Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

Cross-over Trials in Clinical Research

STATISTICS IN PRACTICE

Advisory Editor

Stephen Senn
University College London, UK

Founding Editor

Vic Barnett
Nottingham Trent University, UK

Statistics in Practice is an important international series of texts which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceutics; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

A complete list of titles in this series appears at the end of the volume.

Cross-over Trials in Clinical Research

Second Edition

Stephen Senn
Department of Statistical Science
and Department of Epidemiology and Public Health
University College London, UK

JOHN WILEY & SONS, LTD

Copyright © 2002 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 1UD, England

Phone (+44) 1243 779777

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

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770571.

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

Other Wiley Editorial Offices

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

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

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

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

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

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

British Library Cataloguing in Publication Data

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

ISBN 0 471 49653 7 2nd edition
(ISBN 0 471 93493 3 1st edition)

Typeset in 10/12pt. Photina by Kolam Information Services Pvt. Ltd, Pondicherry, India.
Printed and bound in Great Britain by Biddles Ltd, Guildford, Surrey
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.

Contents

Preface to the Second Edition

The reception of the first edition of this work was much better than I dared hope. I took two uncompromising positions on the subject of carry-over and these went against much conventional wisdom. Despite this, many seemed to find the book helpful, and it is as a result of this positive response that a second edition is possible.

First, I condemned all strategies that relied on pre-testing for carry-over as a means of determining the final form of the analysis of the treatment approach. Such an approach had been extremely common when dealing with the *AB*/*BA* design, but I considered that the implications of Freeman's (1989) devastating examination of the two-stage procedure made it untenable. One reviewer misread this as meaning that I also disapproved of the *AB*/*BA* design itself, but this is incorrect. It is an opinion I never expressed. In fact, I consider that, where circumstances permit, an *AB*/*BA* cross-over is an extremely attractive design to use.

Second, I expressed extreme scepticism concerning common approaches to adjusting for carry-over that relied on simplistic models for it, in particular assuming that the carry-over from an active treatment into an active treatment would be the same as into placebo. Although I worked primarily in phases II to IV whilst employed by CIBA-Geigy, I came into contact with statisticians who worked in phase I on pharmacokinetic-pharmacodynamic (PK/PD) modelling, and it puzzled me that an approach that would have been considered naïve and wrong in an earlier stage of development could be accepted as reasonable later on. In this connection it seemed to me that the criticisms of the standard carry-over model which had been made by Fleiss (1986b, 1989) were unanswerable except by abandoning it.

I consider that with the nearly ten years that have passed since the first edition, these positions look more and more reasonable. In particular, it now seems to be more or less universally the case amongst those who research into the methodology of planning and analysing cross-over trials that the two-stage procedure has been abandoned as illogical. General medical statistics textbooks have lagged behind in this respect—but, within the pharmaceutical industry at least, it seems to be well understood. For example, the ICH E9 (International Conference on Harmonisation, 1999) statistical guidelines, whilst discussing cross-over trials, do not require a two-stage analysis, despite the fact that as recently as the late 1980s, industry-based guidelines were recommending it. My PhD student, Sally Lee, did a survey of members of Statisticians in the Pharmaceutical Industry (PSI) in 2001 and found that in many firms this procedure was no longer being used. The position on adjusting for carry-over in higher-order designs has moved more slowly. Here my own view would still be characterized as extreme by some medical statisticians, although many pharmacokineticists and others involved in PK/PD modelling would regard it as reasonable and, indeed, natural. Nevertheless, the position looks less extreme than it did at the time of the first edition. Thus, in revising the book I have seen no need to revise these positions. Indeed, part of the revision consists of new material supporting them.

A feature of the first edition was that, although a great deal of space was devoted to explaining (mainly for didactic rather than practical reasons) how analysis could be carried out with a pocket calculator, the only statistical package whose use was described was SAS®. The second edition now includes, in appendices to several chapters, descriptions and code for performing analyses with GenStat® and S-Plus®. I am greatly indebted to Peter Lane and Roger Payne for help with the former and to my former colleagues at CIBA-Geigy, Andreas Krause and Skip Olsen, for help with the latter. I am also very grateful to Kelvin Kilminster and Julie Jones for help with SAS®. Any infelicities of coding that remain are my fault. Please note also that where I say that a particular analysis cannot be carried out with a particular package, I mean that my search of the help file and manual has failed to find a way to do it. No doubt all these packages have resources I did not discover.

Also included now are descriptions of analysis with Excel®, in particular with the help of the excellent add-in StatPlus® (Berk and Carey, 2000), and, for non-parametrics, with StatXact®. As regards the latter, I am particularly grateful to CYTEL for having provided me with a copy of this superb software.

In his generally positive review of the first edition (Gough, 1993) the late Kevin Gough remarked that it was a shame that recovering inter-block information in incomplete blocks designs had not been included. This has now been rectified. I have also added descriptions of the use of SAS® and GenStat® for finding sequences for designs, analysis of frequency data using Poisson regression, an explanation of how to remove the bias inherent in the two-stage procedure (together with a recommendation to avoid it altogether!), Bayesian analysis of the *AB*/*BA* design, permutation tests, more material on binary data, including random effect modelling, and survival analysis, as well as reviews of more recent literature on most topics.

