indicating that fluconozole is a better treatment, a negative difference is indicating that amphotericin B is better.

The non-inferiority margin has been set at −15 per cent. Figure 12.4 displays the non-inferiority region and we need the (two-sided) 95 per cent confidence interval, or the one-sided 97.5 per cent confidence interval, to be entirely within this non-inferiority region for non-inferiority to be established.

Example 12.1: (Continued)

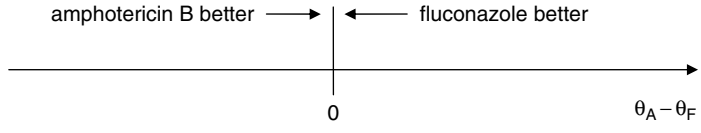


Figure 12.3 Difference in relapse rates

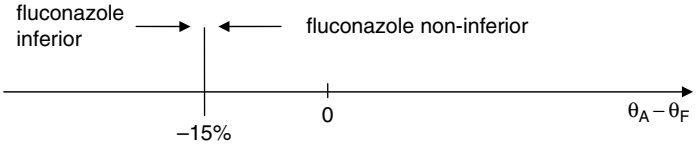


Figure 12.4 Definition of non-inferiority

12.4 A *p*-value approach

Although conventional *p*-values have no role to play in equivalence or non-inferiority trials there is a *p*-value counterpart to the confidence intervals approach. The confidence interval methodology was developed by Westlake (1981) in the context of bioequivalence and Schuirmann (1987) developed a *p*-value approach that was mathematically connected to these confidence intervals, although much more difficult to understand! It nonetheless provides a useful way of thinking, particularly when we come later to consider type I and type II errors in this context and also the sample size calculation. We will start by looking at equivalence and use $\pm\ \triangle$ to denote the equivalence margins.

Within the framework of hypothesis testing, the null and alternative hypotheses of interest for equivalence when dealing with means are as follows:

$$H_0: \ \mu_1 - \mu_2 \leq -\triangle \text{ or } \mu_1 - \mu_2 \geq \triangle$$
$$H_1: \ -\triangle < \mu_1 - \mu_2 < \triangle$$

In this case, the alternative hypothesis states that the two treatments are clinically equivalent; the null hypothesis is saying that the two treatments are not equivalent. Note that the alternative encapsulates the 'objective'; we are trying to disprove the null in order to establish equivalence.

The hypotheses stated above can be expressed as two separate sets of hypotheses corresponding to the lower and upper ends of the equivalence range:

$$H_{01}: \ \mu_1 - \mu_2 \leq -\triangle \qquad H_{11}: \ \mu_1 - \mu_2 > -\triangle$$

$$H_{02}: \mu_1 - \mu_2 \geq \Delta \qquad H_{12}: \mu_1 - \mu_2 < \Delta$$

Undertaking two tests at the 2.5 per cent level, one for H_{01} versus H_{11} and one for H_{02} versus H_{12}, can be shown to be mathematically connected to the confidence interval approach developed earlier. In particular, if each of these tests gives significant *p*-values at the 2.5 per cent significance level, then the 95 per cent confidence interval for the difference in the means will be entirely contained within the equivalence margins $\pm\Delta$. Conversely if the 95 per cent confidence interval is contained within the equivalence margins then each of the above sets of tests will give *p*-values significant at the 2.5 per cent level. The two sets of hypotheses above are both one-sided comparisons; the first set is looking to see whether the treatment difference, $\mu_1 - \mu_2$, is either $\leq$ or $> -\Delta$, while the second set is looking to see if $\mu_1 - \mu_2$ is either $\geq$ or $< \Delta$. The approach using *p*-values is therefore known as the *two, one-sided tests approach*. Following on from the earlier quote specifying the role of confidence intervals, ICH E9 states:

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Operationally, this is equivalent to the method of using two simultaneous one-sided tests to test the (composite) null hypothesis that the treatment difference is outside the equivalence margins versus the (composite) alternative hypothesis that the treatment difference is within the margins.'

With the *p*-value methodology we are rejecting the null hypothesis H_0 in favour of the alternative hypothesis H_1, providing the two (one-sided) *p*-values are $\leq$ 2.5 per cent. We have then established equivalence and we can talk in terms of the treatments being *significantly equivalent*. The terminology sounds almost contradictory, but is a correct statement. If either of the two *p*-values is above 2.5 per cent then the treatments are *not significantly equivalent*.

For non-inferiority the one-sided comparison:

$$H_0: \mu_1 - \mu_2 \leq -\Delta \qquad H_1: \mu_1 - \mu_2 > -\Delta$$

yields a *p*-value which links with the one-sided 97.5 per cent confidence interval for establishing non-inferiority. If the *p*-value from this test is significant at the 2.5 per cent level then the one-sided 97.5 per cent confidence interval will be entirely to the right of the non-inferiority margin $-\Delta$ and vice versa. If we see this outcome then we can talk in terms of the new treatment being *significantly non-inferior* to the active control. Alternatively if we get a non-significant *p*-value then the new treatment is *not significantly non-inferior*.

Using a 2.5 per cent significance level for non-inferiority in this way may initially appear to be out of line with the conventional 5 per cent significance level for superiority. A moments thought should suffice however, to realise that in a test for superiority we would never make a claim if our treatment was significantly worse than placebo, we would only ever make a claim if we were significantly better than placebo, so effectively we are conducting a one-sided test at the 2.5 per cent level to enable a positive conclusion of superiority for the active treatment.

In practice I would always recommend using confidence intervals for evaluating equivalence and non-inferiority rather than these associated *p*-values. This is because the associated *p*-values tend to get mixed up with conventional *p*-values for detecting differences. The two are not the same and are looking at quite different things. The confidence interval approach avoids this confusion and provides a technique that is easy to present and interpret.

12.5 Assay sensitivity

One concern with equivalence and non-inferiority trials is that a positive conclusion of equivalence/non-inferiority could result from an insensitive trial by default. If, for example, equivalence is established then this could mean either that the two treatments are equally effective, or indeed equally ineffective. If chosen endpoints are insensitive, dosages of the drugs too low, patients recruited who are not really ill and the trial conducted in a sloppy fashion with lots of protocol deviators and dropouts, then the treatments will inevitably look very similar! Clearly we must ensure that a conclusion of equivalence/non-inferiority from a trial is a reflection of the true properties of the treatments. The regulatory guidelines (ICH E10) talk in terms of *assay sensitivity* as a requirement of a clinical trial that ensures this.

ICH E10 (2001): 'Note for Guidance on Choice of Control Group in Clinical Trials'

'Assay sensitivity is a property of a clinical trial defined as the ability to distinguish an effective treatment from an ineffective treatment.'

Of course assay sensitivity applies in the same way to trials which evaluate superiority, but in those cases things take care of themselves. A conclusion of superiority by definition implies that the trial is a sensitive instrument, otherwise superiority would not have been detected.

Ensuring assay sensitivity is best achieved through good design and trial conduct to a high standard. From a design perspective, the current design should be in line with trials historically that have shown the active control to be effective. The following aspects should be carefully considered:

- Entry criteria; the patients should preferably be at the moderate to severe end of the disease scale.
- Dose and regimen of the active control; in line with standard practice.
- Endpoint definition and assessment; use established endpoints.
- Run-in/washout period to exclude ineligible patients; avoids diluting the patient population.

Further, the trial conduct should protect against any compromise of assay sensitivity and the following in particular should be avoided:

- poor compliance
- use of concomitant medication that could interfere with response to the treatment
- poor application of diagnostic criteria
- unblinding
- unnecessary drop-outs
- lack of objectivity and consistency in the evaluation of endpoints

It is also possible at the analysis stage to use the trial data to support assay sensitivity. Ebbutt and Frith (1998) investigate the mean change from baseline in both test treatment and active control groups and observe that the magnitude of these changes are in line with what one would expect historically in terms of the effect of the active control over placebo. In a trial where the primary endpoint is a binary outcome then, seeing a response rate in the active control group similar to response rates seen historically in placebo-controlled trials that demonstrated the active control to be an effective treatment, supports assay sensitivity. In contrast, if the response rate in the current trial is higher or lower than expected, then assay sensitivity could be drawn into question. A higher rate may indicate a population that is less severe than that used previously, while a lower rate could indicate an insensitive method of response evaluation.

Finally, one sure way to investigate assay sensitivity is to include a placebo group as a third arm in the trial. This allows direct assessment of assay sensitivity by comparing the active control with placebo where statistically significant differences would need to be seen. However, including a placebo arm will only be possible where it is ethically and practically reasonable to do so and in many equivalence/non-inferiority settings this will not be the case. A related point is that there are some therapeutic settings which are unsuitable for equivalence/non-inferiority trials unless a placebo arm is included, for example depression, anxiety and allergic rhinitis, where established effective drugs do not

consistently demonstrate effects over placebo. The inclusion of a placebo arm would allow direct evaluation of assay sensitivity.

A further phrase that is used in this area is *historical evidence of sensitivity to drug effects*. This idea, introduced initially in ICH E10, refers to the ability of effective treatments to consistently show an advantage over placebo in appropriately designed and conducted clinical trials. As mentioned in the previous paragraph, there are certain therapeutic settings where this is not the case.

12.6 Analysis sets

In superiority trials, the full analysis set is the basis for the primary analysis. As discussed in Section 7.2, the regulators prefer this approach, in part, because it gives a conservative view of the new treatment. In equivalence/non-inferiority trials, however, it is not conservative and will tend to result in the treatments looking more similar than, in reality, they are. This is because the full analysis set will include the patients who have not complied with the medication schedules and who have not followed the study procedures and the inclusion of such patients will tend to weaken treatment differences.

For equivalence and non-inferiority trials, therefore, the regulators like to see analyses undertaken on both the full analysis set and the per-protocol set with positive conclusions being drawn from both. In this sense these two analyses are considered co-primary. There is a common misconception here that for equivalence/non-inferiority trials the per-protocol set is primary. This is not the case. The per protocol set is still potentially subject to bias because of the exclusion of randomised patients and so cannot supply the complete answer; both analysis sets need to be supporting equivalence/non-inferiority in order to have a robust conclusion.

CPMP (2000): 'Points to Consider on Switching Between Superiority and Non-inferiority'

'In a non-inferiority trial, the full analysis set and the per-protocol analysis set have equal importance and their use should lead to similar conclusions for a robust interpretation.'

Similar comments apply to therapeutic equivalence trials also.

12.7 The choice of Δ

One of the most difficult aspects of the design of equivalence and non-inferiority trials, with the exception of bioequivalence, is the choice of the margin(s).

12.7.1 Bioequivalence

The equivalence margins for bioequivalence specified by both the FDA (2001) 'Statistical Approaches to Establishing Bioequivalence' and the CPMP (2001) 'Note for Guidance on the Investigation of Bioavailability and Bioequivalence' require that the ratio of the geometric means, μ_1^{GM}/μ_2^{GM}, for the two treatments lie between 0.80 and 1.25. This requirement applies to both AUC and C_{max}. A deviation from the rules for therapeutic equivalence is that 90 per cent confidence intervals are used rather than 95 per cent leading to a more relaxed requirement.

The reason why ratios of geometric means are used in this context is as discussed in Section 11.4; the distributions of AUC and C_{max} tend to be positively skewed and the log transformation is applied to recover normality.

By taking logs, the above condition, $0.8 < \frac{\mu_1^{GM}}{\mu_2^{GM}} < 1.25$, can be translated into the following requirement for $\mu_1^* - \mu_2^*$ where μ_1^* is the mean AUC on the log scale for the test treatment and μ_2^* is the mean AUC on the log scale for the active control:

$$\ln(0.80) < \mu_1^* - \mu_2^* < \ln(1.25)$$

$$\text{or equivalently} \qquad -0.22 < \mu_1^* - \mu_2^* < 0.22$$

12.7.2 Therapeutic equivalence

The rules for therapeutic equivalence are different from those for bioequivalence. The choice of margin will be a mixture of statistical and clinical reasoning. Strict equivalence is appropriate when we want to consider essential similarity or where the test treatment is to be used as an exact replacement for the new treatment. In these cases Δ should be chosen to be a completely irrelevant difference from a clinical point of view. Ebbutt and Frith (1998) based their choice of $\Delta = 15$ l/min on several considerations:

- Previous trials with the beta-agonist salmeterol had given an average difference from placebo in PEF of 37 l/min while the effect of inhaled steroids was of the order of 25 l/min and Δ was chosen as a proportion of those effects.
- Typically, a mean improvement of 70 l/min would be seen following treatment with a short acting beta-agonist and Δ was chosen to be about 20 per cent of this.
- Discussion with practitioners suggested that 15 l/min was clinically irrelevant.

In the next section we will discuss the non-inferiority setting and many of the considerations there also apply to the situation of equivalence.

12.7.3 Non-inferiority

In the context of using a non-inferiority trial to demonstrate that a test treatment is efficacious, the following provides a statistical approach for the choice of $\triangle$. Consider, in the setting of hypertension, the trials that historically compared the active control with placebo. In a meta-analysis (see Chapter 15) suppose that the 95 per cent confidence interval for the active control treatment effect in terms of the fall in diastolic blood pressure was (4.5 mmHg, 10.3 mmHg). Clearly $\triangle$ would need to be chosen to be considerably less than 4.5 mmHg, otherwise we could find ourselves simply developing another placebo! Defining $\triangle$ to be one half ($= 2.25$ mmHg) or one third ($= 1.50$ mmHg) of the lower bound of this confidence interval would give us statistical confidence coming out of our non-inferiority trial with a positive result, that the test treatment is at worst either 2.25 mmHg or 1.50 mmHg less efficacious than the active control and that the test treatment still maintains a clear advantage over placebo. This value for Δ would then need to be additionally justified on clinical grounds, that is as an irrelevant difference clinically.

Further, the meta-analysis has produced a confidence interval for the reference product against placebo while the non-inferiority trial will give a confidence interval for the test product against the reference product. It is then possible, by combining the standard errors, to indirectly produce a confidence interval for the test product against placebo and this interval could then be used to judge the clinical importance of the test treatment effect as follows. If $\bar{x}_R^* - \bar{x}_P^* = 7.4$ mmHg represents the treatment difference between the reference product mean and the placebo mean in the meta-analysis while $\bar{x}_T - \bar{x}_R = -0.6$ mmHg equals the observed difference in the test and reference treatments in the non-inferiority trial, then assuming that the conditions of the trials are similar (termed constancy – see section 12.7.5) an estimated difference between test treatment and placebo means is, by adding these two effects together, equal to 6.8 mmHg. The standard error attached to each of these differences, say se_1 and se_2 can be combined to give a standard error (se) for $\bar{x}_T - \bar{x}_R$ using the formula $se = \sqrt{se_1^2 + se_2^2}$. The 95 per cent confidence interval for the true test treatment effect is then given approximately by $\bar{x}_T - \bar{x}_R \pm 2 \times se$.

In situations where we are trying to show no important loss of efficacy of a test treatment to an active control, it is not possible to be entirely prescriptive about methods for the choice of $\triangle$. For example, if a new treatment provides an advantage over the existing treatment in terms of safety, the price we are prepared to pay for this in terms of efficacy will clearly depend on the extent of the safety advantage. In these cases it is not appropriate to think in terms of preserving a proportion of the effect of the active control over placebo.

Also if the active control effect over placebo is large, then preserving a proportion in this sense does not fit with the objectives of the non-inferiority evaluation:

CHMP (2005): 'Guideline on the Choice of Non-Inferiority Margin'

'The choice of delta for such an objective cannot be obtained by looking only at past trials of the comparator against placebo. Ideas such as choosing delta to be a percentage of the expected difference between active and placebo have been advocated, but this is not considered an acceptable justification for the choice . . . To adequately choose delta an informed decision must be taken, supported by evidence of what is considered an unimportant difference in the particular disease area.'

As the regulators point out, the choice of $\triangle$ where we are looking to demonstrate no important loss of efficacy needs to be based on clinical reasoning.

12.7.4 The 10 per cent rule for cure rates

It has been relatively common practice, historically, to use a $\triangle$ of 10 per cent for cure rates.

CPMP (1999): 'Note for Guidance on Clinical Evaluation of New Vaccines'

'In individual trials, $\triangle$ can often be set to about 10 per cent, but will need to be smaller for very high protection rates'

CPMP (2003): 'Note for Guidance on Evaluation of Medicinal Products Indicated for Treatments of Bacterial Infections'

'In most studies with antibacterial agents in common indications this ($\triangle$) should likely be 10 per cent, but may be smaller when very high cure rates are anticipated.'

The message here appears consistent; 10 per cent is likely to be acceptable except when rates are high, say > 90 per cent and although these regulatory guidelines are from Europe, the FDA position has been similar. In more recent times, however, the regulators have been happy to deviate from this 10 per cent, in both directions! In a rare disease in which only one or a small number of treatments currently exist, the regulators may be willing to relax the 10 per cent to 12.5 per cent or even 15 per cent. In contrast, for common diseases the regulators may suggest a tighter $\triangle$ arguing that in the interests of public health, the new treatment would only be acceptable if its performance was very close to the active control.

There are one or two other specific therapeutic settings where there is more guidance on the choice of $\triangle$. For example:

CPMP (2003): 'Points to Consider on the Clinical Development of Fibrinolytic Medicinal Products in the Treatment of Patients with ST Segment Evaluation Acute Myocardial Infarction (STEMI)'

'Mortality: In the recent past differences of 14 per cent relative or 1 per cent absolute (whichever proves smallest) have been accepted. These margins were based on 'all cause mortality' rates at day 30 close to 6.5–7 per cent.'

One final point on the choice of delta which is relevant in all therapeutic settings, and certainly in relation to cure rates, is that the regulators could very well change their minds about what is and what is not acceptable for $\triangle$ if the performance of the active control in the trial deviates from what was expected. For example, suppose that a $\triangle = 10$ per cent was chosen with an expected cure rate of 85 per cent for the active control. If in the trial the cure rate on the active control turns out to be 93 per cent then the regulators may view 10 per cent as too large a value and may suggest a reduction to say 5 per cent for this particular trial.

12.7.5 Biocreep and constancy

One valid concern that regulators have is the issue of so-called *biocreep*. Demonstrating that a second generation active treatment is non-inferior to the active control may well mean that the new treatment is slightly less efficacious that the active control. Evaluating a third generation active to the now established second generation active may lead to a further erosion of efficacy and so on, until at some stage a new active, whilst satisfying the 'local' conditions for non-inferiority, is, in reality, indistinguishable from placebo. The FDA discussed this issue many years ago specifically in relation to anti-infectives (FDA (1992) 'Points to Consider on Clinical Development and Labeling of Anti-Infective Drug Products').

The issue of *constancy* concerns the conditions under which the current active control trial is being conducted compared to the conditions under which the active control was established historically. Things may well have changed. For example, the nature of the underlying disease or the effectiveness of ancillary care may be such that the active control performs rather differently now than it did when the original placebo-controlled trials were undertaken. This may well be true, for example, for antibiotics where populations of patients will have developed resistance to certain treatments. If this were the case then the current non-inferiority trial could lead to a misleading conclusion of effectiveness for the new active when in fact the comparator treatment is ineffective.

Both of these elements are causing nervousness amongst regulators. So much so that, for example, the FDA Anti-Infective Drugs Advisory Committee (AIDAC) have recently recommended that the non-inferiority design should no longer

be used in trials for acute bacterial sinusitis (CDER Meeting Documents; Anti-Infective Drugs Advisory Committee (October 29, 2003), www.fda.gov).

12.8 Sample size calculations

We will focus our attention to the situation of non-inferiority. Within the testing framework the type I error in this case is as before, the false positive (rejecting the null hypothesis when it is true), which now translates into concluding non-inferiority when the new treatment is in fact inferior. The type II error is the false negative (failing to reject the null hypothesis when it is false) and this translates into failing to conclude non-inferiority when the new treatment truly is non-inferior. The sample size calculations below relate to the evaluation of non-inferiority when using either the confidence interval method or the alternative *p*-value approach; recall these are mathematically the same.

The sample size calculation requires pre-specification of the following quantities:

- Type I error, which will usually be set at 2.5 per cent
- Power = I – type II error, which would be 80 per cent or 90 per cent in most cases
- $\triangle$, the non-inferiority margin

The remaining quantities would depend on the primary endpoint and the design; assume we are dealing with the parallel group case.

For a continuous endpoint we would need:

- The standard deviation of the endpoint
- The anticipated true difference in the two mean values

For a binary endpoint we would need:

- The response rate in the active control group
- The anticipated true difference in the response rates

Usually we conduct these calculations assuming no difference between the treatments in terms of means (or rates), but this is not always a realistic assumption. It is good practice to at least look at the sensitivity of the calculation to departures from this assumption.

Example 12.2: Evaluating non-inferiority for cure rates

In an anti-infective non-inferiority study it is expected that the true cure rates for both the test treatment and the active control will be 75 per cent. Δ has been chosen to be equal to 15 per cent. Using the usual approach with a one-sided 97.5 per cent confidence interval for the difference in cure rates a total of 176 patients per group will give 90 per cent power to demonstrate non-inferiority. Table 12.1 gives values for the sample size per group for 90 per cent power and for various departures from the assumptions.

Table 12.1 Sample sizes per group

		Cure rate (test treatment)			
		65%	70%	75%	80%
Cure rate (active control)	65%	213	115	70	46
	70%	460	197	105	63
	75%	1745	418	176	91
	80%	∞	1556	366	150

When the cure rates are equal, the sample size decreases as the common cure rate increases. When the test treatment cure rate is above the active control cure rate then the test treatment is actually better than the active control and it is much easier to demonstrate non-inferiority. When the reverse happens, however, where the test treatment cure rate falls below that of the active control, then the sample size requirement goes up; it is much more difficult under these circumstances to demonstrate non-inferiority. It is also worth noting that when the test treatment is truly 15 per cent, the value for Δ in this example, below the rate in the active control, demonstrating non-inferiority is simply not possible.

As with sample size in superiority trials we generally power on the basis of the per-protocol set and increase the sample size to account for the non-evaluable patients. This is particularly important in non-inferiority trials where the full analysis set and the per-protocol set are co-primary analyses. Note also, as before in superiority trials further factoring up may be needed if there are randomised patients who are being systematically excluded from the full analysis set, as is the case, for example, in anti-infective trials.

There is a perception that non-inferiority trials are inevitably larger than their superiority counterparts. Under some circumstances this is true, but is by no

means always the case. One crucial quantity in the sample size calculation for a non-inferiority trial is $\triangle$, which plays a role similar to the crd in a superiority sample size calculation. The sample size (this is also true for equivalence) is inversely proportional to the square of $\triangle$. If $\triangle$ is small then the sample size will be large, and the constraints placed upon us by regulators together with the clinical interpretation of 'irrelevant differences' tend to make $\triangle$ small in such trials. In a superiority trial, the choice of the clinically relevant difference to detect is an internal, clinical, sometimes commercial, decision that is under the control of the trialists and we are at liberty to power a trial on the basis of a large value. The net effect of these considerations is that non-inferiority trials tend to be larger than trials designed to demonstrate superiority. However, and in contrast to this, if we truly feel that the test treatment is, in reality, somewhat better than the active control then assuming such a positive advantage can have the effect of considerably reducing the sample size, as seen in the above example.

12.9 Switching between non-inferiority and superiority

In a clinical trial with the objective of demonstrating non-inferiority, suppose that the data are somewhat stronger than this and the 95 per cent confidence interval is not only entirely to the right of $-\triangle$, but also completely to the right of zero as in Figure 12.5; there is evidence that the new treatment is in fact superior.

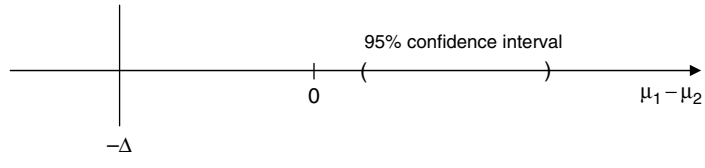


Figure 12.5 Concluding superiority in a non-inferiority trial

What conclusions can we draw in this situation? Can we conclude superiority even though this was not the original objective? Well generally the answer is yes, superiority can be claimed.