As was the case with the first edition, I have marked some passages with an asterisk (*). This is either because the material is rather more difficult than the general level of the book or because, whatever its theoretical interest, its utility to the analyst is low.

In the time since the first edition I have acquired as co-authors in various papers on cross-over trials Pina D'Angelo, Farkad Ezzet, Dimitris Lambrou, Sally Lee, Jürgen Lilienthal, Francesco Patalano, Diane Potvin, Bill Richardson, Denise Till and, on several occasions, Andy Grieve. I am most grateful to all of these for their collaboration. I am also particularly grateful to Andy Grieve for many helpful discussions over the years on this topic, to John Nelder for many helpful discussions on modelling and to John Matthews, Gary Koch, Dimitris Lambrou, Kasra Afsaranijad and the late Kevin Gough for helpful comments on the first edition. My thanks are due also to Schein Pharmaceuticals for having sponsored some research into carry-over in three-period designs.

Since the first edition was published, I have left the pharmaceutical industry. This has made it much more difficult for me to obtain data to illustrate the use of cross-over trials. I am therefore particularly grateful to Sara Hughes and Michael Williams of GlaxoSmithKline for providing me with data to illustrate the use of Poisson regression, to John Guillebaud and Walli Bounds of University College London for providing me with data to illustrate generalized linear mixed models and to Bob Shumaker of Alpharma for giving permission to cite data from Shumaker and Metzler (1998) to illustrate individual bioequivalence.

At Wiley I thank Helen Ramsey, Siân Phillips, Rob Calver, Richard Leigh and Sarah Corney for their support, and in particular Sharon Clutton for having encouraged me to undertake the revision. It was with my wife Victoria's support and encouragement that I came to write the first edition and, now as then, I should like to thank her for setting me on the path to do so in the first place.

Finally, I should like to leave this word of encouragement to the reader. Cross-over trials are not without their problems and there are indications and occasions where their use is not suitable. Nevertheless, cross-over trials often reach the parts of drug development and medical research other designs cannot reach. They are extremely useful on occasion, and this is always worth bearing in mind.

Stephen Senn
Harpenden
January 2002

Preface to the First Edition

This book has been written for two types of reader. The first is the physician or biologist who carries out or wishes to carry out cross-over trials and either has to analyse these trials himself or needs to interpret the analyses which others perform for him. I have kept in my mind whilst writing the book, as a concrete example of such a person, a hospital pulmonologist carrying out research into the treatment of asthma but with a lively amateur interest in statistics. The second type of reader is the applied statistician with no particular experience as of yet in cross-over trials but who wishes to learn something about this particular field of application. The concrete example I have kept in mind here is that of a statistician working in the pharmaceutical industry whose experience to date has been largely with parallel group trials but now has to work with cross-overs.

Obviously, the needs and capabilities of these two sorts of readers are not identical. The former will occasionally find the book hard going and the latter will sometimes find it elementary and wonder why I have been at such pains to describe with words what I might have said with ease with symbols. I beg each type of reader's pardon but ask them to use the book according to their taste. The former will find that no harm comes of skipping what little algebra there is in the book. The latter may easily cut the discussion of what is grasped at once. In particular I advise the former not to bother reading Chapter 2 unless inspired by the rest of the book to do so. The other passages I have marked with an asterisk (*) are those whose practical importance is not sufficient to justify study by those who are anxious to master relevant techniques in the minimum of time.

In Chapter 10, I have covered various matters which others consider important but I do not. In my opinion this chapter may also be omitted by the reader who is only concerned with practical matters.

I have also deliberately used different styles in writing the book. Few topics in clinical trials are as controversial as the cross-over trial. The history of the two-stage analysis of the *AB*/*BA* design and the recent revolution in attitudes to it brought about largely by Freeman's (1989) paper are a good example. It would be dishonest and misleading for any author in this field always to maintain the impersonal voice of neutral infallibility. Accordingly, although I have either adopted the passive form or used 'we' where I feel confident that most or at least many statisticians would agree with what I have to say, where I am aware that some authorities on cross-over trials would disagree, I have deliberately used 'I'. In such cases, the intention is not to irritate the reader but to warn him that others hold different opinions.

I also apologize to all female readers for only using masculine personal and possessive pronouns and adjectives. I am, of course, aware that many physicians, biologists and statisticians, not to mention patients, are female. In using 'he' and 'his' I plead guilty to the sin of linguistic laziness but I have not meant to offend.

There are many practical worked examples in the book. Most come from one very large project comprising over 100 clinical trials, most of them cross-overs, with which I have been involved. This is deliberate. I do not wish to help perpetuate the notion that definitive statements regarding treatments are usually produced using single trials. In drug development in particular, and in science more generally, it is the cumulative effect of analysing studies with different designs and related objectives which constitutes the true advance in our knowledge. This is particularly true in cross-over trials where the problem of carry-over may make the interpretation of individual studies problematic but where the diseases studied are of necessity not life threatening and therefore the ethical problems associated with repeating studies are not severe.

I must also point out that, although in all cases I have presented what I believe to be reasonable analyses of the trials I have used as examples, the analysis presented is in most cases neither the only reasonable approach nor necessarily the one used for reporting the trial. My task in using these examples for writing this book has been made much easier than it would otherwise have been by the fact that the trials have already been reported on. I should like to thank my colleagues Nathalie Ezzet, Nicole Febvre, Roman Golubowski, Roger Heath, Hartwig Hildebrand, Denise Till and Elizabeth Wehrle, as well as Rolf Hambuch of the Institute for Biometrics in Freiburg in Breisgau and Anne Whitehead of Reading University, for work on these trials.

I also have to thank my colleagues Petra Auclair and Hartwig Hildebrand for collaborating on papers whose results I have partly incorporated in this book, Harald Pohlmann for contributing SAS® code and Bernhard Bablok, Farkad Ezzet, Nathalie Ezzet, Albert Kandra and Gunther Mehring for helpful comments. I am also very grateful to Vic Barnett, Peter Freeman and Robin Prescott for comments on individual chapters, to Tony Johnson for extremely helpful and encouraging comments on every aspect of the book and to Gilbert Rutherford for careful reading of the first draft. The basic idea for Chapter 6 grew out of a hint of Gary Koch's, and Peter Freeman proposed an analysis for the *AB*/*BA* design with single baselines which had not occurred to me. I am also grateful to my employer, CIBA-GEIGY Ltd, for permission to reproduce data from trials and to CIBA-GEIGY Ltd and my superior, Jakob Schenker, for general support. The views expressed in the book and the errors which remain are mine alone and those who have helped and supported me bear no responsibility for them.

Finally, I should like to thank my wife, Victoria, for having encouraged me to write the book in the first place.

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

1 Introduction

1.1 THE PURPOSE OF THIS CHAPTER

In clinical medicine, cross-over trials are experiments in which subjects, whether patients or healthy volunteers, are each given a number of treatments with the object of studying differences between these treatments. The commonest of all such designs is one in which approximately half of the patients are first given an active treatment or *verum* and on a subsequent occasion a dummy treatment or *placebo* whereas the rest of the patients are first given placebo and then on a subsequent occasion verum. This is a simple example of a type of design which we shall consider in detail in Chapter 3.

The purpose of this chapter, however, is simply to provide some gentle exposition, in very general terms, of some features of cross-over trials. In particular we shall:

- define cross-over trials;
- explain why they are performed;
- mention clinical specialties for which they are useful;
- point to some dangers and difficulties in performing them;
- as well as explain some general attitudes which will be adopted throughout the book.

Methods for analysing cross-over trials will not be dealt with in this chapter but form the subject matter of Chapters 3 to 7 inclusive. The fact that we defer the issue of analysis until later enables us to begin the discussion of cross-over trials with the help of a very famous (but relatively complex) example, of considerable historical interest, which we now consider below.

1.2 AN EXAMPLE

Example 1.1 Cushny and Peebles (1905) reported the results of a clinical trial conducted on their behalf by Richards and Light of the effect of various optical isomers on duration of sleep for a number of inmates of the

Michigan Asylum for Insane at Kalamazoo. Three treatments in tablet form were examined:

- laevorotatory hyoscine hydrobromate, 0.6 mg (which we shall refer to either as *L-hyoscine HBr* or *laevo-hyoscine*);
- racemic hyoscine hydrobromate, 0.6 mg (*R-hyoscine HBr* or *racemic hyoscine*);
- laevorotatory hyoscyamine hydrobromate, 0.6 mg (*L-hyoscyamine HBr* or simply *hyoscyamine*).

Patients were given each of these treatments on a number of evenings and also studied on a number of control nights for which no treatment had been administered. According to the authors,

> As a general rule a tablet was given on each alternate evening . . . (on) the intervening control night . . . no hypnotic was given. Hyoscyamine was thus used on three occasions and then racemic hyoscine, and then laevo-hyoscine. Then a tablet was given each evening for a week or more, the different alkaloids following each other in succession (Cushny and Peebles, 1905, p. 509.)

Table 1.1 summarizes the results in terms of hours of sleep for the patients studied.

Remark These data are of particular historical interest not only because Cushny was a pioneer of modern pharmacology and did important work on optical isomers (Parascondola, 1975) but because they were quoted (incorrectly) by Student (1908) in his famous paper, 'The probable error of a mean'. The data were in turn copied from Student by Fisher (1990a), writing in 1925, and used as illustrative material in *Statistical Methods for Research Workers*.

Table 1.1 (Example 1.1) Number of observations and mean hours of sleep by treatment and patient in a trial of hypnotics.