CPMP (2000): 'Points to Consider on Switching Between Superiority and Non-Inferiority'

'If the 95 per cent confidence interval for the treatment effect not only lies entirely above $-\triangle$ but also above zero, then there is evidence of superiority in terms of statistical significance at the 5 per cent level ($p < 0.05$). In this case, it is acceptable to calculate the exact probability associated with a test of superiority and to evaluate whether this is sufficiently small to reject convincingly the hypothesis of no differenc . . . Usually this demonstration of a benefit is sufficient for licensing on its own, provided the safety profiles of the new agent and the comparator are similar.'

There are no multiplicity arguments that impact on this switch. Essentially we can think of the two 'tests', a test of non-inferiority followed by a test of superiority, as a hierarchy and we will only be considering the second providing the first one gives a significant result. In fact the safest thing to do in this setting is to pre-specify a hierarchy with non-inferiority followed by superiority in the protocol; there can then be no dispute about such a switch in the eyes of regulators.

Following calculation of the exact *p*-value for superiority the 95 per cent confidence interval allows the clinical relevance of the finding to be evaluated. Presumably, however, any level of benefit would be of value given that at the outset we were looking only to demonstrate non-inferiority.

For superiority the full analysis set is usually the basis for the primary analysis so the emphasis in the superiority claim would then need to be based around this.

Example 12.1 (continued): Fluconazole compared to amphotericin B in preventing relapse in cryptococcal meningitis

This example has been presented previously in this chapter. The initial objective was to demonstrate the non-inferiority of fluconazole compared to amphotericin B in the prevention of cryptococcal meningitis in patients with AIDS. The non-inferiority margin was set at −15 per cent for the difference in the relapse rates.

The 95 per cent confidence interval for $\theta_A - \theta_F$ was in fact (7 per cent, 31 per cent) and this is entirely to the right, not only of −15 per cent, but also of zero. In this case a claim for the superiority of fluconazole is supported by the data. The authors concluded that non-inferiority had been established, but additionally there was evidence that fluconazole was more effective than amphotericin B:

'These data allow us to conclude that fluconazole was at least as effective as weekly amphotericin B . . . Indeed, the 19 per cent difference in the probability of relapse at one year . . . suggests that fluconazole was more effective than amphotericin B in preventing a relapse of crytococcal disease in this population of patients.'

Moving in the opposite direction, that is concluding non-inferiority in a superiority trial, is much more difficult, as this would generally require pre-specification of a non-inferiority margin. Such pre-specification would usually not have been considered in a trial designed to demonstrate superiority. However, if a conclusion of non-inferiority would be a useful outcome then it could be appropriate to consider such pre-specification. Switching from superiority to non-inferiority, however, presents further problems. Assay sensitivity may well be one reason why

the trial, initially designed to detect superiority, has not done so. The design would need to be very robust and the data compelling if assay sensitivity were to be upheld.

Given the various possibilities regarding switching there may well be a strong argument in any active control comparison to always go for non-inferiority. If the data then turn out to be stronger and support superiority then this additional conclusion can be made. There are some drawbacks, however, with this way of thinking:

- Non-inferiority trials are more difficult to design; assay sensitivity and the choice of non-inferiority margin are just two of the issues that would additionally need to be considered.
- Non-inferiority trials can often require large sample sizes.
- Designing the trial as a non-inferiority evaluation may give a negative perception within and outside of the clinical trial team.

Nonetheless this may be a strategy worth considering under some circumstances.

13

The analysis of survival data

13.1 Time-to-event data and censoring

In many cases an endpoint directly measures time from the point of randomisation to some well-defined event, for example time to death (survival time) in oncology or time to rash healing in Herpes Zoster. The data from such an endpoint invariably has a special feature, known as censoring. For example, suppose the times to death for a group of patients on a particular treatment in a 24 month oncology study are as follows:

$$14\ 7\ 24^*\ 15\ 3\ 18\ 9^*\ 10\ 24^*\ 9\ldots$$

Here the first patient died after 14 months from the time of randomisation, the second patient after 7 months. The third patient however is still alive at the end of the study while patients 4, 5 and 6 died after 15, 3 and 18 months respectively. Patient 7 was lost to follow-up at 9 months and patient 8 died after 10 months. Patient 9 is also still alive at the end of the trial while finally patient 10 died after 9 months. As can be seen, the primary endpoint, survival time, is not available for all of the patients. It is not that we have no information on patients 3, 7 and 9 but we do not have complete information, we know only that their survival times are at least 24, 9 and 24 months, respectively. These patients provide what we call *censored observations*. Unfortunately we cannot even do some simple things. For example, it is not possible to calculate the mean survival time. You might say, well can't we just ignore the fact that these observations are censored and calculate the average of the numerical values? Well you could, but clearly this would give an underestimate of the true mean survival time since eventually the actual survival times for patients 3, 7 and 9 will be greater than the numerical values in the list. Can't we just ignore the censored values and calculate the mean of those that remain? Again this calculation would give an underestimate of the

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

true mean, the censored values tend to come from the patients who survive a long time and ignoring then would systematically remove the patients that do well.

It is this specific feature that has led to the development of special methods to deal with data of this kind. If censoring were not present then we would probably just takes logs of the patient survival times and undertake the unpaired t-test or its extension ANCOVA to compare our treatments. Note that the survival times, by definition, are always positive and frequently the distribution is positively skewed so taking logs would often be successful in recovering normality.

The special methods we are going to discuss in this section were first developed primarily in the 1970s and applied in the context of analysing time to death and this is why we generally refer to the topic as 'survival analysis'. As time has gone on, however, we have applied these same techniques to a wide range of time-to-event type endpoints. The list below gives some examples:

- Time to rash healing in Herpes Zoster
- Time to complete cessation of pain in Herpes Zoster
- Disease-free survival
- Time to first seizure in epilepsy
- Time to alleviation of symptoms in flu
- Time to cure for an infectious disease

Throughout this section we will adopt the conventions of the area and refer to survival analysis and survival curves, accepting that the methods are applied more widely to events other then death.

Censoring in clinical trials usually occurs because the patient is still alive at the end of the period of follow-up. In the above example, if this were the only cause of censoring then all the censored observations would be equal to 24 months. There are, however, other ways in which censoring can occur, such as lost to follow-up or withdrawal. These can sometimes raise difficulties and we will return to discuss the issues in a later section. Also, at an interim analysis the period of follow-up for the patients still alive in the trial would be variable and this would produce a whole range of censored event times; our methodology needs to be able to cope with this.

In the next section we will discuss Kaplan–Meier curves, which are used both to display the data and also to enable the calculation of summary statistics. We will then cover the logrank and Gehan–Wilcoxon tests which are simple two group comparisons for censored survival data (akin to the unpaired t-test), and then extend these ideas to incorporate centre effects and also allow the inclusion of baseline covariates.

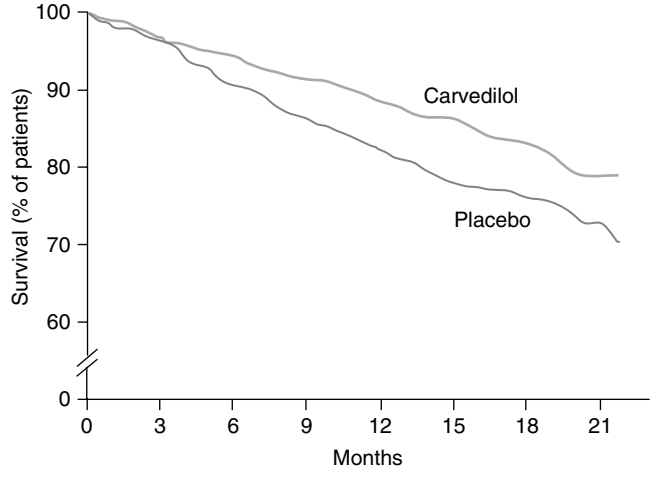
13.2 Kaplan–Meier (KM) curves

13.2.1 Plotting KM curves

Kaplan and Meier (1958) introduced a methodology for estimating, from censored survival data, the probability of being event-free as a function of time. If the event is death then we are estimating the probability of surviving and the resultant plots of the estimated probability of surviving as a function of time are called either *Kaplan–Meier (KM) curves* or *survival curves*.

Example 13.1: Carvedilol in severe heart failure

A placebo-controlled randomised trial reported by Packer *et al.* (2001) investigated the effect of carvedilol on survival time in severe heart failure. Figure 13.1 shows the survival curve for each of the two treatment groups following the early termination of the trial at a planned interim analysis.



No. of Patients at Risk

Placebo	1133	937	703	580	446	286	183	114
Carvedilol	1156	947	733	620	479	321	208	142

Figure 13.1 Kaplan–Meier survival curves in placebo and carvedilol groups (Packer M, Coats AJS, Fowler MB, *et al.* for the Carvedilol Prospective Randomised Cumulative Survival Study Group, 'Effect of carvedilol on survival in severe chronic heart failure', *New England Journal of Medicine*, **344**, 1651–1658. © (2001) Massachusetts Medical Society.)

The Kaplan–Meier method looks at the patterns of deaths over time to estimate the probability of surviving. The censored values make a contribution to this estimation process; if a patient is censored at 12 months, then that patient is involved in estimating the probability of surviving up to and including 12 months, but not beyond. The form of the estimated survival curves is a so-called step function, with steps down occurring at those time points where there are deaths. As patients die or are censored the number of patients remaining alive in the trial in each of the treatment groups is diminishing. The probabilities of surviving are, as a consequence, estimated from fewer and fewer patients as time progresses and once these groups of patients become small, the estimated curves become a little unstable. That is why generally you will see more variability at the longer follow-up times. To give information in relation to this it is common practice to record the *number of patients at risk* at various time points; these are the numbers of patients alive and in the trial at those time points. In the Packer *et al.* (2001) study there were 1133 patients randomised to placebo and 1156 patients randomised to carvedilol; at 12 months following randomisation there were 446 patients at risk in the placebo group compared to 479 in the carvedilol group. By 21 months the risk sets comprised 114 patients in the placebo group and 142 in the carvedilol group.

13.2.2 Event rates and relative risk

It is straightforward to obtain the estimated probability of surviving for various key time points from the Kaplan–Meier estimates. In the Packer *et al.* (2001) example, the estimated survival probability at 12 months in the carvedilol group was 0.886 compared to 0.815 in the placebo group, an absolute difference of 7.1 per cent in the survival rates. A standard error formula provided by Greenwood (1926) enables us to obtain confidence intervals for these individual survival rates and for their differences.

The estimated risk of dying in the first 12 months is then 0.114 in the cavedilol group compared to 0.185 in the placebo group. This enables the calculation of a relative risk at 12 months as $0.114/0.185 = 0.62$ and the relative risk reduction is 38 per cent. Similar calculations can be undertaken at other time points.

13.2.3 Median event times

We have mentioned earlier in this chapter that it is not possible to calculate the mean survival time. It is, however, usually possible to obtain median survival times from the Kaplan–Meier curves. The median survival time for a particular

group corresponds to the time on the x-axis when the survival probability on the curve takes the value 0.5. In order for this statistic to be obtained, the survival curves must fall below the 0.5 value on the y-axis and for the example above this has not happened. In such cases we use the survival rates as summary descriptions of the survival experience in the groups. We will see a further example later where the curves do fall below the 0.5 point. When the median times are available it is also possible to obtain associated standard errors and confidence intervals.

It is usual to estimate and plot the probability of being event-free, but there will be occasions when interpretation is clearer when the opposite of this, cumulative incidence (or cumulative probability of experiencing the event by that time), is plotted. This is simply obtained as 1 – probability of being event-free. Pocock *et al.* (2002) discuss issues associated with the interpretation of these plots. These authors point out that interpretation in the conventional type of plot, when the event rates are low, can be exaggerated visually by a break in the y-axis, so take care!

13.3 Treatment comparisons

The Kaplan–Meier curves do not of themselves provide a formal, p-value, comparison of the treatments. This comparison of the survival curves is undertaken using either the logrank test or the Gehan–Wilcoxon test. We will look at these two test procedures in turn.

The *logrank test* was developed by Peto and Peto (1972). The p-value resulting from this test is not a comparison of the survival curves at a particular time point (although such a test could be constructed), but a test comparing the two complete curves. A statistically significant p-value indicates that the survival experience in the two groups is different. The *Gehan–Wilcoxon test* (Gehan (1969)) (sometimes referred to as the *generalised Wilcoxon test*) is an alternative procedure for producing a p-value comparison of two survival curves. Why do we need two different tests? Well the reasons relate to the basic shape of the curves that we are comparing. Both tests provide valid p-values, but these two tests are designed in different ways to pick up different patterns of treatment differences.

Figure 13.2 provides two sets of hypothetical population survival curves. In part a) the curves are seen to separate out gradually over the period of follow-up. In contrast, the curves in part b) separate out fairly rapidly, but then start to converge later on in time.

The curves in part a) represent, over the period of follow-up, a 'permanent' treatment effect. There is a long-term advantage in survival for one group

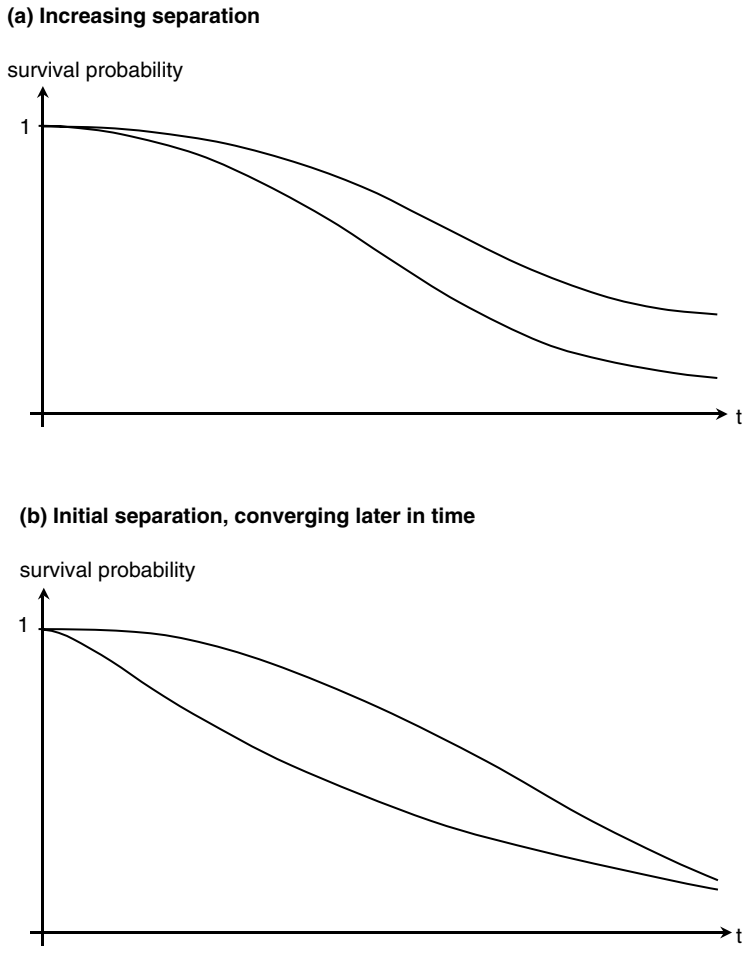


Figure 13.2 Patterns of survival curves

compared to the other group. The interpretation of the curves in part b), however, is different. Here we see a short-term benefit of one treatment group over the other, but as time goes on this relative benefit diminishes leaving little long-term effect. This pattern of differences represents a delay in the occurrence of the event in one group compared to the other group.

The two tests mentioned earlier are designed to look for these different patterns. The logrank is best able to pick up the longer-term differences as seen in Figure 13.2 a) while the Gehan–Wilcoxon test focuses on short-term effects or a delay as seen in Figure 13.2 b). The appropriate test to use depends upon what kind of differences you are expecting to see.

In the Packer *et al.* (2001) trial in Figure 13.1 we see longer-term differences displayed over the 21 month follow-up period and the logrank test is entirely appropriate here. The p-value from the logrank test was 0.00013, a highly significant result.

Generally, it is these kinds of patterns we see in long-term cardiovascular and oncology trials. In other applications, however, such long-term effects are not anticipated and indeed do not fit with the objectives of the study. For example, in flu trials, where the primary endpoint is the time to alleviation of symptoms, we are looking for rapid resolution of symptoms in the active group compared to placebo. By the time we get to eight days, most of the patients in each of the two groups have only minor symptoms remaining, so the probability of being event-free (still having symptoms) is the same in the two groups. In this case it is the Gehan–Wilcoxon test that is best suited to picking up differences. Similar comments often apply to disease-free survival in advanced cancer trials where the best we can hope for is that the test treatment delays the event (death or disease recurrence). It is unfortunately the case that the logrank test dominates this area and there are many applications that have failed to detect important short-term differences between 'survival' curves as a result of using a test that is insensitive to the detection of those differences. One of many examples is Okwera *et al.* (1994) who compared two treatments (thiacetazone and rifampicin) for pulmonary tuberculosis in HIV-infected patients. The survival curves at 300 days following randomisation were separated by more than 10 per cent (77 per cent surviving in one group compared to around 66 per cent in the second group). In other words, 10 per cent more patients, at least, were alive at 300 days in the rifampicin arm compared to the thiacetazone arm, a clear and important benefit of one treatment over the other. At 600 days, however, the survival curves had come together showing no longer-term benefit. The quoted p-value from the logrank test was > 0.50. Now one could argue whether a short-term benefit is clinically important, but the issue here is that the test used was not sensitive in terms of picking up differences between the curves. The Gehan–Wilcoxon test would have stood a much better chance of yielding statistical significance and at least generated some interest in discussing the implication of those short-term differences.

One question that often arises is; which test should I use as I don't know what kind of effect I am going to see? My short answer to this question is that in a confirmatory setting you should know! By the time you reach that stage in the drug development programme your knowledge of the disease area and the treatment, in combination with the endpoint, should enable accurate prediction of what should happen. Of course, earlier on in the programme you may not know and in this exploratory phase it is perfectly valid to undertake both tests to explore the nature of the effects.

13.4 The hazard ratio

13.4.1 The hazard rate

In order to be able to understand what a hazard ratio is, you first need to know what a hazard rate is. The *hazard rate* (function) is formally defined as the conditional death (or event) rate calculated through time. What we mean by this is as follows. Suppose in a group of 1000 patients in month 1, 7 die; the hazard rate for month 1 is 7/1000. Now suppose that 12 die in month 2; the hazard rate for month 2 is 12/993. If now 15 die in month 3 then the hazard rate for month 3 is 15/981 and so on. So the hazard rate is the death (event) rate for that time period amongst those patients still alive at the start of the period.

There are several things to note about the hazard rate. Firstly, it is unlikely that the hazard rate will be constant over time. Secondly, even though we have introduced the concept of the hazard rate as taking values for monthly periods of time, we can think in terms of small time periods with the hazard rate being a continuous function through time.

The hazard rate can be estimated from data by looking at the patterns of deaths (events) over time. This estimation process takes account of the censored values in ways similar to the way such observations were used in the Kaplan–Meier curves.

Figure 13.3 shows a schematic plot of two hazard rates corresponding to two treatment groups in a randomised trial. As can be seen from this plot, the hazard rates in each of the two treatment groups start off just after randomisation ($t = 0$) at a fairly modest level and then increase to a peak after a certain period of time, say one year, and then begin to decrease. This is telling us that at one year the death rates in each of the groups are at their maximum; prior to that they have been steadily increasing and following one year the death rates are tending

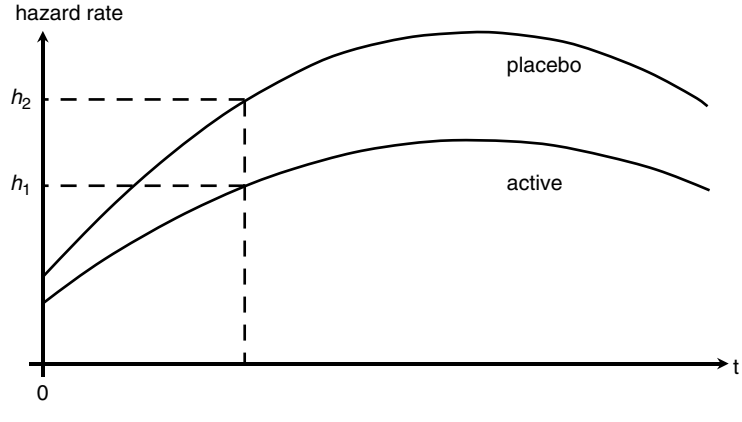


Figure 13.3 Hazard rates for two groups of patients

to tail off. It is also clear from the plot that the death rate in the placebo group is always higher than the death rate in the active group.

13.4.2 Constant hazard ratio

Even though the individual hazard rates seen in Figure 13.3 are not constant, it would be reasonable to assume, wherever we look in time, that the ratio of the hazard rates is approximately constant. In fact, these hazard rates have been specifically constructed to behave in this way. When this is the case, the ratio of the hazard rates will be a single value, which we call the *hazard ratio*. We will denote this ratio by λ so that $\lambda = h_1/h_2$.

It is convention for the hazard rate for the test treatment group to appear as the numerator and the hazard rate for the control group to be the denominator.

A hazard ratio of one corresponds to exactly equal treatments; the hazard rate in the active group is exactly equal to the hazard rate in the placebo group. If we adopt the above convention and the event is death (or any other undesirable outcome) then a hazard ratio less than one is telling us that the active treatment is a better treatment. This is the situation we see in Figure 13.3. A hazard ratio greater than one is telling us that the active treatment is a poorer treatment.

Even if the hazard ratio is not precisely a constant value as we move through time, the hazard ratio can still provide a valid summary provided the hazard rate for one of the treatment groups is always above the hazard rate for the other group. In this case the value we get for the hazard ratio from the data represents an average of that ratio over time.

Confidence intervals for the hazard ratio are straightforward to calculate. Like the odds ratio (see Section 4.5.5), this confidence interval is firstly calculated on the log scale and then converted back to the hazard ratio scale by taking anti-logs of the ends of that confidence interval.

13.4.3 Non-constant hazard ratio

However, it is not always the case, by any means, that we see a constant or approximately constant hazard ratio. There will be situations, as seen in Figure 13.4, when the hazard rate for one group starts off lower than the hazard rate for a second group and then as we move through time they initially move closer together, but then a switch occurs. The hazard rate for the first group then overtakes that for the second group and they continue to move further apart from that point on.

In this case it clearly makes no sense to assign a single value to the hazard ratio. The hazard ratio in this case will start off below one, say, increase towards one as

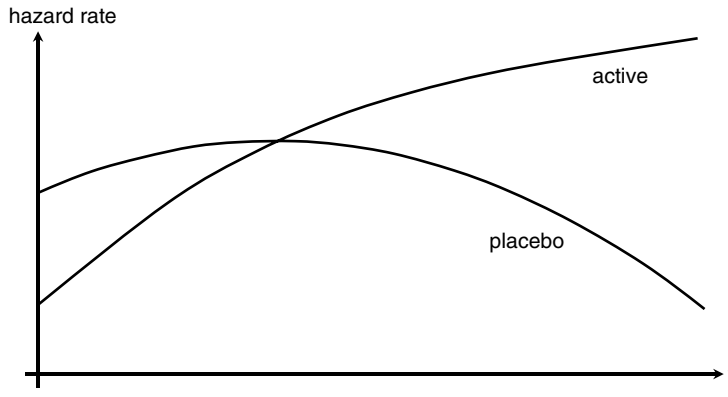


Figure 13.4 Hazard rates for two groups of patients where the hazard ratio is not constant

we move through time, but then flip over and start to increase above and away from one. The hazard ratio still exists, but it is not constant and varies over time. See Kay (2004) for further discussion on these points.

13.4.4 Link to survival curves

In an earlier section we saw two different patterns for two sets of survival curves. In Figure 13.2 a) the survival curves move further and further apart as time moves on. This pattern is consistent with one of the hazard rates (think in terms of death rates) being consistently above the other hazard rate. This in turn corresponds to a fairly constant hazard ratio, the situation we discussed in Section 13.4.1. So a constant hazard ratio manifests itself as a continuing separation in the two survival curves as in Figure 13.2 a). Note that the higher hazard rate (more deaths) gives the lower of the two survival curves.

In Figure 13.2 b) the survival curves move apart very early on in time and then start to converge later on in time. This pattern is consistent with one of the hazard rates being above the other initially; this is the survival curve that falls most rapidly due to the high death rate early on. In order for the survival curves to begin to converge, however, a reversal of the death rates in the two groups needs to take place, so that the death rate in the group that did well initially now starts to increase. This is the only way that the catch up can take place, to ensure that the probability of surviving beyond the end of the observation period is approximately equal in the two groups. In turn this pattern is consistent with a hazard ratio that is not constant. Here the hazard ratio starts off well below one, increases to one and then moves above one through time as the death rates in the two groups reverse.