	Treatment							
	Control		0.6 mg L-Hyoscyamine HBr		0.6 mg L-Hyoscine HBr		0.6 mg R-Hyoscine HBr	
Patient	Number	Mean	Number	Mean	Number	Mean	Number	Mean
1	9	0.6	6	1.3	6	2.5	6	2.1
2	9	3.0	6	1.4	6	3.8	6	4.4
3	8	4.7	6	4.5	6	5.8	6	4.7
4	9	5.5	3	4.3	3	5.6	3	4.8
5	9	6.2	3	6.1	3	6.1	3	6.7
6	8	3.2	4	6.6	3	7.6	3	8.3
7	8	2.5	3	6.2	3	8.0	3	8.2
8	7	2.8	6	3.6	6	4.4	5	4.3
9	8	1.1	5	1.1	6	5.7	5	5.8
10	9	2.9	5	4.9	5	6.3	6	6.4
11	—	—	2	6.3	2	6.8	2	7.3

These data thus have the distinction of having been used in the paper which inaugurated the modern statistical era (since it was the first to deal explicitly with small sample problems) and also in what is arguably the single most influential textbook written on the subject. (See Plackett and Barnard, 1990, and Senn and Richardson, 1994, for historical accounts.)

The particular feature of these data which is of interest here, however, is that they were obtained by giving each of a number of subjects a number of treatments to discover something about the effects of individual treatments. They thus come from what we would now call a *cross-over trial* (sometimes also called a *change-over trial*) which we may now define as follows.

Definition A cross-over trial is one in which subjects are given sequences of treatments with the object of studying differences between individual treatments (or sub-sequences of treatments).

Remark It is probable that the word cross-over has come to be used for trials in which patients are given a number of treatments, because in the commonest type of trial of this sort (see Section 1.1 above) two treatments *A* and *B* (say) are compared. Patients are given either *A* or *B* in the first period and then 'crossed over' to the other treatment. In more complicated designs, however, such simple exchanges do not occur but the word cross-over is nevertheless employed. The essential feature of the cross-over is not crossing over *per se* but is as captured in our definition above.

Further remark Note that the fact that patients in a given clinical trial are assigned to sequences of treatment does not alone cause the trial to be a cross-over. For example in clinical trials in cancer it is usual for patients to be given many treatments: some simultaneously, and some sequentially. In a trial investigating a new therapy, patients might well be assigned in the first instance either to a standard first-line treatment or to the new therapy with the purpose of studying the difference of the effects of treatment on remission. Patients who failed to respond or suffered a relapse would then be given alternative therapies and so on. At the end of the trial the difference between the effects of the new and alternative therapy on time to remission (or relapse) might form the object of an analysis. We could then regard the patients as having been allocated different *treatments* with the purpose of studying differences between them. Alternatively, we might study the effect on survival as a whole of allocating the patients to the different sequences (starting with new or alternative therapy). We would then be examining the difference between *sequences*. For neither of these two ends could we regard ourselves as having conducted a cross-over trial.

Before going on to consider cross-over trials in more detail some general points may usefully be made using the Cushny and Peebles data quoted in Example 1.1.

The first point to note is the *ethical* one. It may reasonably be questioned as to whether the initial investigation of substances of this sort ought to be carried out in the mentally ill who may be not in a position to give their free consent on the basis of an understanding of the potential risks and benefits of the trial. Such agreement on the part of the patient is referred to as *informed consent* in the literature on clinical trials. (It should be noted, however, that Cushny and Peebles tried the drugs out on themselves first. A similar step is undertaken in modern pharmaceutical development where so-called Phase I trials are undertaken with healthy volunteers in order to establish tolerability of substances.) Ethical considerations provide an important constraint on the design of all clinical trials and cross-over trials are no exception. This is a point which should constantly be borne in mind when designing them. In particular the fundamental right, which should not only be granted to all patients but also made clear to them, to be free to withdraw from a trial at any time, is one which can, if exercised, cause more problems of interpretation for cross-over trials than for alternative designs.

The second point to note concerns *purpose*. The trial had a specific scientific purpose. Cushny and Peebles wished to discover if there were any differences between two optical isomers: laevorotatory (L) and dextrorotatory (D) hyoscine HBr. For practical reasons the differences had to be inferred by comparing the effect of L-hyoscine HBr to the racemic form (the mixture of L and D), R-hyoscine HBr. This pharmacological purpose of the trial was of more interest to Cushny and Peebles than any details of treatment sequences, patient allocation or analysis. I mention this point because in my opinion some of the methodological research in cross-over trials over the past few decades can be justified more easily in terms of mathematical interest *per se* rather than in terms of its utility to the practising scientist.

Nevertheless, the third point to note concerns *sequences* of treatments. These were not necessarily wisely chosen and in any case are not clearly described. If we label a control night, C, L-hyoscyamine HBr, X, R-hyoscine HBr, Y, and L-hyoscine HBr, Z, it seems that the general rule was to use a sequence:

$$X\ C\ X\ C\ X\ C\ Y\ C\ Y\ C\ Z\ C\ Z\ C\ Z\ C\ X\ Y\ Z\ X\ Y\ Z\ X\ Y\ Z$$

(Preece, 1982). This would certainly produce the number of observations recorded for patients 1 and 2, although not for any other. If there were any general tendency for patients to improve or deteriorate such a scheme would bias comparisons of treatments since, for example, Z is on average given later than X.

Despite this criticism, the fourth point, which relates to the proper *interpretation* of this trial, is to note that the conclusions of Cushny and Peebles (1905), which are that 'hyoscyamine is of no value in the dose given as a hypnotic, while the laevorotatory and racemic forms of hyoscine have about the same influence in inducing sleep' (p. 509), being based on the right data, are probably not unreasonable. This may be seen by studying Figure 1.1. The figure gives the

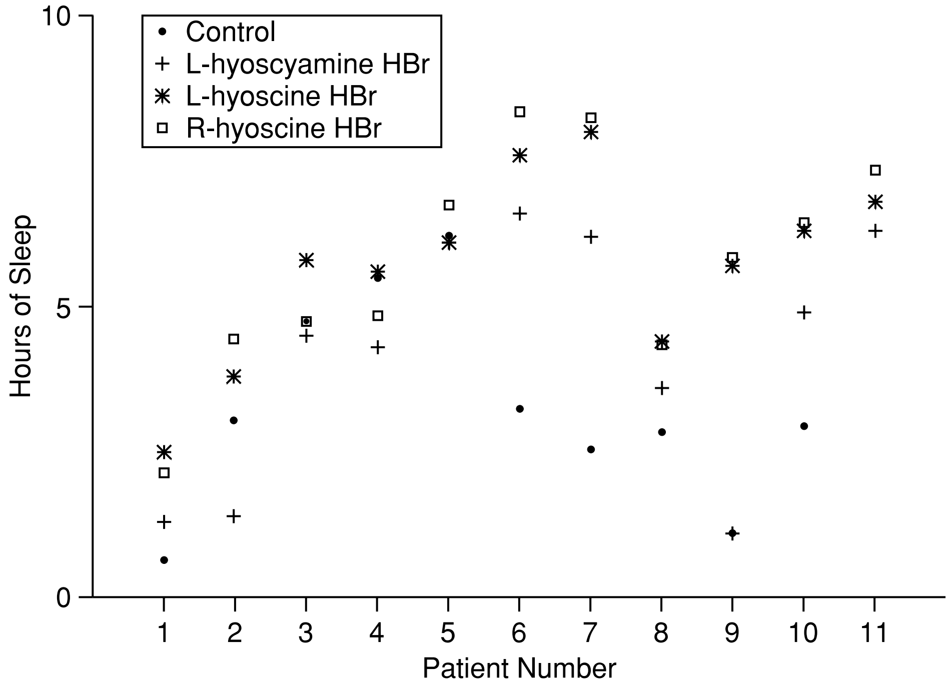


Figure 1.1 (Example 1.1) The Cushny and Peebles data. Mean hours of sleep for 11 patients for three active treatments and a control.

mean hours of sleep for the three treatments as well as the controls for patients number 1 to 10. If we compare the four results for each patient with each other, we shall see that the on the whole the values for the two forms of hyoscine are the highest of the four obtained but similar to each other. On the other hand, Student and Fisher, using incorrectly labelled data, concluded that there was a difference between optical isomers of hyoscine HBr. The moral is that the contribution to correct conclusions made by good data is greater than that made by sophisticated analysis. (In making this point I mean no disrespect to either Student or Fisher.)

The final point concerns *conduct* of the experiment. It may be noted that the patients did not each receive an equal number of treatments. Whether this was through careless planning or accident in execution one cannot say but the result is that the data bear the hallmarks of a real experiment: the data are imperfect. Missing observations continue to be one of the major problems in interpreting clinical trials, and cross-overs are no exception.

1.3 WHY ARE CROSS-OVER TRIALS PERFORMED?

We mentioned in Section 1.2 that not all trials in which patients are assigned to sequences of treatments are cross-over trials. For the trial to be a cross-over the sequences have to be of incidental interest and the object of the trial must be to study differences between the individual treatments which make up the sequences.

This was, in fact, the purpose of the trial in hyoscines reported as Example 1.1 above. Here the sequence in which the patients were given the drugs was not of interest. In fact, as we may deduce from their conduct and reporting of the trial, Cushny and Peebles (1905) probably considered the sequence in which the individual treatments were allocated as being of no consequence whatsoever. (As we pointed out above this is not always a wise point of view to take but, on the other hand, not always as disastrous as some modern commentators imply.) The purpose of the trial was to investigate the difference between the individual treatments. It is this which makes it a cross-over trial.

It is instructive to consider an alternative procedure that might have been used above. Each patient could have been assigned one treatment only. We should then have a *parallel group* trial. If we ignore the observations for patient 11 above it would thus have been necessary to study 40 (i.e., 4 × 10) patients to have obtained as many mean results per treatment as were obtained above. Even so the information would not have been as useful. In looking at Figure 1.1 it is noticeable that on the whole (there were some exceptions) patients who had high control values had high values for the three treatments. This point can be brought out by recasting the data (as Peebles and Cushny did, in fact, themselves) in the form of differences to control as has been done in Table 1.2 below. The data are also shown in this form in Figure 1.2. (No control values for patient 11 having been recorded, he is omitted from this table and figure.)