For the pattern of survival curves in Figure 13.2 b), it makes no sense to calculate the hazard ratio because it is not constant, it does not take a single value. A better way to summarise the performance in the two groups would be to use median survival times or relative risk at particular time points of interest. We mentioned in Section 13.3 that the logrank is specifically designed to provide a *p*-value for a pattern of survival curves as in Figure 13.2 a). As we have now seen, this pattern corresponds to a constant hazard ratio; indeed the logrank test is essentially a test of the hypothesis $H_0 : \lambda = 1$ against the alternative $H_1 : \lambda \neq 1$.

13.4.5 Calculating KM curves

It is worth revisiting the calculation of the Kaplan–Meier curve following on from the discussion of the hazard rate, in order to see, firstly, how censoring is accounted for, and secondly, how the two are linked in terms of the calculation.

As in Section 13.4.1, consider a group of 1000 patients at the point of randomisation. Suppose in month 1, 7 die, then the hazard rate for month 1, as before, is 7/1000. Now assume that in addition 10 are censored at the end of month 1 while amongst the 983 (= 1000 − 7(deaths) −10 (censorings)) remaining in the study at the beginning of month 2, 15 die in month 2; the hazard rate for month 2 is 15/983. Next assume that 8 patients are censored at the end of month 2, so that 960 patients are in the study at the beginning of month 3, and that of these 12 die in that month; the hazard rate for month 3 is 12/960 and so on. The calculation of the hazard rate through time takes account, not only of the number of patients dying, but also of the censoring patterns. Patients with censored observations contribute to the denominator of the hazard rate calculation in the months prior to the time at which they are censored.

Now to the calculation of the Kaplan–Meier survival probabilities. Using the above example, the estimated probability of surviving beyond month 1 is $1 - (7/1000) = 0.9993$. The probability of surviving beyond month 2 is the probability of surviving beyond month 1× the probability of surviving beyond month 2 amongst those alive at the start of month 2, and from the data this is estimated to be $0.9993 \times (1 - (15/983)) = 0.984$. Continuing this process the probability of surviving beyond month 3 is the probability of surviving beyond month 2× the probability of surviving beyond month 3 amongst those alive at the start of month 3; this is estimated by $0.984 \times (1 - (12/960)) = 0.972$. In general then if h_t is the hazard rate for month t then the estimated probability of surviving beyond month t is equal to:

$$(1 - h_1) \times (1 - h_2) \times \ldots \times (1 - h_t)$$

Note also here that the numbers 1000, 983 and 960 are the numbers of patients alive in the trial, in that group, at the start of months 1, 2 and 3 respectively, the risk sets.

This calculation in relation to both the hazard rate and the survival probabilities has been undertaken at intervals of one month. In practice we use intervals which correspond to the unit of measurement for the endpoint itself, usually days, in order to use the total amount of information available.

13.5 Adjusted analyses

In Chapter 6 we covered methods for adjusted analyses and analysis of covariance in relation to continuous (ANOVA and ANCOVA) and binary and ordinal data (CMH tests and logistic regression). Similar methods exist for survival data. As with these earlier methods, particularly in relation to binary and ordinal data, there are numerous advantages in accounting for such factors in the analysis. If the randomisation has been stratified, then such factors should be incorporated into the analysis in order to preserve the properties of the resultant *p*-values.

13.5.1 Stratified methods

Both the logrank and Gehan–Wilcoxon tests can be extended to incorporate stratification factors, for example, centres and baseline covariates. These methods provide *p*-value comparisons of treatments allowing for the main effects of centre and baseline covariates. Although possible, extensions of these procedures to evaluate the homogeneity of the treatment effect, that is, the investigation of treatment-by-centre or treatment-by-covariate interactions, is not so straightforward. Consequently, we tend to build the covariates into the modelling through analysis of covariance methods that will be covered in the next two sections.

13.5.2 Proportional hazards regression

The most popular method for analysis of covariance is the *proportional hazards model*. This model, originally developed by Cox (1972), is now used extensively in the analysis of survival data to incorporate and adjust for both centre and covariate effects. The model assumes that the hazard ratio for the treatment effect is constant.

The method provides a model for the hazard function. As in Section 6.6, let z be an indicator variable for treatment taking the value one for patients in the active group and zero for patients in the control group and let x_1, x_2, etc. denote the covariates. If we let $\lambda(t)$ denote the hazard rate as a function of t (time), the main effects model takes the form:

$$\ell n\,(\lambda(t)) = a + cz + b_1x_1 + b_2x_2 + \cdots$$

As before, the coefficient c measures the effect of treatment on the hazard rate. If $c < 0$ then the log hazard rate, and therefore the hazard rate itself, in the active group is lower than the hazard rate in the control group. If $c > 0$ then the reverse

Example 13.2: Genasence™ in the treatment of advance malignant melanoma

In this study reported by Bedikian *et al.* (2006), several potential baseline prognostic factors were included in a proportional hazards model. These factors were:

- ECOG score (0 versus 1 or 2)
- LDH (elevated versus not elevated)
- Metastatic site:
    - Liver versus visceral other than liver
    - Non-visceral versus visceral other than liver
- Gender (female versus male)
- Platelet count (elevated versus not elevated)
- Alkaline phosphatise (elevated versus not elevated)

In the model each of these factors was coded 0 or 1, for example:

$$x_1 = 1(\text{if ECOG score} = 0) \text{ or} = 0(\text{if ECOG score} = 1 \text{ or } 2)$$
$$x_2 = 1(\text{if LDH elevated}) \text{ or} = 0(\text{if LDH not elevated})$$

For metastatic site, the factor is at three levels and so we need two indicator variables:

$$x_3 = 1(\text{if liver}) \text{ or } = 0(\text{if visceral other than liver})$$
$$x_4 = 1(\text{if non-visceral}) \text{ or } = 0(\text{if visceral other than liver})$$

Treatment-by-centre and treatment-by-covariate interactions can be investigated by including cross-product terms in the model as with binary data and logistic regression, although it is more usual to evaluate these potential interactions visually, as in the example below. All of the remarks made previously in Section 6.7 regarding regulatory aspects of the inclusion of covariates apply equally well to the survival data setting and the proportional hazards model.

is true, the active treatment is giving a higher hazard rate and if $c = 0$ then there is no treatment difference. An analysis using this model then largely centres around testing the hypothesis $H_0 : c = 0$ and subsequently presenting an estimate for the treatment effect. The structure of the model is such that c is the log of the hazard ratio and the anti-log, e^c, is the (adjusted) hazard ratio. This, together with a confidence interval for the hazard ratio, gives us a measure of the treatment effect, adjusting for baseline factors.

Example 13.1 (Continued): Carvedilol in severe heart failure

The proportional hazards model was fitted to the survival data overall and additionally within subgroups defined according to key baseline prognostic factors in order to evaluate the homogeneity of the treatment effect. Figure 13.5 shows these hazard ratios together with 95 per cent confidence intervals. Such plots were discussed earlier in Section 10.8 in relation to subgroup testing. The data in Figure 13.5 indicate that there is a consistency of treatment effect across the various subgroups investigated.

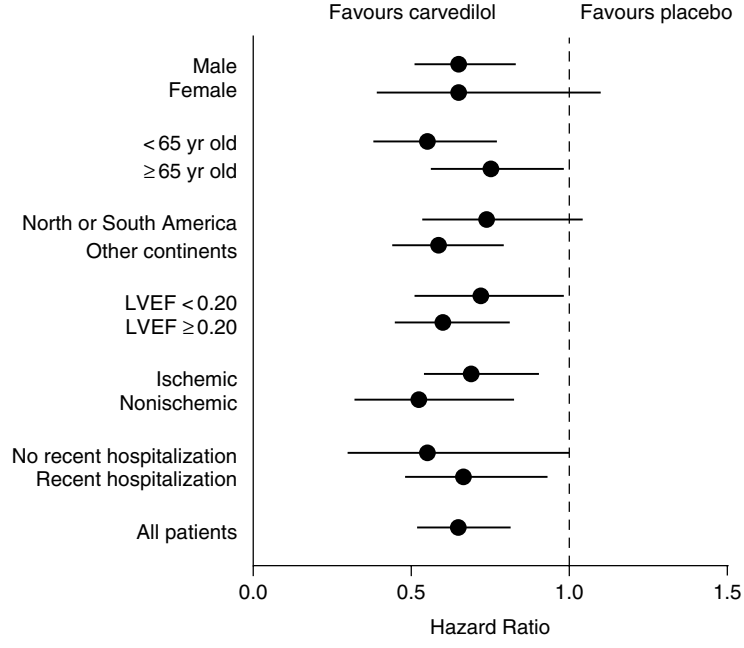


Figure 13.5 Hazard ratios (and 95 per cent confidence intervals) for death in subgroups defined by baseline characteristics (Packer M, Coats AJS, Fowler MB, *et al.* for the Carvedilol Prospective Randomised Cumulative Survival Study Group, 'Effect of carvedilol on survival in severe chronic heart failure', *New England Journal of Medicine*, **344**, 1651–1658. Copyright (2001) Massachusetts Medical Society.

The proportional hazards model, as the name suggests, assumes that the hazard ratio is a constant. As such it provides a direct extension of the logrank test, which is a simple two treatment group comparison. Indeed if the proportional hazards model is fitted to data without the inclusion of baseline factors then the p-value for the test $H_0 : c = 0$ will be essentially the same as the p-value arising out of the logrank test.

13.5.3 Accelerated failure time model

As we have already seen, there will be settings where the pattern of differences between treatment groups does not conform to proportional hazards, where the hazard ratio is not a constant, single value. Such situations are best handled by using an alternative model to incorporate baseline factors. The a*ccelerated failure time model* is an analysis of variance technique which models the survival time itself, but on the log scale:

$$\ell nT = a + cz + b_1 x_1 + b_2 x_2 + \cdots$$

We mentioned earlier, in Section 13.1, that if we did not have censoring then an analysis would probably proceed by taking the log of survival time and undertaking the unpaired t-test. The above model simply develops that idea by now incorporating covariates etc. through a standard analysis of covariance. If we assume that ℓnT is also normally distributed then the coefficient c represents the (adjusted) difference in the mean (or median) survival times on the log scale. Note that for the normal distribution, the mean and the median are the same; it is more convenient to think in terms of medians. To return to the original scale for survival time we then anti-log c, e^c, and this quantity is the ratio (active divided by control) of the median survival times. Confidence intervals can be obtained in a straightforward way for this ratio.

Example 13.3: Time to alleviation of symptoms in flu

The dataset used in this example comprised two phase III placebo-controlled studies used in the license application for oseltamivir (Tamiflu™). See Patel, Kay and Rowell (2006) for further details. Pre-defined symptoms consisted of cough, nasal congestion, headache, sore throat, fatigue, myalgia and feverishness and each of these was assessed on an ordinal scale: none, mild, moderate, severe. Alleviation was recorded as the day on which all symptoms

Example 13.3: (Continued)

were recorded as either none or mild. The Kaplan–Meier curves for time to alleviation of symptoms in the two treatment groups diverged rapidly with 50 per cent of the Tamiflu^{TM} patients having alleviation of symptoms within 72 hours compared to only about 30 per cent in the placebo group. At 192 hours (4 days), however, the proportions of patients still with symptoms in the two groups were almost the same and just over 15 per cent. These curves displayed the shapes seen in Figure 13.2 b). The median times to alleviation of symptoms from the Kaplan–Meier curves were 78.3 hours in the Tamiflu^{TM} arm compared to 112.5 hours in the placebo arm, a ratio of 0.70 corresponding to a reduction of 30 per cent in the Tamiflu^{TM} arm. The sccelerated failure time model, as described earlier, was fitted to these data with an indicator variable for 'smoker' and several indicator variables for geographical region; the coefficients of treatment and smoking were as follows (Table 13.1):

Table 13.1 Accelerated failure time model for flu data

Term	Estimated coefficient	Standard error	p-value
Treatment c	-0.342	0.074	< 0.0001
Smoking b_1	0.007	0.082	0.93

Smoking had a non-significant effect ($p = 0.93$) on time to alleviation of symptoms, but there was a highly significant treatment effect ($p < 0.0001$). The ratio of the adjusted (for smoking and region) median times to alleviation of symptoms obtained from the accelerated failure time model was $e^{-0.342} = 0.71$, corresponding to a 29 per cent reduction in the median time in the Tamiflu^{TM} group. Note that this value is very similar to that obtained directly from the Kaplan–Meier curves.

13.6 Independent censoring

One important assumption which we have made in the analyses presented so far is that the censoring process is independent of the process associated with the event that we are looking at. This is the assumption of *independent censoring*. If, for example, patients were being withdrawn from the trial, and therefore giving censored survival times at the time of withdrawal, because their health was deteriorating, then this assumption would be violated. Doing

this would completely fool us as to the true hazard rate, for example. In the extreme, if this happened with every patient, we would never see any deaths!

Usually censoring occurs at the end of the pre-defined observation/follow-up period and in these cases the issue of independent censoring is not a problem; the censoring process is clearly unconnected with the underlying process for the events. When patients are lost to follow-up or withdrawn, however, there is the potential problem of bias. Unfortunately there is no easy way to deal with this. The best approach is to analyse the data as it is collected and presented, but then undertake one or more sensitivity analyses. For example, for patients who withdraw for negative reasons (deteriorating health, suffering a stroke) assume that the censoring is in fact a death, for patients who withdraw for positive reasons (disease-free for six months) assign the maximum length of follow-up to those patients and censor them at that time. This is the worst case scenario for the withdrawals for negative reasons and the best case scenario for the withdrawals for positive reasons. If the conclusions are essentially unchanged when these assumptions are made then you can be assured that this conclusion is fairly robust to any violation of the assumption of independent censoring. An alternative, and certainly this is possible with a hard endpoint such as death, is to continue follow-up, even though the patient has withdrawn from treatment, to obtain the required information on the time to death.

Example 13.1 (Continued): Carvedilol in severe heart failure

A total of 12 patients (six in each of the two treatment groups) underwent cardiac transplantation during the study. One issue was how to deal with these patients in the analysis as receiving a transplant will impact on the survival prospects for that patient. In the primary analysis these patients were censored at the time of transplantation. As a sensitivity analysis, the eventual time to death (or censoring at the end of follow-up) was included in the analysis; the conclusions were essentially unchanged.

13.7 Sample size calculations

The power of a study where the primary endpoint is time-to-event depends not so much on the total patient numbers, but on the number of events. So a trial with 1000 patients with 100 deaths has the same power as a trial with only 200 patients, but with also 100 deaths. The sample size calculation for survival data is therefore done in two stages. Firstly, the required number of patients suffering events is

calculated, and secondly, this is factored upwards by an assumed proportion of patients who are expected to suffer events, to give the required number of patients for the trial.

Example 13.4: Sample size calculation for survival data

On a standard treatment it is expected that the five year survival rate will be 35 per cent. It is desired to detect an increase in this rate to 55 per cent with 90 per cent power in a 5 per cent level test. The required number of deaths to give this level of power is 133. The overall survival rate at five years is expected to be 45 per cent (the average of 35 per cent and 55 per cent) and therefore we expect to see 45 per cent of patients dying. It follows that $133/0.45 = 296$ patients need to be recruited overall; or 143 patients per group. We would then expect to see 81 deaths in one group and 52 in the second group. With rounding this gives the required number of events needed.

There is of course no guarantee that things will turn out as expected and even though we require 133 events in our example, 296 patients may not give us that. An alternative way of designing the trial would be to continue follow-up until the required number of events is observed. This is frequently done as a safer option to ensure a certain level of power although it does add another element of uncertainty and that is the duration of the study. Careful management of the whole trial process is then needed.

Example 13.1 (Continued): Carvedilol in severe heart failure

This trial adopted a design which continued follow-up until the required number of events was observed:

'The sample size was estimated on the basis of the following assumptions: the one-year mortality in the placebo group would be 28 per cent; the risk of death would be altered by 20 per cent as a result of treatment with carvedilol; the study would have 90 per cent power (two-sided $\alpha = 0.05$) to detect a significant difference between the treatment groups. Since it was recognised that the estimate of the rate of events might be too high, the trial was designed to continue until 900 deaths had occurred.'

The sample size methodology above depends upon the assumption of proportional hazards and is based around the logrank test as the method of analysis. If this assumption is not appropriate and we do not expect proportional hazards then the accelerated failure time model may provide an alternative framework for the

analysis of data. In this case it is no longer strictly true that the power of the study depends only on the number of events (and not directly on the number of patients). However, this still gives a solution that is approximately correct, so we first of all undertake a sample size calculation based on the unpaired t-test for the difference in the medians (means) on the log scale and then factor up by the expected number of patients dying to give the total number of patients needed. Alternatively the trial follow-up continues until the specified numbers of deaths have been seen. The resulting power in either case will be slightly higher than the required power.

14

Interim analysis and data monitoring committees

14.1 Stopping rules for interim analysis

In Chapter 10 we spoke extensively about the dangers of multiple testing and the associated inflation of the type I error. Methods were developed to control that inflation and account for multiplicity in an appropriate way.

One area that we briefly mentioned was interim analysis, where we are looking at the data in the trial as it accumulates. The method due to Pocock (1977) was discussed to control the type I error rate across the series of interim looks. The Pocock methodology divided up the 5 per cent type I error rate equally across the analyses. So, for example, for two interim looks and a final analysis, the significance level at each analysis is 0.022. For the O'Brien and Fleming (1979) method most of the 5 per cent is left over for the final analysis, while the first analysis is at a very stringent level and the adjusted significance levels are 0.00052, 0.014 and 0.045.

These two methods are the most common approaches seen in pharmaceutical applications. A third method, which we see used from time to time, is due to Haybittle (1971) and Peto *et al.* (1976). Here a significance level of 0.001 is used for each of the interims, again leaving most of the 5 per cent left over for the final analysis. For two interims and a final, the adjusted significance level for the final analysis is in fact 0.05 to two decimal places, for three interims we have 0.049 'left over'. Clearly these methods have little effect on the final evaluation, but associated with that there is also little chance of stopping at an interim stage.

Each of these methods assumes that the analyses take place at equally spaced intervals. So, for example, if the total sample size is 600 and we are planning two interims and a final, then the interims should take place after 200 and 400

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

patients have been observed for the primary endpoint, in order for these adjusted significance levels to apply. Small deviations from this make little difference; conducting analyses after say 206 and 397 patients, for example, would essentially work fine. More major deviations, however, would have an impact. Also we may, for practical reasons, not want the analyses to be equally spaced and we will see an example of this later in this chapter. In both these cases we would need to use a more general methodology, α *spending functions*. Application of these methods requires some fairly sophisticated computer software. There are several packages available for these calculations including ADDPLAN (www.addplan.com), EasT (www.cytel.com) and S-Plus (www.insightful.com).

On a final point, for survival data, information is carried through the numbers of patients with events and not the number completing the trial. So for example, if the trial is continuing until we have seen 900 deaths, then the interim analyses would be conducted after, respectively, 300 and 600 deaths. This is what is meant by equally spaced in this case.

14.2 Stopping for efficacy and futility

There are several potential reasons why we would want to stop a trial on the basis of the data collected up to that point:

- overwhelming evidence of efficacy (or harm) of the test treatment
- futility; the data is such that there is little chance of achieving a positive result if the trial were continued to completion
- safety; collective evidence that there are serious safety problems with the test treatment

We will deal with safety and this final point separately in Section 14.3.

14.2.1 Efficacy

Each of the schemes outlined in Section 14.1 for dividing up the 5 per cent type I error can be applied for the evaluation of efficacy in theory. In practice, however, we would only want to be stopping for efficacy if the evidence was absolutely overwhelming. For this reason the O'Brien and Fleming scheme looks to be the most useful in that it has a sliding scale of adjusted significance levels starting from very stringent, through less stringent, to something close to 5 per cent for the final analysis. The Pocock scheme pays too big a price for the early looks and at the final analysis the adjusted significance level is well below 5 per cent. It

would be very unfortunate indeed if, with two interims and a final analysis, the final analysis gave $p = 0.03$. Statistical significance could not be claimed because the p-value has not fallen below the required 0.022. The Haybittle and Peto *et al.* schemes would also not be particularly suitable, giving so little chance of stopping early.

Note that the schemes we are discussing here are based upon the calculation of standard two-tailed p-values. Clearly, we would only be claiming overwhelming efficacy if the direction of the observed treatment effect were in favour of the test treatment. If the direction of the treatment effect were in the opposite direction then we would still be stopping the trial, but now concluding differences in the negative direction and for certain endpoints, for example, survival time this would constitute harm. It could also be appropriate in such cases to consider stopping for harm on the basis of somewhat weaker evidence and using an overall one-sided significance level of 0.025 for efficacy and an overall one-sided significance level of, say, 0.10 for harm. It is straightforward to construct such asymmetric schemes by undertaking two calculations; one with a one-sided significance level of 0.025 (for efficacy) and the other with a one-sided significance level of 0.10 (for harm) and in each case dividing the adjusted two-sided significance levels by two. A further generalisation would allow the use of separate types of schemes for the splitting up of the overall one-sided significance levels: say O'Brien and Fleming for efficacy and Pocock for harm. Such a structure would sensibly be more cautious early on for harm and prepared to stop on modest evidence rather than imposing a much more stringent adjusted significance level using an O'Brien and Fleming scheme.

Example 14.1: Separate one-sided schemes

With two interims and a final analysis looking at both overwhelming efficacy and harm, the adjusted one-sided O'Brien–Fleming significance levels for efficacy of 0.0003, 0.0071 and 0.0225 would give an overall one-sided significance level of 0.025. Using a Pocock scheme for harm with an overall conservative one-sided significance level of 0.10 would use the adjusted significance levels of 0.0494 at each of the interims and the final analysis.

14.2.2 Futility and conditional power

In addition to looking for overwhelming efficacy there is also the possibility of stopping the trial for futility at an interim stage. It may be, for example, that the analysis of the interim data is not at all in favour of the test treatment and were the trial to continue there would simply be no real possibility of obtaining a

positive (statistically significant) result. For commercial reasons it may simply be better to abandon the trial to save on additional costs and resources.

There are several approaches to evaluating futility but the most common method is based on conditional power. At the design stage we may have based the sample size calculation on a power of, say, 90 per cent to detect a certain level of effect, the clinically relevant difference, d, arguing that a difference less than d was of little or no clinical importance. It is possible at an interim stage to recalculate the power of the trial to detect a difference d given that we already have a proportion of the data on which the final analysis will be based. Suppose that this so-called *conditional power* was equal to 20 per cent. In other words, were we to continue this trial (and the treatment difference that we have observed to date reflected the true treatment difference) then there would only be a 20 per cent probability of seeing a significant p-value at the end. Under these circumstances it may not be worth continuing. In contrast, if the conditional power turns out to be 50 per cent then it may be worth carrying on with the trial. The cut-off that is chosen, which should be pre-specified, is based on commercial risk/benefit considerations but generally speaking is around 20 to 30 per cent.

This method of calculating the conditional power assumes that the observed difference between the treatments at the interim stage is the true difference, termed the conditional power *under the current trend*. It is also possible to calculate conditional power under other assumptions, for example, that the true treatment difference in the remaining part of the trial following the interim analysis is equal to d. These calculations under different assumptions about how the future data should behave will provide a broad basis on which to make judgements about terminating the trial for futility.

14.2.3 Some practical issues

It is not a requirement that a trial must have an interim analysis, either for efficacy or for futility. In most long-term trials, however, where there is the opportunity for an interim evaluation then it may be something worth putting in place. The interim can involve only efficacy, only futility, or both and may indeed involve some other things as well, such as a re-evaluation of sample size (see Section 8.5.3).

For practical reasons, however, the number of interims should be small in number. Undertaking interims adds cost to the trial and they also need to be very carefully managed. In particular, the results of each interim must be made available in a timely way in order for go/no-go decisions to be made in good time. Remember the trial does not stop to allow the interim to take place, recruitment and follow-up continues. It has been known in more than one trial for total recruitment to be complete before the results of the interim analysis become available – this is obviously a situation that you would want to avoid, the interim then becomes a

completely pointless (and expensive) exercise. Do not overburden the clinical trial with lots of interims, two at most is my recommendation.