Just presenting the data in this form is revealing. Immediately it highlights the relatively poor performance of L-hyoscyamine HBr compared to the two forms of hyoscine HBr. (Even more revealing for this purpose, of course, would be calculating the difference between these treatments for each patient.) This feature of the data has been brought out by using every patient as his own control, a device which permits a particular source of variation, *between-patient*

Table 1.2 Mean hours of sleep per patient expressed as a difference from the mean obtained for the control.

	Treatment		
Patient	0.6 mg L-Hyoscyamine HBr	0.6 mg L-Hyoscine HBr	0.6 mg R-Hyoscine HBr
1	0.7	1.9	1.5
2	−1.6	0.8	1.4
3	−0.2	1.1	0.0
4	−1.2	0.1	−0.7
5	−0.1	−0.1	0.5
6	3.4	4.4	5.1
7	3.7	5.5	5.7
8	0.8	1.6	1.5
9	0.0	4.6	4.7
10	2.0	3.4	3.5

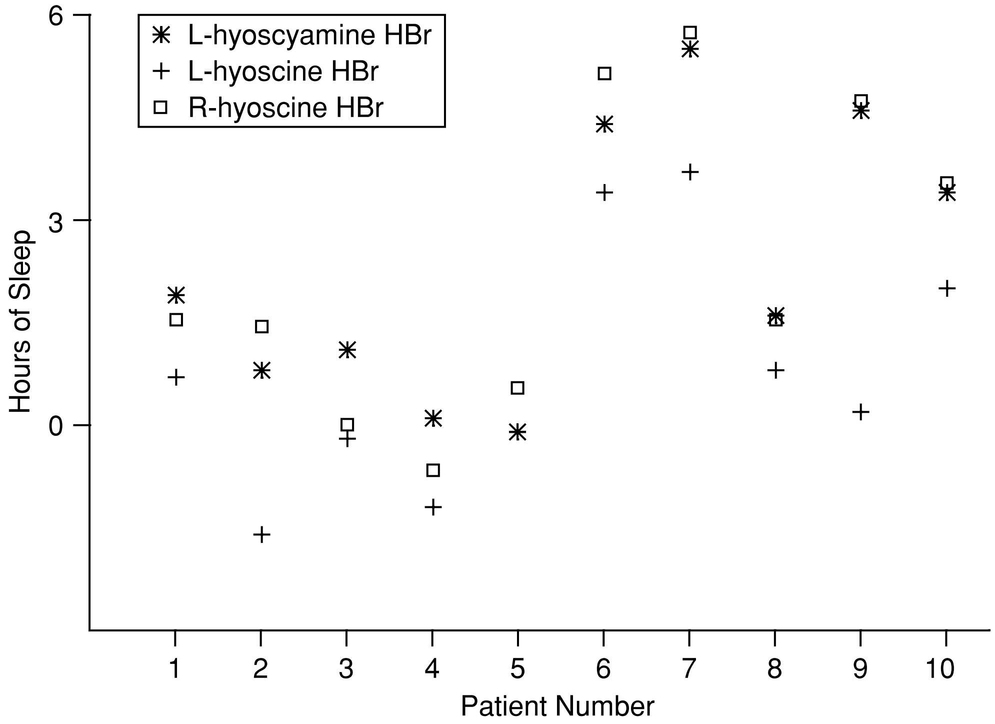


Figure 1.2 (Example 1.1) Mean hours of sleep for three active treatments expressed as a difference from control.

variation, to be eliminated. Thus, we can see that, although when the L-hyoscine HBr values and the control values from Table 1.1 are mixed together there is considerable overlap, seven of the values under treatment being lower than the highest control value, yet only one of the differences, that for patient 5, is negative.

These then are the main reasons why a cross-over trial may be preferred to a parallel group trial. First, to obtain the same number of observations fewer patients have to be recruited. Second, to obtain the same precision in estimation fewer observations have to be obtained. A cross-over trial can thus lead to a considerable saving in resources.

1.4 WHAT ARE THE DISADVANTAGES OF CROSS-OVER TRIALS?

There are disadvantages as well as advantages to the cross-over trial when compared to the parallel group trial. It is worth considering what these are.

First, there is the problem of *drop-outs*. These are patients who discontinue their programme of treatment before the trial is complete. Drop-outs cause difficulties for analysis and interpretation in parallel group trials as well but here at least the time until discontinuation for a patient may yield information which can be recovered. In cross-over trials this is extremely difficult to do and of course the patient can provide no direct information on the treatments he did not even start if, for example, he drops out during the first treatment period.

Second, there are many conditions, or *indications*, for which cross-over trials would be a quite unsuitable approach. Obviously any disease in which there is a non-negligible probability that the patient will die during the period of observation is totally unsuited for study through a cross-over trial but so, more generally, is any condition in which the patient may be expected to suffer considerable deterioration or improvement during the course of treatment. This, for example, usually makes infectious diseases (which are diseases which may on the one hand prove fatal and on the other for which the patient may be cured), an unsuitable field for cross-over trials.

Third, there is a problem which is related to that above, namely that of *period by treatment interaction*, a phenomenon which occurs if the effect of treatment is not constant over time. If it is likely that the period in which a treatment is given will modify to any important degree the effect of that treatment then not only may a given cross-over trial become difficult to interpret but the very problem itself may be difficult to detect. There is thus the danger that the investigator or 'trialist' may confidently make incorrect assertions. One such cause of period by treatment interaction is that of *carry-over*, which may be defined as follows.

Definition Carry-over is the persistence (whether physically or in terms of effect) of a treatment applied in one period in a subsequent period of treatment.

Remark If carry-over applies in a cross-over trial we shall, at some stage, observe the simultaneous effects of two or more treatments on given patients. We may, however, not be aware that this is what we are observing and this ignorance may lead us to make errors in interpretation. This topic will be covered in more detail below and will not be discussed further at this point.

Fourth, there is the problem of *inconvenience to patients*. Cross-over trials may place the patients at particular inconvenience in that they are required to submit to a number of treatments and the total time they spend under observation will be longer. It should be noted, however, that this particular feature can sometimes be turned to advantage in that it may be of interest for a patient to have the opportunity to try out a number of treatments for himself in order to gain personal experience of their effects.

Finally, there is a difficulty of *analysis*. Although there is a considerable and growing literature on cross-over trials, it is true to say that there are a number of problems still lacking totally satisfactory algorithms for their solution. For example, a type of measurement commonly encountered in clinical trials is the so-called 'ordered categorical outcome'. Such outcomes are obtained when measurements are made using a rating scale such as: poor, moderate, good. There are no easy ways of analysing such outcomes for cross-over trials with three or more treatments.

1.5 WHERE ARE CROSS-OVER TRIALS USEFUL?

Cross-over trials are most suited to investigating treatments for ongoing or chronic diseases: for such conditions where there is no question of curing the underlying problem which has caused the illness but a hope of moderating its effects through treatment. A particularly suitable indication is asthma, a disease which may last for a lifetime and remain relatively stable for years. Rheumatism is another suitable condition, as is migraine. Mild to moderate hypertension and epilepsy (chronic seizures) are also conditions in which cross-over trials are frequently employed.

Even within these areas, however, the use of a cross-over design may be more or less appropriate according to the question being investigated. Thus *single-dose trials*, in which the patient is given a single administration of the treatment under study at any particular time, even if he may be subsequently crossed over to other treatments, are usually more suitable than long-term trials in which the patient is given regular repeated administrations of the same treatment. For the latter, the danger of patients dropping out, the possibility that repeated dosing may cause difficulty with carry-over and the sheer total amount of time that may be necessary to give one patient a number of therapies may make the cross-over trial an unwise choice.

Again certain types of therapy may lend themselves more easily to cross-over trials. Thus in asthma, bronchodilators (a class of drugs which has a rapid, dramatic and reversible effect on airways) are more suitable candidates than are steroids, which have a less marked but also more persistent effect.

Cross-over trials are also very popular in single-dose *pharmacokinetic* and *pharmacodynamic* studies in healthy volunteers as well as in Phase I tolerability studies and trials for *bioequivalence*. We shall not define or discuss such studies further here. Pharmacokinetics, pharmacodynamics and bioequivalence are topics which are discussed in Chapters 7 and 10.

1.6 WHAT ATTITUDE TO CROSS-OVER TRIALS WILL BE ADOPTED IN THIS BOOK?

The basic attitude towards cross-over trials which will be adopted in this book is one of cautious optimism. There are certain problems which may occur with cross-over trials which are less likely to cause problems with parallel trials but it is quite wrong to regard cross-over trials as uniquely problematical as has been suggested in some commentaries. There are many areas in which cross-over trials are particularly useful and in such cases what the practitioner needs are simple, useful, techniques for analysing a variety of outcomes as well as practical advice regarding planning of trials. *I shall assume*, in fact, that in designing a trial the practitioner has a particular background scientific question in which he is interested (as had Cushny and Peebles) and that his interest in a given analytical technique is purely in terms of its utility to him in answering this question. A consequence of this is that a particular attitude will be adopted towards the problem of carry-over which I shall now explain.

1.7 CARRY-OVER

The problem of *carry-over* or more generally *period by treatment interaction* is taken by many commentators to be the outstanding problem of cross-over trials. It is worth considering first of all, therefore, whether this problem can cause difficulties elsewhere as well. This point can be examined with the help of an example.

Salbutamol is a standard treatment (at the time of writing undoubtedly the most popular of its class, that of beta-agonists) for patients suffering from asthma. If a new beta-agonist is to be introduced on to the market it will certainly be tested at some stage or other against salbutamol. Consider the case of a trial of a more newly developed beta-agonist, formoterol, against salbutamol.

Suppose this were done using a very simple type of cross-over in which half the patients were given salbutamol for a fixed period followed (possibly) by a period in which no treatment was given (a so-called wash-out, to be defined more precisely in Section 1.8) and then given formoterol, the other half being given formoterol first, followed by the wash-out and then the treatment with salbutamol. This sort of design is sometimes referred to as the *two-treatment, two-period cross-over*, or alternatively (and more precisely) as the *AB/BA cross-over* (where in this case *A* could be salbutamol and *B* formoterol). A particular problem which could occur with this trial is that the effect of one or other of the treatments in the first period might be such that by the end of the wash-out the patients would not be in the state they would have been in had they not been given treatment.