The results of interim analyses are not formally binding, but it would be a very brave decision to continue a trial when the decision coming out of the interim was to stop. If the trial data at the interim has given a statistically significant result then there would clearly be ethical problems in continuing randomising patients when one treatment has been shown to be overwhelmingly superior. The investigators would not be willing to continue with the study. For futility this is less of an issue; there are no real ethical problems with continuing the trial with two treatments that evidence suggests are somewhat similar. There are, however, technical problems when a trial is designed with pre-planned stopping boundaries for futility and then the boundaries are ignored. The type I error of 5 per cent, which has been carefully preserved with the interim analysis plan, is slightly inflated.

It is almost self-evident that all analyses of the kind we are discussing here must be pre-planned in a detailed way. The regulators, in particular, are very unhappy with unplanned interims or interims that are ill-defined. Such situations would give rise to major problems at the time of submission.

14.2.4 Analyses following completion of recruitment

Analyses of data are sometimes undertaken by the sponsor following completion of recruitment, but before follow-up of all patients according to the study schedule for a variety of reasons:

- The sponsor is looking for preliminary information to allow for strategic planning.
- There is a specific request from regulators for this analysis.
- To accelerate regulatory approval. The data may be sufficiently compelling for a regulatory submission to be made.

These analyses are not interims in the formal sense. Generally there is no 'stopping rule' associated with them and the trial will continue to completion irrespective of what the results look like. Great care needs to be taken with regard to these analyses and in particular the question needs to be asked; is this analysis going to compromise the integrity of the study as a whole? If the answer to this question is potentially yes, then the analysis is ill-advised. Given that there is no stopping rule, it is very difficult to know what kind of price to pay in terms of α. One recommendation is to treat this as a so-called 'administrative' analysis and use an adjusted α of 0.001 leaving the final (primary analysis)

unaffected with an α of 0.05. A further potential for bias to be introduced is associated with dissemination of the results. Will the investigators and others who are made aware of the interim results behave in a way that could compromise the ability of the trial to reach valid conclusions? This aspect of dissemination will be a common theme throughout this chapter and needs to be carefully controlled.

14.3 Monitoring safety

In addition to considerations of efficacy and futility, it will usually be appropriate in most long-term trials to consider safety in an ongoing way. This is not new and we have always, for example, looked at accumulating data on individual serious adverse events and considered stopping (or modifying) trials if these are indicative of problems with the trial or with the treatments.

As time has gone on we have increasingly done this in a very structured way. Data Monitoring Committees have a major role here and we will consider various aspects of their structure and conduct later. For the moment we will just focus on associated statistical methodologies. Usually this ongoing safety monitoring is done by looking at various aspects of safety; adverse events (serious and non-serious), vital signs, key laboratory parameters, physical examination, ECGs etc., both in individual cases and overall across the treatment groups or the study as a whole. I hesitate to say that this is done in an informal way, because it is taken very seriously, but what I mean is that there is usually little formal statistical structure wrapped around the process. Yes, we may put *p*-values on the adverse events, suitably grouped, but these are simply used as flags for potential problems. A discussion ensues and decisions are taken by members of the Data Monitoring Committee. This is not necessarily a bad thing, it fits with a broad requirement to look at many different aspects of safety across the trial as a whole.

If at the design, stage and based on the nature of the disease and the treatment, specific potential safety issues can be identified then more formal rules can be set-up. These rules are unlike those considered earlier for overwhelming efficacy. For safety it would not usually be appropriate to look just at one or two interim stages, we must look more frequently than that. Bolland and Whitehead (2000) propose a *sequential plan* for monitoring a particular safety event (for example death or major bleed) which allows the evaluation of the significance of treatment differences at regular intervals yet preserves a pre-specified overall type I error rate. In preserving an overall type I error rate, this method avoids the false positive at a specified level and over-reacting to unfolding events. Evaluating the accumulating data in relation to the plan on a monthly basis or following recruitment of every ten patients are the kinds of settings that one might consider.

14.4 Data Monitoring Committees

14.4.1 Introduction and responsibilities

In this section we will cover several aspects, particularly in relation to statistical issues, associated with Data Monitoring Committees (DMCs). This is not meant to be a comprehensive coverage of the area and the reader is referred to the book by Ellenberg *et al.* (2003) for an excellent and exhaustive coverage that in addition contains a plethora of case studies. There are two guidelines, one from the FDA (2006) 'Establishment and Operation of Clinical Trial Data Monitoring Committees' and one from the CHMP (2005) 'Guideline on Data Monitoring Committees', which outline the roles and responsibilities of DMCs in the regulatory environment. DMCs are also referred to as Data Monitoring Boards (DMBs) and Data and Safety Monitoring Committees/Boards (DSMCs/DSMBs).

It is clearly important that the trial sponsors remain blind to the accumulating data within the separate treatment groups and the main reason for having a DMC is to enable trial data to be looked at without compromising that blinding. The main responsibilities of a DMC will vary depending on the particular circumstances. The first responsibility, however, will always be to protect the safety of the trial participants. There may be additional responsibilities associated with interim analyses, if these are to be incorporated in the trial, and at all times in these collective activities the DMC will have a responsibility to protect to the scientific integrity of the trial. The DMC may or may not be involved in interim analyses associated with efficacy if indeed there are any and the committee may only be looking at safety. While this is sometimes the case it is not ideal. It is difficult for DMCs to make recommendations based on safety in isolation in that the absence of efficacy makes it impossible to make a risk–benefit judgement. There should therefore be provision for the committee to access efficacy data should they request it.

CHMP (2005): 'Guideline on Data Monitoring Committees'

'In most cases, safety monitoring will be the major task for a DMC. Even if the safety parameters monitored are not directly related to efficacy, a DMC might need access to unblinded efficacy information to perform a risk/benefit assessment in order to weigh possible safety disadvantages against a possible gain in efficacy.'

In protecting the scientific integrity of the trial the DMC must ensure that all interim analyses and safety monitoring activities are conducted in an appropriate way to protect blinding and the type I error to avoid problems with multiplicity. There is also the responsibility to oversee the overall conduct of the trial to make sure that it will fulfil its objectives.

FDA (2006): 'Establishment and Operation of Clinical Trial Data Monitoring Committees'

'A DMC will generally review data related to the conduct of the study (that is, the quality of the study and its ultimate ability to address the scientific questions of interest), in addition to data on effectiveness and safety outcomes. These data may include, among other items:

- *Rates of recruitment, ineligibility, non-compliance, protocol violations and dropouts, overall and by study site;*
- *Completeness and timeliness of data;*
- *Degree of concordance between site evaluation of events and centralised review;*
- *Balance between study arms on important prognostic variables;*
- *Accrual within important subsets.'*

It is important that interaction between the sponsor and the DMC is kept to an absolute minimum to avoid any inadvertent communication of unblinded information. Sponsor exposure to unblinded interim results, except in terms of the strict rules concerning formal interim analyses, can seriously compromise the scientific validity of the study. The sponsor is in the position of being able to influence the future conduct of the trial and exposure to interim results could influence aspects of that, leading to bias. There can also be pressure from the sponsor to provide interim results for planning purposes; taking decisions about future production facilities, agreeing budgets for further trial activity and so on. Such pressure should be resisted as it could lead to the integrity of the trial being seriously undermined. Where this need is compelling then communication should be managed in a very tight way. Further discussion of this point is provided in the FDA guideline in Section 6.5.

A DMC is usually needed in long-term trials in life-threatening diseases and sometimes in non-life-threatening diseases where there are potential safety concerns. It may also be necessary to have DMCs in studies in specific and vulnerable or fragile populations such as children, pregnant women or the very elderly, but DMCs are not usually necessary in phase I and early phase II trials or in short-term studies where the goal is relief of symptoms.

14.4.2 Structure

The independence of the committee from the sponsor is important and there should also be no conflicts of interest amongst the participants, for example,

holding equity in the sponsor's company or a direct competitor's company. The members of the committee should also not otherwise be involved in the study, for example, as investigators or in the case of the statistician, the analysis of the data. The DMC consists of at least three participants, one of which will be a statistician; the remaining participants will be clinicians with expertise in relevant clinical disciplines associated with the disease under study or with the potential side effects. If the trial is an international study then it is advisable to have members of the committee from the spread of geographical regions in which the trial is to be conducted. It is often this final point which determines the ultimate size of the DMC.

There will also be at least one other statistician involved closely with the activities of the DMC and this is the statistician who supplies data tables to the committee for their deliberations. This statistician should also not be otherwise involved in the trial as they will potentially be supplying unblinded information to the DMC and attending their meetings. In the way that these things tend to be organised these days this individual may be part of a CRO that is providing this service (and potentially other services) to the sponsor. See Pocock (2004) and the FDA guideline for further discussion on this and related points. The DMC should also receive details of individual SAEs, usually in the form of narratives and these will often be supplied directly from the sponsor. These patients can be unblinded by the independent statistician if this has not already been done.

Data tables produced for the DMC should contain separate summaries by treatment group, with the treatment groups labelled A and B (partially blinded). A separate sealed envelope or a password-protected electronic file should be provided to the members with decodes for A and B to enable the DMC members to be completely unblinded. This may seem an elaborate process, but it protects against inadvertent unblinding.

FDA (2006): 'Establishment and Operation of Clinical Trial Data Monitoring Committees'

'A common approach is presentation of results in printed copy tables using codes (for example, Group A and Group B) to protect against inadvertent unblinding should a report be misplaced, with separate access to the actual study arm assignments provided to DMC members by the statistical group responsible for preparing DMC reports.'

The activities of the DMC should be covered by a charter, prepared in advance of running the trial. The charter should detail the participants, their responsibilities, the format and conduct of meetings, communication pathways with the sponsor, decision-making, confidentiality, indemnity and conflict of interest issues together with details regarding the format of the data supplied to the DMC by the

independent statistician and the supply of other data, for example details of SAEs. It is advisable to involve the DMC members as early as possible in the review and possibly the construction of this document in order to gain clear buy-in to their role in the successful conduct of the trial.

14.4.3 Meetings and recommendations

The DMC should meet at a set of pre-defined time points during the course of the trial, typically following completion of a proportion of the patients. For example, four meetings could be organised, following completion of 25 per cent, 50 per cent, 75 per cent of the patients and finally at trial completion. If the trial is to involve interim analyses then some of the meetings will revolve around those. Summary tables should be supplied in conjunction with all meetings. Meetings of the committee will usually be organised in open and closed sessions. The open sessions, which will also involve, for example, members of the sponsor company and the steering committee, will cover general issues such as recruitment rates, timelines and the presentation of summary demographic and baseline data tables. Details can also be given during these open sessions on the progress and results from other trials within the drug development programme. The closed sessions will involve the members of the DMC plus the independent statistician supplying data to the DMC, if required. Minutes of both closed and open sessions of these meetings should be kept and the closed minutes should be stored securely and confidentially by the chair of the DMC until trial completion. Electronic copies of the data sets on which interim analyses are based should also be retained as these may be requested by regulators.

Outside of the regular planned meetings, details of all SAEs should be supplied to the DMC in real time and the DMC members should arrange to discuss these, by email or teleconference for example, as and when they feel necessary.

Recommendations to the sponsor coming out of the regular DMC meetings will be one of the following:

- Trial to continue unchanged.
- Modification of the protocol to protect the safety of the trial participants.
- Termination of the trial.

Clearly if the recommendation is anything other than ‘continue unchanged’ then additional information would need to be supplied to support these recommendations. The recommendations are not binding on the sponsor, although as mentioned earlier it would be very unusual to see these recommendations being ignored.

Example 14.2: ESTAT trial in acute stroke

This was a multi-centre, pan-European, randomised double-blind placebo-controlled clinical trial in acute stroke to evaluate the effect of ancrod, a natural defribrinogenating agent (Hennerici *et al.* (2006)). The primary endpoint was based on the Barthel Index; a favourable score of 95 or 100 or a return to the pre-stroke level at three months was viewed as a success. The primary method of statistical analysis was based on a logistic model including terms for treatment, age category, baseline Scandinavian Stroke Scale and centre.

The proposed sample size was 600 patients and two interims were planned after 200 and 400 patients (completing 3 months follow-up) using the O'Brien and Fleming scheme with adjusted two-sided significance levels of 0.00052, 0.014 and 0.045. A futility rule was also introduced, based on conditional power (under the current trend) being below 30 per cent for the trial to be stopped.

Safety was a concern and a formal rule was implemented (Bolland and Whitehead (2000)) with an overall one-sided type I error of 0.025 and 90 per cent power to detect an excess death rate in the ancrod group of 29 per cent compared to 18 per cent on placebo. This sequential plan was updated following the recruitment of every 20 patients.

14.5 Adaptive designs

14.5.1 Sample size re-evaluation

In Chapter 8 we spoke about the calculation of sample size and in Section 8.5.3 revisiting this sample size calculation as the trial data accumulates.

In theory it is possible to both increase the sample size and decrease the sample size in light of these interim calculations. Decreasing the sample size, however, is rarely done. The choice of the sample size is usually not just an issue of ensuring enough power for the evaluation of efficacy; it is also about having enough patients in the active group to provide a large safety database for the development plan as a whole. Also there will be considerations for secondary endpoints, so reducing sample size is often ruled out for reasons other than power; nonetheless, in principle, it can be done.

Maintaining the blind is critical in these re-evaluations, otherwise a price in terms of the type I error, α, may need to be paid as looking at unblinded data could be viewed as a formal interim comparison of the treatments. There may

also be external factors that result in a change in sample size. Such considerations do not involve any unblinding of the trial data and so there is no price to pay in terms of the type I error.

Generally speaking these sample size considerations are independent of any formal interim comparisons of the treatments, be they efficacy or futility. Given the need to maintain blinding it is also advisable *not* to involve the DMC in these re-evaluations; their potential knowledge of the unfolding data in an unblinded way could well influence their view on a change in the sample size – they know too much!

Example 14.2 (continued): ESTAT trial in acute stroke

The original sample size was based on the primary endpoint, success on the Barthel Index or return to pre-stroke level. In a 5 per cent level two-sided test, a total of 293 patients per group were needed to give 95 per cent power to detect an improvement in success rate from 35 per cent in the placebo group to 50 per cent in the ancrod group. To allow for a small number of dropouts the planned recruitment figure was rounded up to 600. The results of a sister trial conducted in the US (Sherman *et al.* (2000)) revealed a smaller than expected treatment difference, 34 per cent in the placebo group and 42 per cent in the ancrod group. On this basis the recruitment figure for ESTAT was revised to 1680. This was based on a placebo response rate of 35 per cent and an absolute treatment difference of 8 per cent.

In line with this, the interim analysis plan was revised and only one interim was to be conducted for both efficacy and futility; after 40 per cent of the patients had completed 3 months of follow-up. Since the two proposed analyses were not equally spaced, α spending functions were needed to revise the adjusted significance levels and these turned out to be 0.0007 and 0.0497.

The trial was, in fact, stopped at the interim analysis due to futility; the conditional power, was well below 30 per cent. In addition, even though the formal boundary for safety was not crossed, the trend was in the direction of excess deaths in the ancrod group.

14.5.2 Flexible designs

Re-evaluating sample size is one example of adapting the design of the trial based on accumulating data from the trial. There is currently substantial interest in the potential of allowing flexibility in this way. These considerations come under the heading of *adaptive* or *flexible designs*.

Undertaking interim analyses for either efficacy or futility or both, together with sample size re-evaluation already provides a range of flexibility in the design and planned correctly these approaches may cover many situations of practical interest. What more could be done? Phillips and Keene (2006) list a range of potential adaptations including the following:

- Dropping/adding a treatment arm.
- Change in the patient population.
- Change in the primary endpoint.
- Sample size re-evaluation based on unblinded data.

Dropping treatment arms is already done for safety reasons. Outside of that, this could be built into a formal interim look based on futility. Van Leth *et al.* (2004) provide an example (the 2NN Study) of adding a treatment arm based upon external considerations following the publication of results from a related trial. Analyses of the resulting data involved consideration of all of the data and secondly only those data from the trial following the change. We will discuss further aspects of this trial later in this section.

A change in the patient population could also be undertaken for safety reasons. If this involves a major change to the inclusion criteria then there could be difficulties in extrapolating the trial results to a clear target population. This would need careful consideration at the data analysis stage; one analysis would invariably be based again on only those patients recruited following the change.

The primary endpoint is that endpoint which provides the best measure of treatment response/success and such choices are based on clinical arguments rather than statistical ones. It seems unlikely that a change in this could be justified.

Finally, it is, in principal, possible to increase the sample size based on the observed treatment difference at an interim stage without affecting the type I error, but great care needs to be taken with regard to dealing with this statistically. Evidence to date suggests that such procedures offer no real advantages over and above a standard interim analysis plan.

A recent European regulatory document (CHMP (2006) 'Reflection Paper on Methodological Issues in Confirmatory Clinical Trials with Flexible Design and Analysis Plan') has given more thought to what is acceptable and what is not, in terms of flexible design. Several common themes are seen to run through this paper:

- Prior to phase III, in an exploratory setting, flexibility in design is much less of an issue and is indeed to be encouraged.

- Phase III is the confirmatory stage and major adaptation outside of pre-planned interim analyses and blinded sample size re-evaluation could potentially undermine the confirmatory nature of the trial.
- Having several adaptations in a phase III setting is unlikely to be acceptable; it compromises the confirmatory nature of the trial.
- A planed adaptation is much more likely to be acceptable than an unplanned adaptation.
- The mathematics for many adaptations is essentially solved and this is not the issue – it is the logistics, control and interpretation that present the major problems.
- If a change in design is made at an interim stage then lack of consistency between the results before the change and after the change could compromise the ability to draw general conclusions from the trial. It is the sponsor's responsibility to confirm this consistency. Example 14.3 discusses this issue in the context of a particular trial.
- Dissemination of information regarding the results from the first part of the trial could cause bias following the change.
- Currently we have little experience with flexible designs; more experience is needed before these designs can be used.

Example 14.3: The 2NN Study

This was an open-label, parallel group, randomised trial in patients with chronic HIV-1 infection reported by van Leth *et al.* (2004). The primary endpoint was treatment failure, a composite endpoint based on virology, disease progression or therapy change. Initially patients were randomised equally to one of the following three groups:

1. N_1 – Nevirapine (once daily)
2. E – Efavirenz
3. $N_1 + E$ – combination of Nevirapine (once daily) and Efavirenz

Five months into the trial (388 patients randomised) another study showed that the effectiveness of nevirapine was related to the minimum concentration and so a fourth arm was added; nivarapine twice daily (N_2). Randomisation was now to N_1, E, $N_1 + E$ and N_2 in the ratio 1:2:1:2. The final sample size was 1216 and in the final analysis, data from both periods

were combined. Is this appropriate? In the treatment groups which were used throughout the study the following failure rates together with 95 per cent confidence intervals were seen (Table 14.1).

Table 14.1 Failure rates in the 2NN study

	Before addition of 4th arm	After addition of 4th arm
N_1	46.6% (37.8%, 55.5%)	39.3% (29.1%, 50.3%)
E	34.4% (26.3%, 43.2%)	39.4% (33.5%, 45.5%)
$N_1 + E$	51.6% (42.5%, 60.6%)	55.4% (44.1%, 66.3%)

The issue here is the lack of homogeneity in the results before and after the change. In particular there is a 12.2 per cent absolute difference between nevirapine (once daily) and efavirenz before the design change and virtually no difference between these two arms after the change. No explanation of why this had occurred was given. This lack of consistency would cause regulators major concern.

15
Meta-analysis

15.1 Definition

Meta-analysis is a formal way of bringing together the results from separate studies to provide an overview regarding a particular treatment or intervention. More generally, the phrase statistical or systematic review/overview is sometimes used to describe an analysis of this kind. Meta-analysis, however, should not be confused with pooling. *Pooling* is a related procedure that simply puts all the data together and treats the data as if they came from a single study. Meta-analysis does not do this; it recognises study-to-study variation and indeed looks to see if the data from the different studies are giving a consistent answer. Meta-analysis is to be preferred to pooling as the following example illustrates.

Example 15.1: Success rates in removing kidney stones

This non-randomised investigation (Julious and Mullee (1994)) compared open surgery with percutaneous nephrolithotomy in removing kidney stones. The following 2×2 contingency tables (Table 15.1) presents the data separately for patients presenting with small stones (< 2 cm diameter) and for patients with larger stones (≥ 2 cm in diameter).

Table 15.1 Contingency tables for kidney stones

Stone diameter < 2 cm	Success	Failure	Total
Open surgery	81 (93%)	6	87
Perc. nephr.	234 (87%)	36	270

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

Example 15.1: (Continued)

Table 15.1 (Continued)

Stone diameter ≥ 2 cm	Success	Failure	Total
Open surgery	192 (73%)	71	263
Perc. nephr.	55 (69%)	25	80

It is clear from each of these tables that open surgery is more successful than percutaneous nephrolithotomy (93 versus 87 per cent for small stones, 73 versus 69 per cent for large stones), irrespective of the size of the kidney stone.

Now however consider pooling the data (Table 15.2).

Table 15.2 Combined data

Diameter < 2 cm and ≥ 2 cm combined	Success	Failure	Total
Open surgery	273 (78%)	77	350
Perc. nephr.	289 (83%)	61	350

This table suggest that percutaneous nephrolithotomy has a higher success rate (83 versus 78 per cent) than open surgery, but we are being fooled; the reverse is true. The pooled table is misleading and it is the separate tables that are the basis of a correct interpretation. We are being misled because of two things: firstly patients with small stones do better than patients with larger stones and secondly patients with small stones tend to receive percutaneous nephrolithotomy. The overall high success rate on percutaneous nephrolithotomy is primarily due to the fact that in the main these were the patients who presented with small kidney stones.

This phenomenon, known as *Simpson's Paradox*, illustrates the dangers of simple pooling. A meta-analysis for the data in the example would be based on the separate tables and compute a treatment difference for those patients with small stones and a treatment difference for patients with larger stones. These two differences would then be averaged to give a valid 'combined' treatment difference.

A second (hypothetical) example illustrates how these problems with pooling could occur with data on adverse events within the context of randomised comparisons.

Example 15.2: Adverse event rates (hypothetical)

In two randomised trials, each comparing two active treatments in terms of the incidence of a particular adverse event, the data were as follows (Table 15.3):

Table 15.3 AE rates in two trials

Treatment A	Treatment B	Difference
10.0% ($n = 100$)	8.0% ($n = 200$)	2.0%
3.5% ($n = 200$)	2.0% ($n = 100$)	1.5%

Pooling the data gives an overall adverse event rate on treatment A of 5.7 per cent, compared to 6.0 per cent on treatment, a pooled difference (A − B) of −0.3 per cent. This is clearly misleading and a more appropriate measure of the treatment difference is given by the average absolute difference of 1.75 per cent.

It is recognised that the studies usually being combined in a meta-analysis will not all be identical and will not all have the same protocol. The dosages may be different, the endpoints may be different, treatment duration may differ, the nature of both the treatment and the comparator may be different and so on. Clearly, the more specific the question being addressed by the meta-analysis the more closely the studies must match, but the breadth of the studies that are combined will depend on the breadth of the question being asked.

15.2 Objectives

Meta-analysis is used in numerous different ways and both within and outside of the regulatory setting.

The technique can provide a quantification of the current state of knowledge regarding a particular treatment both in terms of safety and efficacy. The *Cochrane Collaboration* uses the methodology extensively to provide systematic overviews of treatments in particular therapeutic areas and to answer general health questions.

In a similar way, meta-analysis can be a useful way to combine the totality of data from studies in relation to a particular treatment, perhaps as the basis for a marketing campaign.

Combining studies can also very effectively increase power for primary or secondary endpoints or for particular subgroups. Individual studies are unlikely to be powered for secondary endpoints and subgroups, and meta-analysis can be

an effective way of increasing power in relation to these. For primary endpoints, increasing the power in this way will improve precision (reduce the standard error) and give narrower confidence intervals enabling clinical benefit to be more clearly defined.

A more recent area of application for meta-analysis is in the choice of the non-inferiority margin, Δ. As mentioned in Section 12.8, Δ is often chosen as some proportion of the established treatment effect (over placebo) and meta-analysis can be used to obtain an estimate of that treatment effect and an associated confidence interval.

Combining studies can be useful in resolving apparently conflicting results. For example, Mulrow (1994) reported a meta-analysis of trials of intravenous streptokinase for treatment in acute myocardial infarction. The complete analysis involved a total of 33 trials reported between 1959 and 1988 and Mulrow presented a cumulative meta-analysis which combined the trials chronologically over time. Of the eight trials reported between 1959 and the end of 1973, five gave odds ratios that favoured intravenous streptokinase (two were statistically significant) while three trials gave odds ratios favouring control (none were statistically significant). In one sense there was a confusing picture emerging with six negative/inconclusive trials out of the first eight conducted. The meta-analysis combination at that point in time, however, gave a clear result, with an odds ratio around 0.75 in favour of streptokinase, and a highly significant p-value of 0.0071.

The technique can also address whether or not the studies provide a consistent result and exploring heterogeneity is a key element of any meta-analysis.

In the applications that follow we will focus on the combination of clinical trials although the methodology can also apply more widely in an epidemiological setting; hence the use of the word 'study' in the chapter so far.

15.3 Statistical methodology

15.3.1 Methods for combination

Each trial that is to be included in the meta-analysis will provide a measure of treatment effect (difference). For continuous data this could be the mean response on the active treatment minus the mean response in the placebo arm. Alternatively, for binary data the treatment effect could be captured by the difference in the cure rates, for example, or by the odds ratio. For survival data, the hazard ratio would often be the measure of treatment difference, but equally well it could be the difference in the two-year survival rates.

Assume that we have decided on the best measure for the treatment effect. If this is expressed as a difference, for example, in the means, then there will be an associated standard error measuring the precision of that difference. If the

treatment effect is captured by a ratio, for example, an odds ratio or a hazard ratio, then there will be an associated standard error on the log scale, the log odds ratio or the log hazard ratio.

Again, whichever measure of treatment effect is chosen, the meta-analysis combination proceeds in a standard way. We average the treatment effect over the m studies being combined. This is not the straight average, but a weighted average, weighted according to the precision for each individual study and this precision is captured by the standard error. For study i let d_i be the treatment effect with associated standard error se_i. The overall estimate of the treatment effect is then:

$$d = (w_1 d_1 + w_2 d_2 + \cdots + w_m d_m) / (w_1 + w_2 + \cdots + w_m)$$

where $w_i = 1/se_i^2$. Essentially weighting by the standard error in this way is weighting by the sample size so that the larger studies are given more weight.

If the treatment effect in each of the individual trials is the difference in the mean responses, then d represents the overall, adjusted mean difference. If the treatment effect in the individual trials is the log odds ratio, then d is the overall, adjusted log odds ratio and so on. In the case of overall estimates on the log scale we generally anti-log this final result to give us a measure back on the original scale, for example as an odds ratio. This is similar to the approach we saw in Section 4.4 when we looked at calculating a confidence interval for an odds ratio.

15.3.2 Confidence Intervals

The methods of the previous subsection give us a combined estimate, d, for the treatment effect. We now need to construct a confidence interval around this estimate. This initially involves obtaining a standard error, se, for d, which is given by:

$$se = \frac{1}{\sqrt{w_1 + w_2 + \cdots + w_m}}$$

From this it is easy to obtain a 95 per cent confidence interval for the overall treatment effect as $(d - 1.96se, d + 1.96se)$.

If this confidence interval is on the log scale, for example with both the odds ratio and the hazard ratio, then both the lower and upper confidence limits should be converted by using the anti-log to give a confidence interval on the original odds ratio or hazard ratio scale.

15.3.3 Fixed and random effects

The *fixed effects model* considers the studies that have been combined as the totality of all the studies conducted. An alternative approach considers the collection of studies included in the meta-analysis as a random selection of the studies that have been conducted or a random selection of those that could have been conducted. This results in a slightly changed methodology, termed the *random effects model*. The mathematics for the two models is a little different and the reader is referred to Fleiss (1993), for example, for further details. The net effect, however, of using a random effects model is to produce a slightly more conservative analysis with wider confidence intervals.

In the remainder of this section we will concentrate on the fixed effects approach, which is probably the more common and appropriate approach, within the pharmaceutical setting.

15.3.4 Graphical methods

An extremely useful addition to the formal method for combining the studies is to represent the data from the individual studies, together with the combination, in a meta-analysis plot.

Note that the confidence intervals in Figure 15.1 are not symmetric around the estimated hazard ratio. This is because confidence intervals for hazard ratios and odds ratio and indeed ratios in general are symmetric only on the log scale (see Section 4.5.5 for further details with regard to the odds ratio). Sometimes we see plots where the x-axis is on the log scale, although it will be calibrated in terms of the ratio itself, and in this case the confidence intervals appear symmetric.

The studies with the highest precision are those with the narrowest confidence intervals and usually these aspects of the different trials are emphasised by having squares at the estimated values, whose size is related to the precision within that trial. These plots, as seen in Figure 15.1, are so-called *Forest*
.

Example 15.3: Meta-analysis of adjuvant chemotherapy for resected colon cancer in elderly patients

Sargent *et al.* (2001) provide a meta-analysis of seven phase III randomised trials, involving a total of 3351 patients, that compares the effects of fluorouracil plus leucovorin (five trials) or fluorouracil plus levamisole (two trials) with surgery alone in patients with stage II or stage III colon cancer.

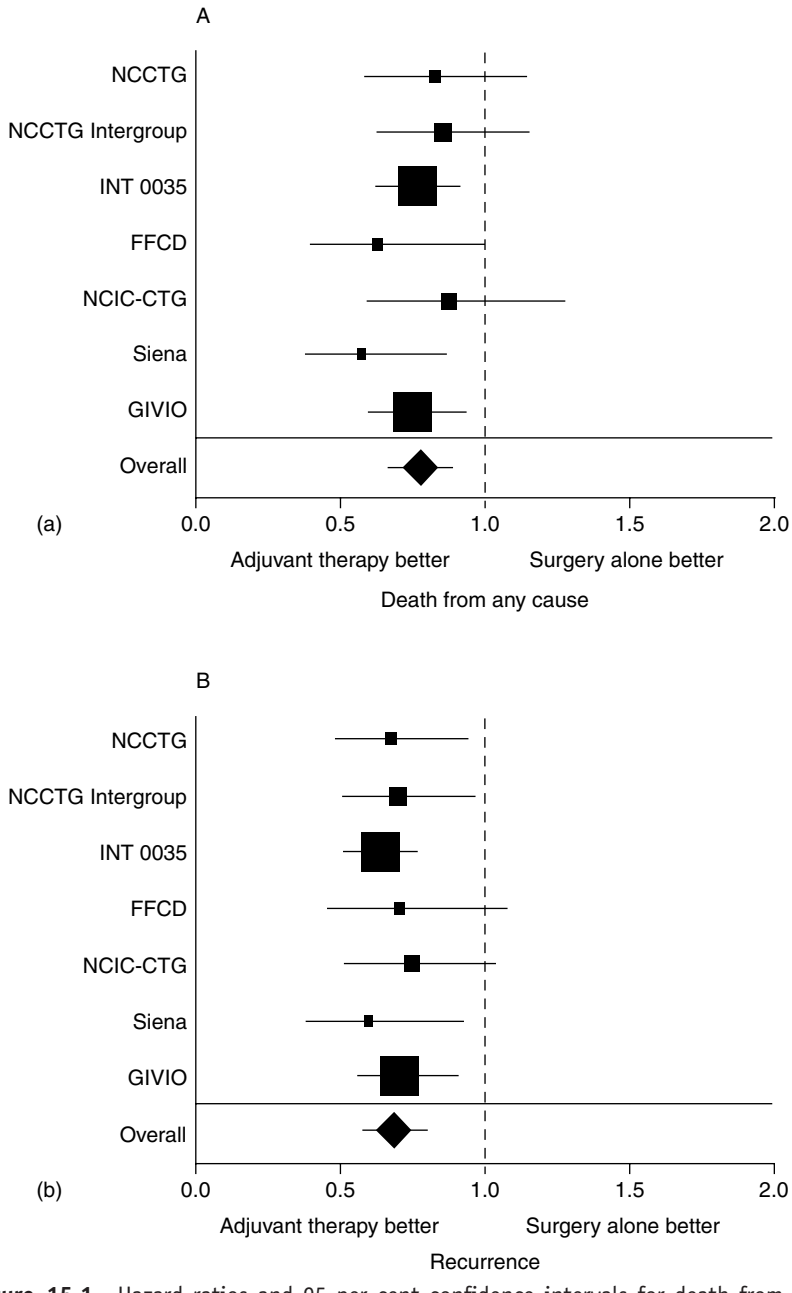


Figure 15.1 Hazard ratios and 95 per cent confidence intervals for death from any cause (panel (a)) and recurrence (panel (b)) by treatment group (Sargent DJ, Goldberg RM, Jacobson SD, MacDonald JS *et al.*, 'A pooled analysis of adjuvant chemotherapy for resected colon cancer in elderly patients', *New England Journal of Medicine*, **345**, 1091–1097. © (2001) Massachusetts Medical Society.)

plots. Each square has area proportional to the size of the study. This helps visually to identify those studies that are providing the most precise information; these are the ones with the most prominent squares.

15.3.5 Detecting heterogeneity

A key element of any meta-analysis is to look for heterogeneity across the studies. This is akin to looking for treatment-by-centre interactions in a multi-centre trial, here we are looking for treatment-by-study interactions. This is done by calculating the statistic:

$$Q = w_1 (d_1 - d)^2 + w_2 (d_2 - d)^2 + \cdots + w_m (d_m - d)^2$$

and comparing the resulting value with the χ^2_{m-1} distribution to obtain the p-value. If the individual studies are giving very different results so that the d_i values are very different, then, on average, the differences $d_i - d$ will be large and Q will be significant on the χ^2_{m-1} scale. A significant p-value, and again we would usually use 0.10 as the cut-off for statistical significance for interactions, indicates that we do not have homogeneity.

Alternatively, but less formally, the plot displaying estimates and 95 per cent confidence intervals can be inspected. Homogeneity is confirmed if there is consistency overall across all of the confidence intervals.

If a lack of homogeneity is seen, the next step is to try to understand why. Clearly, by definition, the studies that have been combined have different protocols and by looking at important features of the studies that reflect those differences, it may be possible to form subgroups of trials that are homogeneous.

15.3.6 Robustness

Within any meta-analysis some trials will be larger than others and because of the way the trials are combined, the larger trials, i.e. those that have higher precision, will tend to dominate. It is helpful therefore to assess the robustness of the overall conclusion by omitting maybe the largest study or studies to see if the result remains qualitatively the same. If it does then the result is robust. If it does not then the overall result is undermined as it is then giving a result that is driven by the largest trial or trials.

15.4 Ensuring scientific validity

15.4.1 Planning

In order to ensure that a meta-analysis is scientifically valid it is necessary to plan and conduct the analysis in an appropriate and rigorous way. It is not sufficient to retrospectively go to a bunch of studies that you like the look of and stick them together!

Ideally the meta-analysis should be pre-planned within a development plan and the rules regarding which trials are to be combined and in what way, set down in advance of running the trials. This will be possible within the regulatory environment, but, of course, will not be possible under other circumstances, for example if the meta-analysis is being undertaken for marketing purposes or if the meta-analysis is part of a post hoc combination of trials to answer a general public health question. Nonetheless, even in these circumstances, the meta-analysis requires very careful planning with rules clearly specified regarding the choice of the studies to be combined and again the methods of combination.

The CPMP (2001) 'Points to Consider on Application with 1. Meta-Analysis; 2. One Pivotal Study' indicates that it is good practice to write a 'protocol' for the meta-analysis:

'When a meta-analysis is included in an application it should be performed in accordance with a protocol . . . '

This document then goes on to list the issues that should be covered by that protocol:

- Objective of the meta-analysis.
- Criteria for inclusion/exclusion of studies.
- Hypotheses and endpoints.
- Statistical methods.
- Approaches to ensure consistent quality of the studies and how to handle poor quality studies.
- Evaluating homogeneity and robustness.

Writing a protocol in this way is important for a meta-analysis that is pre-specified within a development plan, but, in order to ensure its integrity, it is maybe

even more critical when an analysis is performed either retrospectively within a regulatory submission or as a marketing exercise.

Normand (1999) expands on many of the issues surrounding the planning of a meta-analysis and the reader is referred to this article for more information.

15.4.2 Publication bias and funnel plots

One particular issue, however, concerns meta-analyses based upon data obtained through a literature search. It is certainly true that a study which has given a statistically significant result is more likely to be reported and accepted for publication; so if we only focused on published studies then we would get a biased view of the totality of the available studies. Eggar and Smith (1998) discuss various aspects of this *publication bias* and its causes. There have been many calls over the years for registers of studies to be set up and in early 2005 the European Federation of Pharmaceutical Industries and Associations (EFPIA), the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA), the Japanese Pharmaceutical Manufacturers Association (JPMA) and the Pharmaceutical Research and Manufacturers of America (PhRMA) issued a joint statement committing to increasing the transparency of research by setting up a Clinical Trial Registry. Although this is a voluntary initiative, most companies are following this guidance and registering their clinical trials. The registry is maintained by the National Library of Medicine in the US and can be found at www.clinicaltrials.gov. In theory this makes it possible to identify all studies sponsored by the industry and potentially avoiding this publication bias.

There is a graphical technique available, introduced by Eggar *et al.* (1997), called a *funnel plot*, which helps to detect the presence of publication bias by plotting the treatment effect (for example, the difference in the means or the odds ratio) in each study on the x-axis against the sample size on the y-axis. Smaller studies will tend to give more variable results in terms of the observed treatment difference while the larger studies should give more consistent results. The resultant plot with all studies included should then appear like a funnel with the wide part of the funnel at the bottom and the narrow part of the funnel at the top. If, however, there is publication bias then the non-significant studies, very often those with smaller sample sizes will be under represented and either the lower left-hand part of the plot or the lower right-hand part of the plot, depending on how the active compared to placebo difference is measured, will be missing. Once this has been detected then it can be compensated for in the statistical analysis. Visually inspecting the funnel plot in this way is somewhat informal, although the technique can provide some reassurance regarding the absence of publication bias. However, it is not a substitute for relentlessly tracking down all of the studies and all of the data.

Example 15.4: Meta-Analysis of trials of magnesium and streptokinase in acute myocardial infarction

Smith and Eggar (1997) in a letter to *The Lancet*, use funnel plots (shown in Figure 15.2) to illustrate publication bias and to link this with outcomes in several large trials conducted subsequent to the trials within those plots.

In the upper part of Figure 15.2 we see a funnel plot of trials evaluating the effect of intravenous magnesium in the treatment of myocardial infarction. Note the absence of small trials with odds ratios greater than one (which would indicate a lack of benefit for intravenous magnesium); this

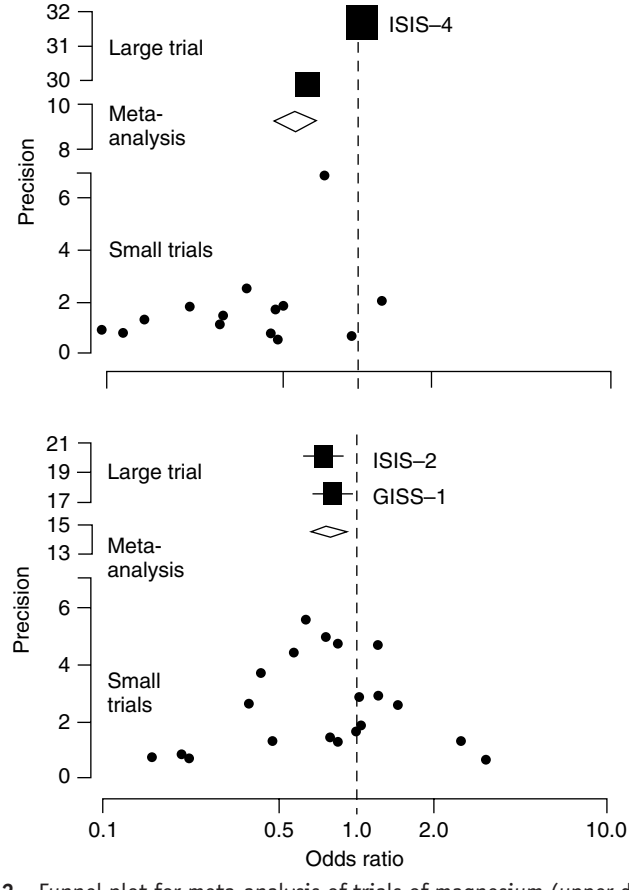


Figure 15.2 Funnel plot for meta-analysis of trials of magnesium (upper diagram) and streptokinase (lower diagram) in acute myocardial infarction. (Smith and Eggar (1997). Reproduced with kind permission from *The Lancet*.)

Example 15.4: (Continued)

indicates the possibility of publication bias and suggests that small trials with negative results have been under reported. The associated meta-analysis, shown as the diamond, gives a positive result overall for the test treatment. Following the meta-analysis there were two trials conducted and these are shown above the meta-analysis diamond as squares. The larger of these, ISIS-4 failed to show a benefit of intravenous streptokinase supporting the presence of bias in the earlier meta-analysis.

In contrast, the lower part of Figure 15.2 shows a funnel plot for trials evaluating streptokinase and the meta-analysis, which combined these trials. The pattern of the individual trial results supports the absence of publication bias and the two large trials shown in the plot, ISIS-2 and GISS-1 show treatment effects entirely in line with the earlier meta-analysis.

15.5 Meta-analysis in a regulatory setting

15.5.1 Retrospective analyses

In a regulatory setting a pre-planned meta-analysis is always going to be more convincing. Often, however, a meta-analysis will only be envisaged either part way through a development plan or at the end of the trial activity once the submission is being put together. It is interesting to note that within the regulatory context, meta-analysis has frequently caused problems for regulators: '*Meta-analysis has long been a source of regulatory discomfort, mostly because of the poor quality of some meta-analyses submitted in applications for licences*' (Lewis (2002)).

The regulators do, however, recognise that pre-planning is not always possible.

CPMP (2001): 'Points to Consider on Application with 1. Meta-Analysis; 2. One Pivotal Study'

'A retrospective specification when the results from all or some of the studies are known should be avoided . . . There are, however situations where the need for a meta-analysis becomes apparent after the results from some or sometimes all studies are known. This is the case when there is a need to put seemingly conflicting results into perspective, or in the exceptional situation where a meta-analysis seems to be the only way to provide reliable proof of efficacy'

It is still important even in this retrospective setting to write a protocol so that the meta-analysis can be performed as objectively as possible. The CPMP Points to Consider paper lists the prerequisites that are needed for such a retrospective analysis.

CPMP (2001): 'Points to Consider on Application with 1. Meta-Analysis; 2. One Pivotal Study'

'Prerequisites for a retrospective meta-analysis to provide sufficient evidence for a claim include:

- *Some studies clearly positive*
- *Inconclusive studies showing positive trends in the primary variable*
- *No statistically significant heterogeneity*
- *Pooled 95 per cent confidence interval well away from zero (or unity for odds ratios, or the pre-defined margin for non-inferiority trials)*
- *A justification that a biased selection of studies and/or endpoints is unlikely*
- *A sensitivity analysis demonstrating robustness of the findings*

For meta-analyses where these requirements are not fulfilled it will prove difficult to get a regulatory acceptance'

15.5.2 One pivotal study

A discussion of the two pivotal trial rule and under what conditions sponsors may be allowed to deviate from that requirement is included in the CPMP (2001) 'Points to Consider on Application with 1. Meta-Analysis; 2. One Pivotal Study' paper that covers meta-analysis, as there are some common issues.

The repeatability of particular findings gives strong scientific support to the existence of a true treatment effect and is essentially the reason why the regulators in general like to see two pivotal trials that are positive, this clearly then constitutes convincing evidence. In conjunction with this, the two trial rule provides an opportunity to investigate the treatment effect in different settings and a demonstration of an effect in both trials adds support to the robustness of that treatment effect. The policy of running two separate trials with the same protocol by simply dividing up the centres is not consistent with this thinking and should be avoided.

The two trial rule, for example in a placebo controlled setting, effectively translates into a very stringent requirement for statistical significance. In a single trial the conventional two-sided type I error rate is 0.05. It follows that in order

to obtain a positive result from such a trial we need the effect to be statistically significant and in favour of the active treatment. The type I error associated with this false positive result is 0.025 (which is 1 in 40). In two trials, therefore, obtaining two false positive results carries a combined false positive rate of $0.025 \times 0.025 = 0.000625$ (which is 1 in 1600). In other words if the active treatment were to be truly ineffective then on only 1 in 1600 occasions would we see two positive trials by chance.

In therapeutic settings where there are practical reasons why two trials cannot be easily undertaken or where there is a major unfulfilled public health need, it may be possible for a claim to be based on a single pivotal trial. The regulatory authorities do allow this, but only under certain conditions.

CPMP (2001): 'Points to Consider on Application with 1. Meta-Analysis; 2. One Pivotal Study'

'In cases where the confirmatory evidence is provided by one pivotal study only, this study will have to be exceptionally compelling, and in the regulatory evaluation special attention will be paid to:

- *The internal validity. There should be no indications of a potential bias*
- *The external validity. The study population should be suitable for extrapolation to the population to be treated*
- *Clinical relevance. The estimated size of the treatment benefit must be large enough to be clinically valuable*
- *The degree of statistical significance. Statistical evidence considerably stronger than* $p < 0.05$ *is usually required . . .*
- *Data quality*
- *Internal consistency. Similar effects demonstrated in different pre-specified sub-populations. All important endpoints showing similar findings*
- *Centre effects. None of the study centres should dominate the overall result, neither in terms of number of subjects nor in terms of magnitude of effect*
- *The plausibility of the hypothesis tested*

Statistical evidence stronger than $p < 0.05$ is open to interpretation but certainly $p < 0.000625$ would be a lower bound on this. In practice, the precise value would depend on the therapeutic setting, and likely remain unspecified.

In all regulatory settings, the regulators are looking for demonstration of a robust treatment effect. Within the context of the intended label there needs to be clear evidence that the treatment is effective in all subpopulations; age groups, disease

severity, race, sex (if appropriate) and so on, and this is what is meant by internal consistency. The two-trial rule gives an opportunity to evaluate the treatment across two different settings, for example different hospital types and different geographies. A single trial will only provide a similar level of assurance if it recruits across the broad range of settings, consistent, with the label followed by a thorough demonstration of the homogeneity of the treatment effect across those settings.

16

The role of statistics and statisticians

16.1 The importance of statistical thinking at the design stage

A clinical trial is an experiment and not only do we have to ensure that the clinical elements fit with the objectives of the trial, we also have to design the trial in a tight scientific way to make sure that it is capable of providing valid answers to the key questions in an unbiased, precise and structured way. This is where the statistics comes in and statistical thinking is a vital element of the design process for every clinical trial.

The following list of areas where statistical thinking is required is not exhaustive, but is meant to give a flavour of the sorts of things that need to be considered:

- What are the key prognostic factors and how, if at all, should these be used to stratify the randomisation?
- Should the randomisation be stratified by centre or by some higher-level factor, for example region or country?
- Would it be appropriate to centralise the randomisation given the nature (for example, complete blinding, partial blinding, no blinding) and complexity of the trial?
- What are the implications for block size in terms of ensuring balance and to prevent inadvertent unblinding?
- How can we choose primary and secondary endpoints in line with the clinical objectives for the trial?

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay
 ISBN 978-0-470-31971-0

- What methods can be used to control variability in order to increase precision?
- Which objectives form part of the confirmatory strategy for the trial and which elements are purely exploratory?
- What statistical testing strategy will provide valid answers to the range of questions being asked, particularly in terms of controlling multiplicity in the confirmatory setting?
- For each of the comparisons being considered is the focus superiority, equivalence or non-inferiority and in the latter two cases how can we choose Δ?
- Is it appropriate to build in an interim analysis given the nature of the trial; is this practical and, if so, how should the interim analysis be defined?
- How many patients are needed to provide answers to the questions being asked and what are the assumptions upon which this calculation is based?
- Do we need to revisit the sample size calculation at some interim stage in the trial?
- What are the implications of the trial procedures on the potential dropout rate and the extent of missing data?
- What impact are the dropouts and missing data likely to have on the definition of analysis sets and in particular our ability to align with the principle of intention to treat?
- Overall, will the trial provide an unbiased and precise estimate of the true treatment effect?

These particular points relate to each individual trial, but equally there will be similar considerations needed at the level of the development plan. In order for the overall, ordered programme of clinical trials to be scientifically sound there needs to be a substantial amount of commonality across the trials in terms of endpoints, definitions of analysis sets, recording of covariates and so on. This will facilitate the use of integrated summaries and meta-analysis for the evaluation and presentation of the complete programme or distinct parts of that programme, and outside of that, will allow a consistency of approach to the evaluation of the different trials.

At both the trial level and the development plan level, statisticians should take time to review the case report forms (CRFs) to make sure, in particular, that the data being collected will be appropriate for the precise, unambiguous and unbiased measurement of primary and secondary endpoints. Other aspects of the data being collected should also be reviewed in light of the way they will be used in the analysis. For example, baseline data will form the basis of covariates to be used in any adjusted analyses, intermediate visit data may be needed for the use of

certain imputations, for example, LOCF and data recorded for the determination of protocol violations will be used to define the per-protocol set.

16.2 Regulatory guidelines

Statistical thinking and practice is very much determined by the regulatory guidelines that are in place. Primarily it is ICH E9 'Statistical Principles for Clinical Trials', published in 1998, which sets down the broad framework within which we operate. In 2001 we saw the publication of ICH E10 'Choice of Control Group' which contained advice on the appropriate choice of concurrent control group and in particular first introduced the concept of assay sensitivity (see Section 12.5) in active control, non-inferiority trials.

Since that time we have seen numerous additional guidelines on specific statistical issues, for example the European (CPMP/CHMP) Points to Consider Papers:

CPMP (2000) Points to Consider on Switching between Superiority and Non-Inferiority
This guideline spelt out the circumstances where it is possible to change the objective of a trial from non-inferiority to superiority if the evidence is sufficiently strong, but clearly stated that switching in the opposite direction would unlikely be possible (see Section 12.9).

CPMP (2001) Points to Consider on Applications with 1. Meta Analysis; 2. One Pivotal Study
This guideline defined the role of meta-analysis within a regulatory submission (see Section 15.5.1) and indicated the circumstances where a single pivotal trial might be acceptable as the basis for registration (see Section 15.5.2).

CPMP (2001) Points to Consider on Missing Data
The use of various procedures for dealing with missing data, such as LOCF, in conjunction with the choice of analysis sets was covered in this guideline (see Section 7.3). The guideline also contained strong recommendations for avoiding missing data by the thoughtful choice of aspects of trial design.

CPMP (2002) Points to Consider on Multiplicity Issues in Clinical Trials
General aspects of multiple testing were considered in this guideline together with discussion on adjustment of significance levels or specific circumstances where adjustment is not needed (see Chapter 10).

CPMP (2003) Points to Consider on Adjustment for Baseline Covariates
It was in this guideline where the use of dynamic allocation was discouraged (see Section 1.4.6). The guideline also covered issues associated with the inclusion of

covariates, the handling of centre effects in the analysis and the investigation of treatment-by-covariate and treatment-by-centre interactions (see Sections 5.3.1, 5.5 and 6.7).

CHMP (2005) Guidance on the Choice of Non-Inferiority Margin
This much awaited guideline provided some general considerations for the choice of the non-inferiority margin. These considerations were not specific, but nonetheless have given us a way of thinking about the choice (see Section 12.7).

CHMP (2005) Guidance on Data Monitoring Committees/ FDA (2006) Establishment and Operation of Clinical Trial Data Monitoring Committees
These documents provided guidance on the set up, operational and working procedures, and the roles and responsibilities of the DMC in a single clinical trial or collection of trials (see Section 14.4).

CHMP (2006) Reflection Paper on Methodological Issues in Confirmatory Clinical Trials with Flexible Design and Analysis Plan
This document has set down some initial thoughts from a regulatory point of view about the issues involved in allowing the design of a clinical trial to be adapted as the trial progresses. Modification of the sample size based on blinded data and stopping for overwhelming efficacy or futility are forms of adaptation that are already well accepted, but this Reflection Paper considers other possibilities that are more controversial.

The regulatory framework is forever changing and undoubtedly our statistical methodology will itself change to meet these new requirements. A new CHMP statistics guideline is planned, for example, in relation to conditional approval while the FDA are considering more therapeutic-specific recommendations in relation to the choice of the non-inferiority margin.

Statistics as a subject is also changing, and new innovations will impact ultimately on regulatory thinking. Currently there is considerable interest on issues associated with adaptive/flexible designs (Phillips and Keane (2006)) and these theoretical considerations will influence the thinking of regulators who need to be sure that certain properties, such as the preservation of type I error, are taken care of. The Reflection Paper mentioned above represents a snapshot regarding current regulatory thinking in this area. As our experience increases and as the theoretical aspects become resolved, appropriate ways of thinking will become established. Another topic that has received a considerable amount of attention in recent years is missing data (see Liu *et al.* (2006) for a review and further references) and perhaps regulatory thinking with regard to this will change as we gain experience in applying more novel approaches.

Other general guidelines such as CHMP (2005) 'Guideline on Clinical Trials in Small Populations' contain some statistical considerations. This particular guideline discusses the relaxation of statistical requirements in this setting including the possibility of a less stringent significance level at the trial level, use of surrogate endpoints, analysis of data incorporating baseline covariates and the use of meta-analysis as the primary evidence of efficacy.

Both the EMEA and the FDA have recognised the need to streamline the drug development process in order to bring new medicines to patients more rapidly; see for example FDA (2004) 'Critical Path Initiative'. The FDA raise (FDA (2006) 'Critical Path Opportunities List') a number of statistical issues that need to be resolved in order to help make the clinical trials process more efficient:

- In the design of active control trials what data should be used to estimate the effect of the active control? Should we look at all previous placebo-control trials, how do we deal with inconsistent results and so on?
- What can be allowed in relation to adapting the trial design based on unblinded data?
- How should we handle missing data; what alternatives exist to LOCF?
- How to deal with various aspects of multiple endpoints, such as the requirement for success on more than a single endpoint; how to deal with correlated endpoints.

Finally most therapeutic specific guidelines contain recommendations that directly impact on statistical considerations, for example in terms of the definition of endpoints, the requirement for more than one primary endpoint, the definition of analysis sets and the choice of Δ. In a particular therapeutic setting it is self-evident that the requisite guidelines should be studied carefully in order to extract relevant information for statistical aspects of design and analysis.

16.3 The statistics process

We have already discussed the role that statistics and statisticians play in the design of clinical trials and programmes of clinical trials. In this section we will look at the manifestation of that planning in terms of the statistical methods section of the protocol and following on from that what happens from a statistical standpoint once the trial is ongoing through to the final reporting of that trial and the regulatory package as a whole.

16.3.1 The Statistical Methods section of the protocol

The Statistical Methods section of the protocol sets down the main aspects of both design and analysis. In particular this section should contain:

- Justification of the sample size (including the possible re-evaluation of sample size once the trial is ongoing)
- Method of randomisation (although block size will not be specified)
- Clear definition and delineation of the primary and secondary endpoints and how these link with the objectives of the trial
- Which aspects of the analysis will be viewed as confirmatory and which will be viewed as exploratory
- How multiplicity will be dealt with within the confirmatory part of the analysis
- Definition of analysis sets (full analysis set, per-protocol set, safety set)
- How missing data will be handled
- Detail regarding the methods of analysis for the primary endpoint(s) including specification of covariates to be the basis of any adjusted analyses
- Overview of statistical methods for the analysis of secondary endpoints
- Methods for handling the safety and tolerability data; the safety data will usually be coded using the MedDRA (*Med*ical *D*ictionary for *R*egulatory *A*ctivities, www.meddramsso.com) coding system in order to aid summary and presentation
- Interim analyses and how the type I error is to be protected within these
- Software to be used for statistical analysis

Only methods set down in the protocol can be viewed as confirmatory and so it is very important to get this section right; mistakes can be costly.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Only results from analyses envisaged in the protocol (including amendments) can be considered as confirmatory.'

16.3.2 The statistical analysis plan

The *statistical analysis plan (SAP)* is a more detailed elaboration of the statistical methods of analysis contained in the protocol. The SAP is written as the trial

is ongoing, but before database lock and the breaking of the treatment codes to unblind those involved in analysing the data. The SAP for studies which are unblinded should be finalised before the statistics team involved in analysing and reporting the data have access to any part of that data.

The SAP will also often contain table templates that allow the precise way in which the statistical analysis will be presented to be set down well in advance of running the analyses on the final trial data.

16.3.3 The data validation plan

Once the CRF is finalised the data management team will be putting together a *validation plan* which will set down the data checks that will be made in conjunction with the data entry process; for example, are the visit dates in chronological order, are the ages of the patients within the range specified in the inclusion criteria, and so on. It is useful for this plan to be reviewed by a statistician for two reasons and especially in terms of issues that relate to the definition of endpoints. Firstly, it is important that the statistics team are aware of what data checks are being undertaken so that at the data analysis stage they can rule out certain potential data problems and be assured of a certain level of data 'quality'. Secondly, the statistician may be able to suggest other specific checks that will help to increase the quality of the data.

16.3.4 The blind review

There is one final opportunity to revisit the proposed methods of statistical analysis prior to the breaking of the blind, or in an unblinded trial, before the statistics group have seen study data. This so-called *blind review* usually takes place around the time of database lock and the following lists some of the aspects of analysis that would generally be considered:

- Precise definition of analysis sets; specifically which patients are to be included and which are to be excluded
- Handling of missing data; finalisation of the algorithms
- Finalisation of algorithms for combining centres should this be required
- Outlier identification and specific decisions taken on how these are to be handled

Under normal circumstances the blind review should take place over a 24 or 48 hour period to limit delays in beginning the analysis proper. The blind review should be documented, detailing precisely what was done.

Sometimes the blind review can throw up data issues that require further evaluation by the data management group with data queries being raised, and these perhaps may result in changes to the database. This sequence of events can cause major headaches and delays in the data analysis and reporting, and so it is important in the planning phase to get the data validation plan correct so that issues are identified and dealt with in an ongoing way.

16.3.5 Statistical analysis

The SAP will have detailed the precise methods of analysis and presentation and should ideally be finalised well before database lock. This enables work to begin in good time on the programming of the analyses. These programs will be tested on 'dirty' data from the trial, so that they can be pretty much finalised before the trial ends, enabling, at least in theory, rapid turnaround of the key analyses.

This is not always as simple as it sounds. In particular, working with dirty data can bring its own problems, including illogical data values that the programs cannot handle. Also, when the final data arrives there may be specific issues and data problems arising that were never picked up at the earlier runs of the programs. Nonetheless, these aspects of planning and program development and validation are essential if we are going to be in a position to complete the statistical analyses and presentations quickly. Also, working with the database in an ongoing way can avoid any surprises occurring following database lock.

The analyses and tables will be a joint effort involving statisticians and statistical programmers. Quality control is an essential component of this part of the process and double programming is frequently undertaken, that is, every analysis and all table entries are reproduced independently by a second programmer and cross-checked against the original. Data listings will also be produced and checked, although the level of checking may not be as rigorous as with the tables. Figures and graphs require a different kind of QC, but certainly the points on these figures and graphs should be verified independently by a second programmer.

16.3.6 Reporting the analysis

ICH E3(1995) 'Structure and Content of Clinical Study Reports' sets down the structure, down to the numbering of the sections and precisely what goes in each of those sections, required within the regulatory setting for reporting each study. Medical writers will work with statisticians to put these reports together.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'. . . statistical judgement should be brought to bear on the analysis, interpretation and presentation of the results of a clinical trial. To this end the trial statistician should be a member of the team responsible for the clinical study report, and should approve the clinical report.'

There will be a number of areas of the clinical report where the statistician will contribute but, in particular, Section 11 'Efficacy Evaluation' and Section 12 'Safety Evaluation' will require statistical oversight. Section 16 of the report contains the Appendices and Subsection 16.1.9 entitled 'Documentation of Statistical Methods' will usually be written by the trial statistician.

Within Section 11, Subsection 11.4.2 entitled 'Statistical/Analytical Issues' contains a series of items covering many of the areas of complexity within most statistical analyses:

- Adjustment for covariates
- Handling of dropouts or missing data
- Interim analyses and data monitoring
- Multi-centre studies
- Multiple comparisons/multiplicity
- Use of an 'efficacy subset' of patients
- Active control studies intended to show equivalence
- Examination of subgroups

Each of these will clearly require input from the statistician.

16.3.7 Pre-planning

A common theme running across almost everything that we do within statistics is the need for pre-planning.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

For each clinical trial contributing to a marketing application, all important details of its design and conduct and the principle features of its proposed analysis should be clearly specified in a protocol written before the trial begins. The extent to which the procedures in

the protocol are followed and the primary analysis is planned a priori will contribute to the degree of confidence in the final results and conclusions of the trial.'

This pre-planning in terms of both design and analysis is predominantly set down in the trial protocol. Pre-planning is one key aspect of the way we design and run our trials that helps to reduce bias. It would be entirely inappropriate to take decisions about methods of analysis based on unblinded looks at the data. Pre-planning also enables us to think through in advance just how we are going to handle the data. This is good discipline and can help us to anticipate problems in advance. A final benefit of pre-planning is a very practical one. Once the trial is complete and the database is locked there is inevitably a 'mad dash' to analyse the data and look at the results. Only with pre-planning and an effective amount of pre-programming and testing of those programs can the statistical analyses be undertaken quickly and without major hitches.

As the trial is ongoing there is also an opportunity to change some of the planned methods of analysis; for example, information that a particular covariate could be important or that a different kind of effect could be seen in a certain subgroup may have become available based on external data from a similar trial that has now completed and reported. Such changes can be incorporated by modifying the SAP and if they represent major changes to the analysis, for example if they were associated with the analysis of the primary endpoint, then a protocol amendment would need to be issued. The reason for this, as mentioned earlier, is that only methods specified in the protocol can be viewed as confirmatory.

In a limited way, there may also be changes in the design as the trial is ongoing, for example resizing of the trial. Such changes represent major design modifications and protocol amendments would be needed to ensure that the modifications fall within what could be considered as pre-planning. As pointed out in ICH E9(1998) in relation to revising the sample size:

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'A revised sample size may then be calculated using suitably modified assumptions, and should be justified and documented in a protocol amendment and in the clinical study report . . . The potential need for re-estimation of the sample size should be envisaged in the protocol whenever possible.'

The final sentence here again emphasises the need for pre-planning with regard to this process wherever possible.

A change in the statistical methods at the data analysis stage, for example including unplanned covariates or using a transformation of the primary endpoint

when one was not planned, would usually be unacceptable. The danger here is that a method may have been chosen that affects the resulting magnitude of the treatment effect. The choice of statistical method should be pre-specified in the statistical analysis plan, and possibly modified at the blind review. Changes to these methods could be acceptable, however, in conjunction with a clearly defined algorithm. For example, the logrank test may be the planned method of analysis for survival data, but if the assumption of proportional hazards is not valid according to some well-defined assessment of that assumption, then treatment comparisons could be based on the Gehan–Wilcoxon test. Alternatively, it can be stated in the protocol that if, on visual inspection of the normal probability plot, the data appears to be positively skewed then the log transformation will be used to recover the normality of the data and stabilise the variance. These would be examples of clearly defined algorithms leading to a well-defined method of analysis for the calculation of the p-value. Of course 'visual inspection' contains an element of subjectivity, but nonetheless regulators can see a clear way through the decision-making process.

A related issue concerns new questions that may arise during the analysis of data. These aspects should be clearly distinguished and would constitute only exploratory analyses.

ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'

'Although the primary goal of the analysis of a clinical trial should be to answer the questions posed by its main objectives, new questions based on the observed data may well emerge during the unblinded analysis. Additional and perhaps complex statistical analysis may be the consequence. This additional work should be strictly distinguished in the report from work which was planned in the protocol.'

16.3.8 Sensitivity and robustness

Statisticians and regulators alike, quite rightly, place great store on robustness and sensitivity analyses. All analyses will be based on certain assumptions regarding the data, such as normality and constant variance, or independent censoring in time-to-event data. Analyses could be potentially affected by the presence of single outlying data points or sensitive to the choice of analysis sets or the handling of missing data. It would be very unsatisfactory if the conclusions drawn from the data were driven by assumptions that were questionable or were unduly influenced by different choices for dealing with specific aspects of the data. Throughout the statistical analysis the sensitivity of the conclusions to assumptions of the kind discussed should be evaluated. The regulators mention these aspects on several

occasions and in relation to numerous aspects of analysis and interpretation. Here are just a few:

- Choice of analysis sets
  ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'
  'In general it is advantageous to demonstrate a lack of sensitivity of the principal trial results to alternative choices of the set of subjects analysed.'

- Missing data
  ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'
  'An investigation should be made concerning the sensitivity of the results of analysis to the method of handling missing values, especially if the number of missing values is substantial.'

- Outliers
  The following quote follows on from that on missing data:
  ICH E9 (1998): 'Note for Guidance on Statistical Principles for Clinical Trials'
  'A similar approach should be adopted to exploring the influence of outliers . . . If no procedures for dealing with outliers was foreseen in the trial protocol, one analysis with the actual values and at least one other analysis eliminating or reducing the outlier effect should be performed and differences between the results discussed.'

In certain cases more specific guidance is given. The FDA, for example, discuss various sensitivity analyses in relation to the analysis of progression-free survival in FDA (2005) 'Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics'.

A general message coming out of this is that, where there are doubts regarding how best to handle specific aspects of the data at the analysis stage, consider a range of different approaches, say two or three, and hopefully demonstrate that the conclusions are unaffected by the approach adopted. If, however, they are affected by this choice then this lack of robustness could undermine the validity of the conclusions drawn.

16.4 The regulatory submission

The statistics group or groups involved in analysing and reporting each of the trials will have a role in compiling the regulatory submission. Both the ISS (integrated summary of safety) and the ISE (integrated summary of efficacy) will involve the analysis and presentation of compiled results across the whole programme of trials. Formal meta-analysis may be employed or alternatively a pooling of data, and the technical aspects of these methods were discussed in Chapter 15.

Once the regulatory submission has been made there will inevitably be questions and issues coming back from the regulators. There may be concerns about the way the data has been handled from a statistical point of view. There may be requests for additional specific analyses to resolve uncertainty. There may be more open issues that the regulators are unhappy with which may require a substantial amount of further analysis. At the extreme end there may be outright rejection and the company may then be back to the drawing board in terms of the product. In all of these cases there will usually be a need for further statistical considerations.

If certain additional analyses have been specifically requested, then providing these should be fairly straightforward. If the questions, however, are more general, then the company may need to respond by providing a series of re-analyses to address the issues. In this case the concept of pre-planning is irrelevant, those deciding on what further analyses to present are unblinded to the data. This scenario of itself creates difficulties. There is a temptation to re-analyse in a number of different ways, but only present back to the regulators those analyses that support the company's position. The best way to proceed in order to avoid potential criticism is to be open with the regulators and present a wide range of analyses that fit with the questions and issues being raised.

In the US the FDA request within the submission an electronic version of the database and this gives them the opportunity to not only re-analyse the data to confirm the results presented within the submission, but also to perform their own alternative analyses. This does not happen in Europe. In the US, therefore, the process following submission is somewhat different and much of the interchange in terms of requesting and supplying alternative analyses is taken care of by the FDA statisticians.

16.5 Publications and presentations

Outside of the clinical report and regulatory setting there will clearly be the need to publish the results of trials in the medical literature and to make presentations at conferences.

In recent years there have been a range of recommendations regarding the structure of publications; how they should be laid out and what they should contain. These have usually been in the form of checklists and all of this has been encapsulated within the CONSORT statement (Moher *et al.* (2001) and Altman *et al.* (2001)). CONSORT is an acronym for *Con*solidated *S*tandards *o*f *R*eporting *T*rials and increasingly many medical journals have adopted this guidance in terms of requiring their clinical trial publications to conform to it. There is a web site which provides up-to-date information and helpful resources and examples; www.consort-statement.org.

The guideline splits down the content of each publication into a series of items, 22 in total, ranging from

1. Title and Abstract
2. Introduction – Background
3. Methods – Participants
4. Methods – Interventions

through to

7. Methods – Sample Size
13. Results – Participant Flow

to

22. Discussion – Overall Evidence

The precise content is too detailed to give complete coverage here, but just to give an impression we will consider two areas; the sample size calculation (item 7) and participant flow (item 13).

The sample size calculation should be detailed in the trial publication, indicating the estimated outcomes in each of the treatment groups (and this will define, in particular, the clinically relevant difference to be detected), the type I error, the type II error or power and, for a continuous primary outcome variable in a parallel group trial, the within-group standard deviation of that measure. For time-to-event data details on clinically relevant difference would usually be specified in terms of the either the median event times or the proportions event-free at a certain time point.

An important aspect of the reporting of any clinical trial is a clear indication of what happened to all of the patients randomised. CONSORT recommends that each publication should contain a diagram showing the flow of participants through the trial; numbers randomised to each of the treatment groups, receiving intended treatment, protocol deviations by treatment group classified by type of deviation and patient groups analysed for primary outcomes. See Figure 16.1 for an example of this.

Example 16.1: Natalizumab plus interferon beta-1a for relapsing multiple sclerosis; the SENTINEL study

Figure 16.1 provides the participant flow in relation to this placebo-controlled trial of natalizumab. (Rudick *et al.* (2006)).

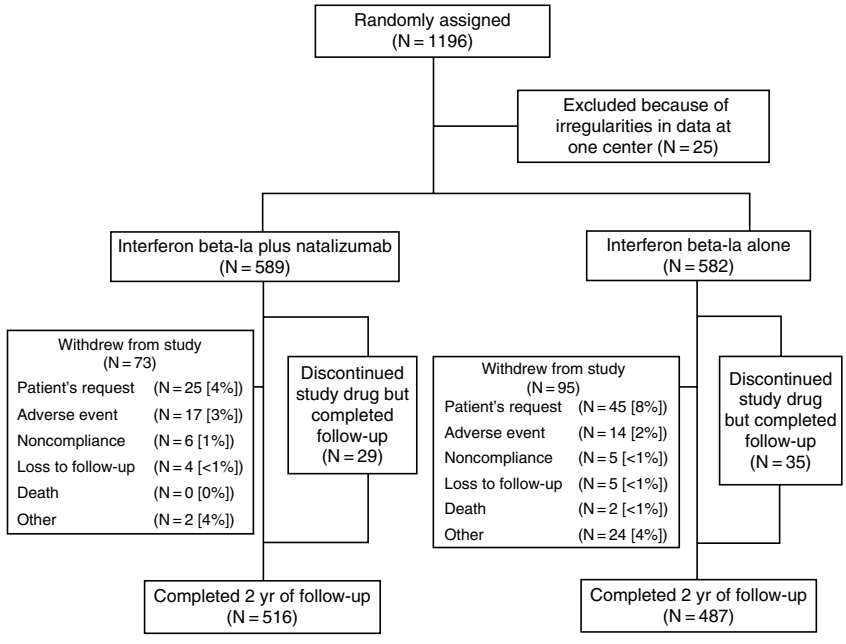


Figure 16.1 Patient disposition in the SENTINEL trial. (Rudick RA Stuart WH, Calabresi PA, Confavreux C *et al.* for the SENTINEL Investigators (2006) 'Natalizumab plus interferon beta-1a for relapsing multiple sclerosis' New England Journal of Medicine, 354, 911–923. © (2006) Massachusetts Medical Society.)

The quality of publications is certainly increasing and in part this is due to guidance of the type already described in this section. It is unfortunately the case, however, that many mistakes are still made, even in leading journals, despite apparently rigorous refereeing procedures. Particular areas of statistics seem to cause consistent difficulty:

- The correct design, analysis and interpretation of non-inferiority trials – remember conventional *p*-values have no role
- Conforming to the principle of intention-to-treat to avoid bias – ITT means all randomised subjects or something very close to that
- Incorrect analysis of time-to-event data in terms of the definition for the origin of the measurement – the point of randomisation is the *only* origin that can be used in a randomised trial
- Adjusting the analysis for baseline factors – with change from baseline as the outcome variable, include baseline as a covariate to avoid regression

towards the mean and note that it is incorrect to use covariates that are measured after randomisation

- The correct statistical methods for combining data in meta-analysis – the summary statistics that are combined must come from independent data sets

It is important to enlist the help of statistical colleagues when putting publications together, not only in terms of the actual analysis, but in terms of the interpretation and reporting in the publication itself. Further, do cast a critical eye over the statistical methodology in the papers you review in order to spot these problem areas and again request the help of your statistical colleagues.

Presentations at conferences do not, of course, go through the same critical review process as publications. Even though abstracts are often submitted and reviewed in advance, mistakes and bad practice will still slip through. It is important to have statistical input when putting these presentations together. Errors in the statistics will invariably get picked up by some members of the audience and the resulting bad press could well be damaging. From the opposite perspective look critically at what is being presented in terms of the statistics and challenge if you feel that inappropriate methods are being used.

References

Altman DG (1991) *Practical Statistics for Medical Research* London: Chapman & Hall

Altman DG (1998) 'Confidence intervals for the number needed to treat' *British Medical Journal*, **317**, 1309–1312

Altman DG, Schultz KF, Moher D, Egger M *et al.* for the CONSORT Group (2001) 'The revised CONSORT statement for reporting randomized trials: explanation and elaboration' *Annals of Internal Medicine*, **134**, 663–694

Arani RB, Soong S-J, Weiss HL, Wood MJ *et al.* (2001) 'Phase specific analysis of herpes zoster associated pain data: a new statistical approach' *Statistics in Medicine*, **20**, 2429–2439

Bedikian AY, Millward M, Pehamberger H, Conry R *et al.* (2006) 'Bcl-2 Antisense (oblimersen sodium) plus dacarbazine in patients with advanced melanoma : The Oblimersen Melanoma Study Group' *Journal of Clinical Oncology*, **24**, 4738–4745

Bland M (2004) 'Cluster randomised trials in the medical literature: two bibliometric surveys' *BMC Medical Research Methodology*, **4**, 21

Bolland K and Whitehead J (2000) 'Formal approaches to safety monitoring of clinical trials in life-threatening conditions' *Statistics in Medicine*, **19**, 2899–2917

Breslow NE and Day NE (1994) *Statistical Methods in Cancer Research, Volume II: The Design and Analysis of Cohort Studies* IARC Publications, 82, New York: Oxford University Press

Brodie MJ, Richens A, Yuen AWC, for UK Lamotrigine/Carbamazepine Monotherapy Trial Group (1995) 'Double-blind comparison of lamotrigine and carbamazepine in newly diagnosed epilepsy' *The Lancet*, **345**, 476–479

Brodie MJ and Whitehead J (2006) 'Active control comparisons: the ideal trial design' *Epilepsy Research*, **68**, 70–73

Byar DP (1980) 'Why data bases should not replace randomised trials' *Biometrics*, **36**, 337–342

Campbell MJ, Donner A and Klar N (2007) 'Developments in cluster randomised trials and Statistics in Medicine' *Statistics in Medicine*, **26**, 2–19

Coronary Drug Project Research Group (1980) 'Influence and adherence to treatment and response of cholesterol on mortality in the coronary drug project' *New England Journal of Medicine*, **303**, 1038–1041

Cox DR (1972) 'Regression models and life tables (with discussion)' *Journal of the Royal Statistical Society, B*, **74**, 187–220

Crawford ED, Eisenberger MA, McLoed DG, Spaulding JT *et al.*(1989) 'A controlled trial of leuprolide with and without flutamide in prostatic cancer' *New England Journal of Medicine*, **321**, 419–424

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay

ISBN 978-0-470-31971-0

Ebbutt AF and Frith L (1998) 'Practical issues in equivalence trials' *Statistics in Medicine*, **17**, 1691–1701

Eggar M and Smith D (1998) 'Meta-analysis bias in location and selection of studies' *British Medical Journal*, **316**, 61–66

Eggar M, Smith GD, Schneider M and Minder C (1997) 'Bias in meta-analysis detected by a simple, graphical test' *British Medical Journal*, **315**, 629–634

Ellenberg SS, Fleming TR and DeMets DL (2003) '*Data Monitoring Committees in Clinical Trials: A Practical Perspective*' New York: John Wiley & Sons, Inc.

Fleiss JL (1993) 'The statistical basis of meta-analysis' *Statistical Methods in Medical Research*, **2**, 121–145

Fleming TR and DeMets DL (1996) 'Surrogate end points in clinical trials: Are we being misled?' *Annals of Internal Medicine*, **125**, 605–613

Ford I, Norrie J and Ahmedi S (1995) 'Model inconsistency, illustrated by the Cox Proportional Hazards model' *Statistics in Medicine*, **14**, 735–746

Gardner MJ and Altman DG (1989) 'Estimation rather than hypothesis testing: confidence intervals rather than *p*-values' In: *Statistics with Confidence* (eds MJ Gardner and DG Altman), London: British Medical Journal, 6–19

Gehan EA (1969) 'Estimating survival functions from the life table' *Journal of Chronic Diseases*, **21**, 629–644

Gillings D and Koch G (1991) 'The application of the principle of intention-to-treat analysis of clinical trials' *Drug Information Journal*, **25**, 411–424

Greenwood M (1926) 'The errors of sampling of the survivorship tables' *Reports on Public Health and Statistical Subjects*, No. 33, Appendix 1. London: HMSO

Grieve AP (2003) 'The number needed to treat: a useful clinical measure or a case of the Emperor's new clothes?' *Pharmaceutical Statistics*, **2**, 87–102

Haybittle JL (1971) 'Repeated assessment of results in clinical trials of cancer treatment' *British Journal of Radiology*, **44**, 793–797

Helsinki Declaration (2004) 'Ethical Principles for Medical Research Involving Human Subjects' WMA General Assembly, Tokyo 2004

Hennerici MG, Kay R, Bogousslavsky J, Lenzi JM, Orgogozo M, for the ESTAT Investigators (2006) 'Intravenous ancrod for acute ischemic stroke in the European Stroke Treatment with Ancrod Trial: a randomised controlled trial.' *The Lancet*, **368**, 1871–1878

Jensen MP, Karoly P, O'Riordan EF, Bland F and Burns RS (1989) 'The subjective experience of pain. An assessment of the utility of 10 indices' *The Clinical Journal of Pain*, **5**, 153–159

Jones B, Jarvis P, Lewis JA and Ebbutt AF (1996) 'Trials to assess equivalence: the importance of rigorous methods' *British Medical Journal*, **313**, 36–39

Julious SA (2004) 'Using confidence intervals around individual means to assess statistical significance between two means' *Pharmaceutical Statistician*, **3**, 217–222

Julious SA (2005) 'Why do we use pooled variance analysis of variance?' *Pharmaceutical Statistics*, **4**, 3–5

Julious SA and Mullee MA (1994) 'Confounding and Simpson's Paradox' *British Medical Journal*, **309**, 1480–1481

Kaplan EL and Meier P (1958) 'Non-parametric estimation from incomplete observations' *Journal of the American Statistical Association*, **53**, 457–481

Kaul S and Diamond GA (2006) 'Good enough: a primer on the analysis and interpretation of non-inferiority trials' *Annals of Internal Medicine*, **145**, 62–69

Kay R (1995) 'Some fundamental statistical concepts in clinical trials and their application in herpes zoster' *Antiviral Chemistry and Chemotherapy*, **6**, **Supplement 1**, 28–33

Kay R (2004) 'An explanation of the hazard ratio' *Pharmaceutical Statistics*, **3**, 295–297

Kay R (2006) Letter to the Editor on 'Phase specific analysis of herpes zoster associated pain data: a new statistical approach' *Statistics in Medicine*, **25**, 359–360

Landis RJ, Heyman ER and Koch GG (1978) 'Average partial association in three-way contingency tables: A review and discussion of alternative tests' *International Statistical Review*, **46**, 237–254

Lewis JA (2002) 'The European regulatory experience' *Statistics in Medicine*, **21**, 2931–2938

Lewis JA (2004) 'In defence of the dichotomy.' *Pharmaceutical Statistics*, **3**, 77–79

Li Z, Chines AA and Meredith MP (2004) 'Statistical validation of surrogate endpoints: Is bone density a valid surrogate for fracture?' *Journal of Musculoskeletal Neuron Interaction*, *4*, 64–74

Liu M, Wei L and Zhang J (2006) 'Review of guidelines and literature for handling missing data in longitudinal clinical trials with a case study' *Pharmaceutical Statistics*, **5**, 7–18

Machin D, Campbell MJ, Fayers PM and Pinol APY (1997) *Statistical Tables for the Design of Clinical Trials* (2nd edn) Blackwell Scientific Publications, Oxford

Mantel N and Haenszel W (1959) 'Statistical aspects of the analysis of data from retrospective studies of disease' *Journal of the National Cancer Institute*, **22**, 719–748

Marshall RJ and Chisholm EM (1985) 'Hypothesis testing in the polychotomous logistic model with an application to detecting gastrointestinal cancer' *Statistics in Medicine*, **5**, 337–344

Matthews JNS, Altman DG, Campbell MJ and Royston P (1990) 'Analysis of serial measurements in medical research' *British Medical Journal*, **300**, 230–235

Meier P (1978) 'The biggest public health experiment ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine' In *Statistics, A Guide to the Unknown* (ed. J Tanur, F Mostellor *et al.*) San Fransisco: Holden Day

Miller DH, Khan OA, Sheremata WA, Blumhardt LD *et al.*, for the International Natalizumab Multiple Sclerosis Trial Group (2003) 'A controlled trial of natalizumab for relapsing multiple sclerosis' *New England Journal of Medicine*, **348**, 15–23

Moher D, Schulz KF and Altman DG (2001) 'The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials' *Annals of Internal Medicine*, **134**, 657–694

Mulrow CD (1994) 'Rationale for systematic reviews' *British Medical Journal*, **309**, 597–599

Nakamura H, Arakawa K, Itakura H, Kitabatake A *et al.*, for the MEGA Study Group (2006) 'Primary prevention of cardiovascular disease with pravastatin in Japan (MEGA study): a prospective randomised controlled trial' *The Lancet*, **368**, 1155–1163

Normand S-LT (1999) 'Meta-analysis: formulating, evaluating, combining, and reporting' *Statistics in Medicine*, **18**, 321–359

O'Brien PC and Fleming TR (1979) 'A multiple testing procedure for clinical trials' *Biometrics*, **35**, 549–556

Okwera A, Byekwaso F, Mugerwa R, Ellner J *et al.* Makerere University–Case Western University Research Collaboration (1994) 'Randomised trial of thiacetazone and rifampicin-containing regimens for pulmonary tuberculosis in HIV-infected Ugandans.' *The Lancet*, **344**, 1323–1328

Packer M, Coats AJS, Fowler MB, *et al.* for the Carvedilol Prospective Randomised Cumulative Survival Study Group (2001) 'Effect of carvedilol on survival in severe chronic heart failure' *New England Journal of Medicine*, **344**, 1651–1658

Patel K, Kay R and Rowell L (2006) 'Comparing proportional hazards and accelerated failure time models: An application in Influenza' *Pharmaceutical Statistics*, **5**, 213–224

Peto R and Peto J (1972) 'Asymptotically efficient rank invariant procedures.' *Journal of the Royal Statistical Society, A*, **135**, 185–207

Peto R, Pike MC, Armitage P, Breslow NE *et al.* (1976) 'Design and analysis of randomised clinical trials requiring prolonged observation of each patient. I. Introduction and design' *British Journal of Cancer*, **34**, 585–612

Phillips AJ and Keene ON on behalf of the PSI Adaptive Design Expert Group (2006) 'Adaptive Designs for Pivotal Trials' *Pharmaceutical Statistics*, **5**, 61–66

Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A *et al.*, for the Herceptin Adjuvant (HERA) Trial Study Team (2005) 'Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer' *New England Journal of Medicine*, **353**, 1659–1672

Pocock SJ (1977) 'Group sequential methods in the design and analysis of clinical trials' *Biometrika*, **64**, 191–199

Pocock SJ (1983) *Clinical Trials: A Practical Approach* Chichester: John Wiley & Sons, Ltd

Pocock SJ (2004) 'A major trial needs three statisticians: why, how and who?' *Statistics in Medicine*, **23**, 1535–1539

Pocock SJ, Clayton TC and Altman DG (2002) 'Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls' *The Lancet*, **359**, 1686–1689

Powderly WG, Saag MS, Cloud GA, Robinson P *et al.*, the NIAID AIDS Clinical Trials Group and the NIAID Mycosis Study Group (1992) 'A controlled trial of fluconazole or amphotericin B to prevent relapse of cryptococcal meningitis in patients with the acquired immunodeficiency syndrome' *New England Journal of Medicine*, **326**, 793–798

Roes KCB (2004) 'Dynamic allocation as a balancing act' *Pharmaceutical Statistics*, **3**, 187–191

Rubin LJ, Badesch DB, Barst RJ, Galie N *et al.* for the Bosentan Randomised Trial of Endothelin Antagonist Therapy Study Group (2002) 'Bosentan theraphy for pulmonary arterial hypertension' *New England Journal of Medicine*, **346**, 869–903

Rudick RA Stuart WH, Calabresi PA, Confavreux C *et al.* for the SENTINEL Investigators (2006) 'Natalizumab plus interferon beta-1a for relapsing multiple sclerosis' *New England Journal of Medicine*, **354**, 911–923

Sargent DJ, Goldberg RM, Jacobson SD, MacDonald JS *et al.* (2001) 'A pooled analysis of adjuvant chemotherapy for resected colon cancer in elderly patients' *New England Journal of Medicine*, **345**, 1091–1097

Schuirmann DJ (1987) 'A comparison of two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability' *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657–680

Senn S (1997) *Statistical Issues in Drug Development* Chichester: John Wiley & Sons, Ltd

Senn S (2002) *'Cross-Over Trials in Clinical Research (2nd edn)'* Chichester: John Wiley & Sons

Senn S (2003) 'Disappointing dichotomies.' *Pharmaceutical Statistics*, **2**, 239–240

Sherman DG, Atkinson RP, Chippendale T *et al.* (2000) 'Intravenous ancrod for treatment of acute ischemic stroke; the STAT study; a randomised controlled trial.' *Journal of the American Medical Association*, **283**, 2395–2403

Smith GD and Eggar M (1997) Letter to the Editor *The Lancet*, **350**, 1182

Stokes ME, Davis CS and Koch GG (1995) *Categorical Data Analysis Using the SAS System* Cary, NC: SAS Institute Inc.

Storosum JG, van Zwieten BJ, Vermeulen HDB, Wohlfarth T *et al.* (2001) 'Relapse and recurrence in major depression: a critical review of placebo-controlled efficacy studies with special emphasis on methodological issues' *European Psychiatry*, *16*, 327–335

Stutchfield P, Whitaker R and Russell I, on behalf of the Antenatal Steroids for Term Elective Caesarean Section (ASTECS) Research Team (2005) 'Antenatal betamethasone and incidence of neonatal respiratory distress after elective caesarean section: pragmatic randomised trial' *British Medical Journal*, **331**, 662–667

van Belle G, Fisher LD, Heagerty PJ and Lumley T (2004) *'Biostatistics: A Methodology for the Health Sciences (2nd edn)'* New Jersey: John Wiley & Sons

van Leth F, Phanuphak P, Ruxrungtham K, Baraldi E *et al.* for the 2NN Study Team (2004) 'Comparison of first-line antiretroviral therapy with regimens including nevirapine, efavirenz,

or both drugs, plus stavudine and lamivudine: a randomised open-label trial, the 2NN Study' *The Lancet*, **363**, 1253–1263
Westlake WJ (1981) 'Bioequivalence testing – a need to rethink (Reader Reaction Response)' *Biometrics*, **37**, 589–594
The Xamoterol in Severe Heart Failure Study Group (1990) 'Xamoterol in severe heart failure' *The Lancet*, **336**, 1–6

Regulatory Guidelines

ICH Guidelines

ICH E3 (1995) 'Note for Guidance on Structure and Content of Clinical Study Reports'
ICH E9 (1998) 'Note for Guidance on Statistical Principles for Clinical Trials'
ICH E10 (2001) 'Note for Guidance on Choice of Control Group in Clinical Trials'

FDA Guidelines

FDA (1992) 'Points to Consider on Clinical Development and Labeling of Anti-Infective Drug Products'
FDA (1998) 'Developing Antimicrobial Drugs – General Consideration for Clinical Trials'
FDA (2001) 'Statistical Approaches to Establishing Bioequivalence'
FDA (2004) 'Critical Path Initiative'
FDA (2005) 'Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics'
FDA (2006) 'Critical Path Opportunities List'
FDA (2006) 'Establishment and Operation of Clinical Trial Data Monitoring Committees'

European Statistics Guidelines

CPMP (2000) 'Points to Consider on Switching between Superiority and Non-Inferiority'
CPMP (2001) 'Points to Consider on Applications with 1. Meta Analyses; 2. One Pivotal Study
CPMP (2001) 'Points to Consider on Missing Data'
CPMP (2001) 'Note for Guidance on the Investigation of Bioavailability and Bioequivalence'
CPMP (2002) 'Points to Consider on Multiplicity Issues in Clinical Trials'
CPMP (2003) 'Points to Consider on Adjustment for Baseline Covariates'
CHMP (2005) 'Guidance on the Choice of Non-Inferiority Margin'
CHMP (2005) 'Guideline on Data Monitoring Committees'
CHMP (2005) 'Guideline on Clinical Trials in Small Populations'
CHMP (2006) 'Reflection Paper on Methodological Issues in Confirmatory Clinical Trials with Flexible Design and Analysis Plan' (Draft)

European Therapeutic Area Guidelines

CPMP (1997) 'Note for Guidance on Medicinal Products in the Treatment of Alzheimer's Disease'
CPMP (1997) 'Note for Guidance on Evaluation of New Anti-Bacterial Medicinal Products'
CPMP (1999) 'Note for Guidance on Clinical Evaluation of New Vaccines'

CPMP (2001) 'Note for Guidance on Clinical Investigation of Medicinal Products for the Treatment of Acute Stroke'

CPMP (2002) 'Note for Guidance on the Clinical Investigation of Medicinal Products in the Treatment of Asthma'

CPMP (2003) 'Note for Guidance on Evaluation of Medicinal Products Indicated for Treatment of Bacterial Infections'

CPMP (2003) 'Points to Consider on the Clinical Development of Fibrinolytic Medicinal Products in the Treatment of Patients with ST Segment Elevation Acute Myocardial Infarction (STEMI)'

CHMP (2006) 'Guideline on Similar Biological Medicinal Products Containing Biotechnology-Derived Proteins as Active Substance: Non-Clinical and Clinical Issues'

Index

Absolute risk reduction (ARR) 88
Accelerated failure time model 207–8, 210
Adaptive designs 223–7
Add-on trials 3
Adjusted analysis 63, 91–110, 142, 246, 247–8, 250, 253
  multiplicity 157
  survival data 204–8
Adjusted significance level 148, 149–51, 153, 155
Allergic rhinitis 181
Alpha error 127
Alpha-spending functions 153, 214
Alternative hypotheses 47, 59, 64, 93, 99
  equivalence 178–9
Alzheimer's disease 23, 120
Analysis of covariance (ANCOVA) 19, 63, 91–110, 142
  ANOVA 101–2, 109
  assumptions 104, 159, 163
  baseline factors 97, 102–3, 110
  continuous data 97–104, 106, 109, 204
  covariates 99–104, 108
  logistic regression 104–6
  non-parametric tests 169, 170
  survival data 194, 204
  transformations 164
Analysis of variance (ANOVA) 19, 108, 142, 204, 207
  ANCOVA 101–2, 109
  assumptions 159, 162
  non-parametric tests 169
  transformations 164
  see also Two-way analysis of variance
Analysis sets 124–6, 246, 249, 250, 256
  ITT 115–18, 121
  multiplicity 148, 158
Anti-infective trials 115, 116, 120, 137
Arithmetic means 164
Assumptions 114, 159, 160, 216, 246, 254–5
  ANCOVA 104, 159, 163
  homogeneity of variance 159, 160
  independent censoring 208–9
  non-inferiority 187–8
  non-parametric tests 168, 169, 170
  normality 159, 160–3, 166, 170
  proportional hazards 210
  sample size 138, 140, 162, 187–8
  t-tests 159, 160–3, 168, 170
Asthma trials 15, 19, 43, 58, 107, 149–50
Averages 27

Back-transformations 164
Balance 6, 8, 9–10, 61, 220
Baseline factors 91–2, 109–10, 122, 125, 223
  ANCOVA 97, 102–3, 110
  design 246
  DMCs 222
  interpreting t-tests 61–2
  logistic regression 96–7, 106
  multiple regression 94–6
  non-parametric tests 170
  publications 257
  randomisation 8–9, 10
  regulatory issues 106–8, 247–9
  sampling 26, 30, 133–4
  simple linear regression 92–4
  survival data 194, 204, 205–7
Beta error 128

Statistical Thinking for Non-Statisticians in Drug Regulation Richard Kay

ISBN 978-0-470-31971-0

Between-patient design 13, 52, 63, 77–8, 79
Bias 11–12, 127, 137, 182, 238–40
  adjusted analyses 102, 106, 108
  design 3–4, 6, 8, 11–12, 21, 226, 245–6, 254
  DMCs 220
  intention-to-treat 113, 115–17, 118, 120, 122–4, 125–6
  interim analysis 218
  meta-analysis 238–40, 241–2
  missing data 118, 120
  paired t-test 61
  publications 238–40, 260
  randomisation 4, 6, 8, 12
Binary data 18–19, 79, 88, 120, 181, 187
  chi-square tests 63–7, 76, 77, 79, 88
  CMH tests 91–2, 109, 204
  design 18–19, 21, 23
  logistic regression 96, 104–6, 109, 204, 205
  p-values and confidence intervals 45–6, 142, 145
  sample size 29, 133, 136, 138, 187
  treatment benefit measures 67–9, 76
Biocreep 186–7
Bioequivalence 14, 17, 173, 178, 182, 183
Blind review 89, 106, 117, 158, 170, 251–2, 255
Blinding 3–4, 89, 109, 223
  design 1, 3–4, 5–6, 12, 223, 226, 245
  DMCs 219–21
  regulatory issues 248–9
  sample size 138, 139
  see also Blind review; Unblinding
Block randomisation 4, 5–6
Blood pressure trials 25, 47
Bonferroni correction 148, 151, 152–4, 155
Breslow–Day test 88

Cancer/oncology trials 7, 18–19, 23–4, 45
Cardiovascular and myocardial infarction trials 18, 22, 73, 186
Carry-over effect 14
Case report forms (CRFs) 246
Categorical data 18–19, 79, 88
  chi-square tests 73–5, 76, 77, 79, 88
  CMH tests 91–2
  logistic regression 96, 97, 104–6
  treatment benefit measures 76–7
Categorisation 23–4
Censored observations 193, 203
Censoring 193–4, 208–9, 255
  independent 208–9
  survival data 193–4, 195–6, 200, 203, 206, 208–9
Central Limit Theorem 30
Central randomisation 4, 8–9
Centre effects 157, 194, 204, 242, 248
Chi-square distribution 65, 74
Chi-square tests 19, 63–7, 71, 75–7, 79–80, 88
  binary data 63–7, 76, 77, 79, 80, 88
  categorical data 73–5, 76, 77, 79, 80, 88
  extensions 77, 79, 80
  Fisher's exact test 71–2
  logistic regression 105
  one degree of freedom test for trend 75
  ordinal data 75–6, 77, 79, 80, 88
Cholesterol trials 18, 22, 26–7, 44–5
Clinical importance 143–5, 150–2, 216
Clinically relevant difference (crd) 135–6, 216, 242
  power 137
  sample size 132–3, 135–6, 139, 189
Cluster randomisation 4, 10–11
Cochrane Collaboration 231
Cochran–Mantel–Haenszel (CMH) test 88, 91–2, 109, 204
Coin tossing 49–52, 56, 72, 127
Combining centres 88–9
Common odds ratio 77, 88
Complete cases analysis 119
Completers analysis 119
Compliance 87, 107, 114–17, 181–2, 220
Composite endpoints variables 23, 150, 226
Conditional power 215–16, 223, 224
Confidence coefficient 41–2, 46
Confidence intervals 39–46, 70–1, 92, 141–4, 170, 175–80, 233
  ANCOVA 99, 102, 104, 109
  bioeqiuvalence 183
  clinical importance 143–4
  design 10, 226
  equivalence 175–6, 178–9, 180
  logistic regression 104
  meta-analysis 232–6, 241
  multiple regression 96
  non-inferiority 176–8, 179–80, 184, 187–90

p-values 58, 141–2, 143–5, 170, 178–80
paired t-test 60
sample size 42–4, 46, 187
single mean 39–44
standard errors 35, 38, 42, 43–4, 45, 70–1
survival data 196, 201, 206
transformations 164
treatment benefit measures 67, 76–7
unpaired t-test 58, 60
Confirmatory trials 16–17, 118, 248, 250, 254
Conflicts of interest 220–1
Confounding 7
Consistency 226–7
CONSORT statement 139–40, 257–8
Constancy 186–7
Constant variance 159, 160, 164
Continuous data 18–20, 29, 47, 82, 97–104
ANCOVA 97–104, 106, 109, 204
ANOVA 204
confidence intervals 45
covariates 107–8
meta-analysis 232
multiple regression 94–6
sample size 133, 135, 138, 140, 187
t-tests 61, 63, 77–9, 83
two-way ANOVA 91, 109
Control groups 2–3, 45, 78, 135, 253
bioequivalence 183
design 1–3, 12, 17–18
equivalence 17–18, 173, 180–1
non-inferiority 17–18, 173, 176, 180–1, 184–9, 191
regulatory issues 247, 249
treatment benefit measures 67–71
Count data 19–20, 29, 45
Covariates 82, 106–8, 110, 246
ANCOVA 99–104, 108
logistic regression 104–6
regulatory issues 106–8, 247–9
survival data 194, 204, 205, 207
Critical value 52, 66
Cross-over trials 12, 13, 14, 61, 78–9, 125
Cross-product term 101, 105
Cryptococcal meningitis trial 177–8, 190–1
Cumulative incidence 197
Cure rates 185–6, 188–9
Data Monitoring Committees (DMCs) 218, 219–22, 224, 248
Database lock 251, 252, 254
Degrees of freedom 42–3, 58, 59, 95, 160
chi-square tests 65, 74, 76
Dependence 93–4, 97, 104
Depression trials 19, 122, 124, 181
Design 1–24, 218, 223–7, 245–7, 248, 249, 254
regulatory issues 4, 17, 23, 247, 248–9
Dirty data 252
Dosage trials 47, 79–80, 174, 180–1, 231
design 3, 17
multiplicity 152, 154
Double-blind trials 2, 4
Dropouts 125–6, 180–1, 220, 224, 246, 253
missing data 120, 121, 125
sample size 139–40
Dynamic allocation 4, 9–10

Efficacy 214–15, 218, 219, 253
design 2, 4, 6, 11, 16–18, 21, 223–5
equivalence 17–18, 174
intention-to-treat 117, 124
multiplicity 149, 152–3
non-inferiority 17–18, 174, 176, 184–5, 186
regulatory issues 248–9, 256
stopping 214–15, 248
Endpoints 20–4, 180–1, 187, 215, 249
see also Primary endpoints; Secondary endpoints
Epilepsy trials 19, 123–4, 194
Equivalence margins 175–6, 182–3, 246
Equivalence 17–18, 145, 173–6, 182–3, 189, 253
assay sensitivity 180–2
confidence intervals 175–6, 178–9, 180
design 17–18, 246
intention-to-treat 117
p-values 145, 175, 178–80
Ethics 17–18, 137, 174, 181, 217
Event rates 135, 137, 196, 197, 231
sample size 137–8, 139
Exclusion criteria 8, 25, 116, 125, 237
Exercise duration trials 18, 22
Expected frequencies (*E*) 64–6, 67, 72, 73, 75–6
Exploratory trials 3, 16–17, 119, 139, 246, 250

F-test 157, 160
False negatives 128, 187
False positives 127, 140, 147–8, 187, 218, 242
Fisher's exact test 71–3, 75
Fixed effects model 234
Flexible designs 223–7, 248
Flu symptom trials 194, 199, 207–8
Follow-up 115, 121, 154, 216–17, 224
  survival data 193–4, 196, 197, 209, 210–11
Forest plots 234
Full analysis set 137, 158, 182, 250
  intention-to-treat 115–17, 118, 125–6
  non-inferiority 182, 188, 190
Funnel plots 238–40
Futility 153, 215–16, 218, 224–5

Gaussian distribution 29–31
Gehan–Wilcoxon tests 194, 197–9, 204, 255
General alternative hypothesis 74, 77, 83, 85
Generalisability 81, 84, 87, 88
Generalised Wilcoxon test 197
Geometric means 164, 183
Global assessment variables 22, 23

Harm 214–15
Hazard rate 200–1, 202, 203–6, 209
Hazard ratio 46, 200–4, 206–7
  constant 201, 202–3
  non-constant 201–2
Helsinki Declaration (2004) 18
Herpes Zoster trials 122–3, 193–4
Heterogeneity 92, 232, 236, 241
  multiplicity 84–6, 87, 88
Hierarchy 21, 150–2, 154, 190
Historical evidence of sensitivity 182
HIV trials 22, 199, 226–7
Homogeneity 12, 16, 84–7, 102–4, 160, 227
  meta-analysis 236, 237, 243
  survival data 204, 206
  variance 159, 160
Hypotheses 17, 47–56, 95, 97, 99–101, 160
  meta-analysis 237, 242
  see also Alternative hypotheses; Null hypotheses

Imbalance 102–3, 106, 108, 109–10, 170
Inclusion criteria 8, 25, 181, 225, 237, 251
  intention-to-treat 116–17, 124
Independence 159
Independent censoring 208–9, 255
Inferences 25–38
Intention-to-treat (ITT) 111–18, 122–4, 246, 259
  missing data 116, 118–22, 125
  population 116
Interim analysis 138, 213–18, 223, 250, 253, 254
  design 223–6, 246
  DMCs 219–20, 222
  multiplicity 148–9, 152–3, 213

Kaplan–Meier (KM) curves 195–7, 203–4
  survival data 194, 195–7, 200, 203–4, 209
  see also Median event times; Relative risk
Kidney stone trials 229–30

Last observation carried forward (LOCF) 119–20, 155
  missing data 119–20, 122
  regulatory issues 247, 249
Least squares 93, 95
  regression line 93–4
Level of effect 129–30, 132
Likert scale 24
Log hazard rate 205
Log hazard ratio 233
Log odds ratio 233
Log ratios 70
Log scale 163–4, 183, 211
  meta-analysis 233, 234
  survival data 201, 206, 210
Log transformations 163–5, 168, 183, 255
Logistic regression 19, 96–7, 104–6, 109, 204–5
Logit or logistic transform 97, 164
Logrank test 197–9
Lower confidence limit 39, 44

Main effects model 99, 104
Mann–Whitney U-test 166–8
Mantel–Haenszel (MH) chi-square test 19, 75–6, 79
Margins 246, 248, 249
  equivalence 175–6, 178–9, 182–3, 246
  non-inferiority margins 176–7, 182–5, 189–91, 278

Mean absolute deviation 29
Means 27–8, 32–8, 162, 164, 169, 187
  confidence intervals 39–46, 71
  covariates 107–8
  difference between two 44–6, 53
  equivalence 175, 178
  interpreting t-tests 62–3
  normal distribution 30–1
  notation 29
  paired t-tests 58–61
  single 39–44
  standard error 32–5, 38
  standard error of difference 35–8
  statistical significance 141–2, 145
  survival data 193–4, 207
  t-test extensions 77
  transformations 163–4
Mean-to-mean variation 35
Median event times 196, 207–8
Medians 27–8, 30, 46, 211, 258
  non-parametric tests 169
Meta-analysis 86, 184, 229–43, 246, 256, 260
  regulatory issues 237–8, 240–3, 247, 249
  see also Forest plots; Funnel plots
Minimisation 4, 9–10
Missing data 118–22, 246, 250, 251, 253, 255–6
  intention-to-treat 116, 118–22, 124
  multiplicity 148, 158
  regulatory issues 247–9
Modified intention-to-treat population 116
Multi-centre trials 81–9, 91–2, 250, 253
  design 20, 246
  randomisation 5, 7–9, 81–2, 156
  regulatory issues 148–9, 155, 247
  t-test extensions 78
Multiple regression 94–6, 99
Multiple sclerosis trials 120, 167–8, 258–60

Neyman–Pearson Lemma 128
Noise 15–16, 27, 59, 78, 99
  chi-square test 66–7
  non-parametric tests 167, 169
  p-values 53, 145
  simple linear regression 93–4
Non-constant variance 160
Non-inferiority 17–18, 117, 173–91, 246, 259
  analysis sets 182
  assay sensitivity 180–2
  biocreep 186
  confidence intervals 176–8, 179
  design 246
  meta-anlysis 232, 241
  multiplicity 154, 158, 190
  p-values 145, 177, 178–80, 187, 190
  regulatory issues 247, 248
  sample size 187–9, 191
  switching between superiority 189–91
Non-parametric tests 159–71
Normal distribution 29–32, 41, 66, 167
Normal probability plot 160–2
Normality 160–3, 183, 194, 255
  assumptions 159, 160–3, 166, 170
  transformations 160, 163–4, 166
Null distribution 65–6, 128, 167
  p-values 50, 52, 54–5
Null hypotheses 47, 59, 72, 77, 93, 99, 127–8
  chi-square tests 64, 74
  equivalence 173, 178–9
  multiplicity 83, 85
  non-inferiority 173, 187
  non-parametric tests 169
  p-values 47, 49–50, 53–6, 141, 145
Number needed to treat (NNT) 67, 69–70, 71, 76, 88

Observed frequencies (*O*) 64–6, 67, 70, 73
Odds 67–8, 77, 105
Odds ratio (OR) 67–8, 71, 76–7, 88, 92, 201
  confidence intervals 46, 70–1, 142
  logistic regression 104–5
  meta-analysis 232–3, 234, 238, 239, 241
  multiplicity 88, 156
One pivotal study 240, 241–2, 243, 247
One sample t-test 58
One-sided confidence intervals 176, 177, 179
One-sided equivalence 174
One-sided significance level 215
One-sided tailed tests 55–6
One-sided type I error 223
One-way analysis of variance (one-way ANOVA) 77–9
Ordered categorical data 19–20, 75–6

Ordinal data 19–20, 23–4, 29, 79, 88, 204
  chi-square tests 75–6, 77, 79, 79, 88
  CMH tests 91–2, 109, 204
  logistic regression 96, 97, 104–6, 109, 204
  treatment benefits 76–7
Ordinal logistic model 97
Osteoporosis trials 22
Outliers 170–1, 251, 256

p-values 47–56, 141–2, 144–5, 178–80, 255, 259
  adjusted analyses 94, 96, 97, 104, 107, 109
  ANCOVA 99–100, 102, 109
  chi-square tests 66–7, 74–5
  confidence intervals 58, 141–2, 143–5, 170, 178–80
  design 10, 13, 16
  equivalence 145, 175, 178–80
  Fisher's exact test 72
  interim analysis 215, 216
  interpreting t-tests 61–3
  meta-analysis 232, 236
  misinterpretation 144–5
  multiplicity 83, 85–6, 88, 147, 149, 152, 154–7
  non-inferiority 145, 177, 178–80, 187, 190
  non-parametric tests 167, 169, 170
  safety 218
  sample size 53, 127, 129–31, 144, 145, 178, 187
  statistical significance 141–5
  superiority 173, 180
  survival data 197, 199, 203, 204, 207
  t-tests 57–60, 77, 159, 160, 162–4, 170
  transformations 163–4
Pain trials 18, 24, 122–3, 194
Paired design 13
Paired t-tests 58–61, 62, 79, 132, 164
  assumptions 159, 168
Pairwise comparisons 78, 79, 80, 148
Parallel groups 12–13, 15, 57, 63, 258
Participant flow 258–9
Patient-to-patient variability 12, 13, 15–16
  changing parameters 135
  confidence intervals 44
  intention-to-treat 126
  multiplicity 84
  p-values 53, 145
  paired t-tests 60
  sampling 26–9, 34, 35
  standard deviation 28–9, 35
  t-test extensions 78
Pearson chi-square test 19, 63–6, 67, 71
  categorical data 73
  ordinal data 75–6
Per-protocol set 117, 137, 247, 250
  equivalence 182
  intention-to-treat 117, 118–19, 126
  multiplicity 158
  non-inferiority 182, 188
  power 137, 140
Phase I trials 220
Phase II trials 220
Phase III trials 225–6, 234–6
Placebos 3–4, 47, 129, 143, 249, 258
  design 2–4, 6, 17–18, 224
  dosage studies 79–80
  meta-analysis 232, 238, 241
  treatment benefit measures 69
Pocock method 153, 213–15, 223
Polychotomous logistic model 97
Pooling 229–31, 241, 256
Population 26–7, 39–45, 59, 116, 225
  sample 25–7, 29–33, 38
Positively skewed data 160–1, 163–4, 169, 183, 194
Post-baseline data 116, 120
Power 6, 16, 128–31, 144, 223–4, 258
  ANCOVA 102
  changing parameters 134–6
  equivalence 173
  intention-to-treat 120, 126
  interim analysis 215–16, 223
  meta-analysis 231–2
  multiplicity 85–6, 131, 156
  non-inferiority 173, 187–8
  non-parametric tests 170
  regulatory issues 136–7
  sample size 127–40, 187–8, 209–11
  survival data 209–11
  type II errors 128, 130, 137, 139
Pre-planning and pre-specification 106, 251, 253–5
  covariates 106
  design 4, 9, 12, 21, 245–7
  interim analysis 216–17
  meta-analysis 237–8, 240–2
  missing data 119, 125
  multiplicity 157, 158

non-inferiority 187, 190
per-protocol sets 126
safety 218
sample size 187
statistical process 249, 251–5, 257
Precision 11–12, 24, 232, 246
standard errors 35, 37–8, 43
Presentations 257–60
Primary endpoints 20–1, 137, 149–52, 174, 181, 187, 249
covariates 108
design 7, 17, 20–1, 23, 224–5, 226, 245–6
interim analysis 214, 223
multiplicity 82, 149–52
sample size 132, 133–4, 139–40, 187, 209
statistical process 250, 254
Probability 47–56, 147, 216
chi-square tests 65, 74, 75
Fisher's exact test 72
logistic regression 96–7
multiplicity 87
non-inferiority 189, 190
paired t-tests 59
power 128, 131, 136
survival data 196, 197–8, 203–4
see also p-values
Proof of concept 17, 22
Proportional hazards model 204, 205–7, 210–11, 255
Proportions 29, 38, 45–6
Protocol 8, 34, 117, 137, 180, 190, 250, 253–5
covariates 106
DMCs 220
intention-to-treat 111, 116, 117, 118–19, 125, 126
meta-analysis 231, 236, 237–8, 240–1
missing data 118–19, 121, 125
multiplicity 149, 151, 152, 155, 157
outliers 171
sample size 138, 139
statistical process 249–50, 253–5, 256
violations 8, 111, 220, 247
Pseudo-centres 82, 88
Publications 257–60
bias 238–40, 259

Qualitative interactions 87
Quantile–quantile plots 160–1, 164–5
Quantitative interactions 87

Random effects model 234
Randomisation 4–11, 250
baseline testing 109
bias 4, 6, 8, 12
covariates 107–8
design 1–2, 4–11, 12, 13, 226, 245
intention-to-treat 111–16, 122–4
meta-analysis 229–30, 231
publications 259, 260
sampling 26, 37
Ranking 166–7, 168, 169, 170
Re-analyses 257
Re-parameterisation 101
RECIST criteria 19, 24
Recruitment 3, 8, 115, 180–1, 217–18, 224–5
DMCs 220, 222
interim analysis 216, 217–18, 223
Regression 63, 92–7, 102, 142, 259
assumptions 159, 163
logistic 96–7
multiple 94–6
simple linear 92–4
transformations 164
Regression towards the mean 107–8
Regulatory issues 106–8, 136–8, 148–9, 240–3, 247–9, 256–7
covariates 106–8, 247–9
design 4, 17, 23, 247, 248–9
interim analysis 217
meta-analysis 237–8, 240–3, 247, 249
multiplicity 148–9, 155, 247
statistical process 249, 250, 252, 255–7
statistical significance 55
Relative risk (RR) 67, 68–9, 71, 76, 88, 142, 196
survival data 196, 203
Relative risk increase (RRI) 69
Relative risk reduction (RRR) 67, 69, 71, 76, 88, 196
Reliability 37–8
Repeatability 241
Repeated measure ANOVA 154–5
Reporting the analysis 252–3
Residuals 162–3
Respiratory problem trials 18, 105
Retrospective analyses 240–1
Risk ratio 68

Robustness 3, 182, 191, 236, 255–6
meta-analysis 236, 237, 241–2
non-parametric tests 169
survival data 209
t-tests 161, 170

Safety sets 125, 250
Safety 218, 250, 253, 256
design 2, 4, 6, 11, 16, 18, 21, 223, 225
DMCs 219–20, 222
intention-to-treat 125
interim analysis 218, 223
multiplicity 149
stopping 214
Salk polio vaccine trial 1–2
Sample distribution 26–7
Sample histogram 26–7
Sample size 25, 32–5, 131–4, 137–40, 187–9, 209–11, 223–4
adjustment 137–8
assumptions 138, 140, 162, 187–8
changing parameters 134–6
chi-square tests 66, 71, 72, 75–6
clinical importance 144
confidence intervals 42–4, 46, 187
design 5, 12, 15, 16, 223–4, 225–6
equivalence 189
intention-to-treat 125, 126
interim analysis 213, 216, 223
multiplicity 85, 138, 153
non-inferiority 187–9, 191
p-values 53, 127, 129–31, 144, 145, 178, 187
power 127–40, 187–8, 210–11
regulatory issues 248
reporting calculation 138–40
standard deviation 28, 132, 137–8, 139–40, 187
statistical process 250, 254, 258
survival data 138, 209–11
t-tests 57, 59, 79, 170
type I and type II errors 127–8, 131, 138, 140, 187
Sampling 25–38, 39–40
distributions 35–6
variation 34, 39
Scatter plots 92–4, 98, 100
Schizophrenia trial 119
Scientific method 49
Score data 19–20, 29, 45
Secondary endpoints 21, 123, 151, 231–2, 250
design 13, 17, 21, 223, 245–6
Sensitivity 117–18, 121, 162, 180–2, 247, 255–6
adjusted analyses 97, 110
design 23
intention-to-treat 117–18, 121
meta-analysis 241
missing data 121
multiplicity 148
survival data 209
see also Assay sensitivity
Sequential plan 218
Serious adverse events (SAEs) 218, 221–2, 231
Shapiro–Wilks test 162
Signal 15–16, 26–7, 59
adjusted analyses 93–4, 99
chi-square tests 64, 67
non-parametric tests 167, 169
p-values 52, 53, 55
Signal-to-noise ratio 15–16, 66–7
adjusted analyses 99, 104
chi-square tests 65, 66–7
non-parametric tests 167, 169
p-values 52, 53–4, 145
unpaired t-tests 57
Significance level 55, 127–8, 249
adjusted 148, 149–51, 153, 155
interim analysis 213–15, 223
Similarity 173, 174
Simple linear regression 92–4, 95–6, 97
Simple randomisation 5
Simpson's Paradox 230
Single-blind trials 4
Square root transformation 164
Standard deviation 28–9, 80, 134–5, 258
chi-square tests 66–7, 79
confidence intervals 41, 42
non-inferiority 187
non-parametric tests 167
power 137
sample size 28, 132, 137–8, 139–40, 187
sampling 29, 30–1, 32, 34–5, 37–8
t-tests 59, 78–9, 159, 160
Standard errors 35–8, 43–4, 66, 184
adjusted analyses 93, 95, 97, 99, 101–2, 104
confidence intervals 35, 38, 42, 43–4, 45, 70–1

difference of two means 53
meta-analysis 232–3
non-parametric tests 167, 169
p-values 53
survival data 196, 197
t-tests 59, 160
Standardisation 162
Statistical analysis plan (SAP) 157, 250–1, 252, 254–5
covariates 106
missing data 125
multiplicity 89
Statistical significance 55, 137, 141–5
equivalence 181
interim analysis 215–17
meta-analysis 232, 236, 238, 241–2
multiplicity 84, 86, 147–8, 150–2, 154, 158
non-inferiority 181, 189
non-parametric tests 167
transformations 164
Stepwise regression 96, 97, 108
Stopping rules 213–17, 218, 222, 223, 248
Stratification 7–8, 91–2, 103, 108, 109, 204
design 4, 7–10
Stroke trials 21, 133, 150, 223, 224
Study-to-study variation 229
Subgroups 155–7, 206, 231–2, 253, 254
ANCOVA 103–4
see also Treatment-by-covariate interactions
Submission 256–7
Success/failure classification 119, 120
Superiority 17–18, 117, 154, 246, 247
equivalence and non-inferiority 173, 180, 182, 188–9
switching between non-inferiority 189–91
Surrogate endpoints 21–2, 249
Survival curves 194, 195–8, 200, 202, 203–4, 210
Survival data 193–211, 255
accelerated failure time model 207–8
adjusted analyses 204–8
censoring 193–4, 208–9
Gehan–Wilcoxon tests 197–9
hazard rate 200–1
hazard ratio 200–3
Kaplan–Meier curves 195, 203–4
logrank test 197–9
median event times 197
proportional hazards model 204–6
relative risk 196
sample size 209–11

t-distribution 53–4, 57–8, 60, 160
t-tests 19, 57–61, 77–9, 159
assumptions 159, 160–3, 168, 170
extensions 77–9
interpretation 61–3
transformations 163–6
two sample 52, 57, 140, 145
see also Paired t-tests; Unpaired t-tests
Test of significance 55
Test statistic 50, 52, 53–5, 74
Time-dependent covariates 107
Time-to-event data 122–4, 255, 258, 259
Transformations 160, 163–6, 169, 170, 171
see also Log transformation
Treatment benefit measures 67–71, 76–7
Treatment-by-centre interactions 84–6, 87, 92, 236, 248
ANOVA 84–8
Treatment-by-covariate interactions 99–101, 102–4, 248
ANCOVA 100
logistic regression 105–6
multiplicity 155–6
non-parametric tests 170
Treatment-by-study interactions 236
Treatment effects/differences 67–71, 84–7, 91–2
ANCOVA 97–104, 109–10
changing parameters 134
design 7, 224–5, 246
logistic regression 104–6
meta-analysis 230, 232–3, 238, 240–2
p-values 141, 144–5
power 128–31
pre-planning 255
sample size 138
standard error 37–8
survival data 197–9, 201, 204, 206, 208
t-tests 59, 61–3
treatment benefit measures 67–71, 76–7
True difference 187
True mean 29, 33, 35, 37, 40–1, 194
True standard deviation 29
Two pivotal trial rule 241–3
Two-sample t-test 52, 57, 140, 145

Two-sided *p*-value 169
Two-sided significance levels 223
Two-sided tests 55–6
Two-sided type I error 241
Two-tailed *p*-values 215
Two-tailed tests 55–6, 130, 135, 138
Two-way analysis of variance (ANOVA) 82–5, 91, 102, 109
Two, one-sided tests approach 179
Type I errors 127–8, 131, 147, 223–5, 241, 248
  DMCs 219
  interim analysis 213, 214, 217, 223
  multiplicity 147–9, 152–3, 154
  non-inferiority 187
  p-values 49, 56, 178
  safety 218
  sample size 128, 131, 138, 140, 187
  statistical process 250, 258
Type II errors 127–8, 136–7, 258
  multiplicity 149
  p-values 49, 178
  power 128, 130, 137, 139
  sample size 127–8, 187

Unblinding 181, 249, 251–2, 254–5, 257
  design 6, 21, 223–5, 245
  DMCs 220–1
Unequal randomisation 4, 6–7, 134
Unpaired t-tests 57–8, 60, 129, 132–3
  ANCOVA 97, 99, 102
  assumptions 159, 160–3, 168
  extensions 78
  interpretation 62
  p-values 52, 141–2
  sample size 132–3, 211
  transformations 163–4
  type I and type II errors 127
Unrestricted randomisation 4, 5
Upper confidence limit 39, 44

Validation plan 251, 252
Variable selection methods 108
Variance 28, 31, 160, 255
  homogeneity 159, 160
  sample size 135, 140
  transformations 164
Visual Analogue Scale (VAS) 24

Washout period 14, 181
Weighting 84, 87, 233
Welch's approximation 157, 160
Wilcoxon rank sum test 166–7
Wilcoxon signed rank test 168–9
Withdrawals 3, 139
  intention-to-treat 111, 123–4, 125–6
  missing data 119, 120, 121
  survival data 194, 208–9
Within-patient design 13, 14, 58, 78–9, 132
Worst case/best case 119, 120, 122, 209

Index compiled by Alison Waggitt