This might occur in a number of ways. For example there might be a physical persistence of the drug or the drug might have shown a curative effect. These are both examples of types of carry-over. In the former case there is the danger of a drug–drug interaction occurring and in the latter the second treatment may appear to benefit the patient when in fact the previous treatment is responsible. The consequence of these types of carry-over would be to bias the estimates of the effect of treatment.

Alternatively, during the time in which the cross-over trial is run the condition of the patients might suffer a *secular change* (some factor other than treatment might slowly be affecting the condition of most patients) and the benefit (or otherwise) of one drug compared to the other might be dependent on the current state of the patient. This would provide a case of period by treatment interaction. Again interpretation might be problematical.

It has been regularly overlooked, however, that for the sort of conditions in which cross-over trials are commonly used similar problems cannot be ruled out for parallel group trials. For example, it is fairly common in *'long-term' parallel-group trials* of asthma to treat patients for a year only (and frequently no longer than three months) with a view to being able to make recommendations regarding much longer periods of therapy. If the results of the trial are to be used with confidence, therefore, it must be believed that the effect of treatment beyond one year is the same as it is during the year. Obviously if the effect of treatment wears off over time—a phenomenon known as 'tachyphilaxis' (Holford and Sheiner, 1981)—then this form of period by treatment interaction may cause the results from such a trial to be quite misleading.

Again, the patients entering a parallel group trial may well have been under previous treatment. In asthma most of them will have been receiving salbutamol for many years. If there is salbutamol carry-over, then only under the very special assumption that this will be purely additive (i.e. that it will be the same regardless of which therapy follows) will this be unproblematical.

In fact, tachyphilaxis is suspected of occurring for beta-agonists, although it has usually been considered (fortunately) to be more important in terms of cardiac side-effects than in terms of efficacy. If, however, for the example above both salbutamol and formoterol showed tachyphilaxis for efficacy, then the results might be misleadingly disadvantageous for salbutamol since patients at the end of one year's apparent treatment with salbutamol would, in fact, if one takes account of their pre-trial experience, have been treated for many more. A possible consequence of this would be that the apparent advantage to formoterol could be reversed at some future date by studying a population which had previously been treated with formoterol.

Thus carry-over and period by treatment interaction are not uniquely a problem for cross-over trials as is sometimes claimed. They may affect parallel group trials as well. Nevertheless there are probably more occasions when carry-over in particular might more plausibly affect cross-over trials. It is worth considering, therefore, what may be done about it.

1.8 WHAT MAY BE DONE ABOUT CARRY-OVER?

It is easier to look first of all at what may not be done.

For many years the standard recommended analysis of the *AB*/*BA* cross-over was the so-called *two-stage procedure* (Grizzle, 1965; Hills and Armitage, 1979). This will be considered in more detail in Chapter 3 below, not because it may be recommended as a form of analysis, but because it is worth studying to bring home the dangers of *pre-testing* (i.e. carrying out preliminary tests of assumptions before choosing a substantive model). For the moment it is sufficient to describe it as follows. The two-stage procedure consists first of all of performing a statistical test on the data to examine the possibility of carry-over having occurred. If it is not judged to have occurred then a within-patient test, whereby each patient's result under treatment A is referred to his result under B, is performed. If it is judged to have occurred then, on the basis that carry-over could not possibly have affected the values in the first period, a between patient test is carried out on the first period values, comparing the results under treatment *A* for one group of patients to the results under *B* of the other group.

The overall performance of this procedure has now been studied in depth by Freeman (1989) and has been shown to be vastly inferior in almost any conceivable circumstance to the simple alternative of always doing the within-patient test. An explanation as to why this is so must wait until Chapter 3 but a simple analogy with medicine may be helpful at this point. The initial test for carry-over in the two-stage procedure is similar to a screening test for a medical condition. It has a false positive and false negative rate associated with it. Furthermore it turns out that the 'cure' one would envisage for a case known to require treatment has a high probability of being fatal when the disease is absent. Because of this the conservative approach of not screening at all turns out to be best.

There are other similar two-stage, or even *multi-stage*, testing procedures which have been proposed for more complicated cross-over designs than the *AB*/*BA* design. It is not known that these approaches definitely perform as badly as the two-stage approach for the *AB*/*BA* cross-over. It is also not known, however, that these approaches are safe and it is known that the problem which arises with the two-stage procedure is potentially a problem for all pre-testing procedures. Accordingly in this book I shall not describe any multi-stage testing procedures.

We shall at some points indicate how tests for carry-over may be performed. Regarding this, however, it is appropriate to issue the following warnings. First, that the reader should on no account consider modifying a proposed analysis for the purpose of estimating or testing a treatment effect on the basis of the result of a test for carry-over performed on the same data. Second, that the reader should be extremely cautious about interpreting the results of such tests. They are virtually impossible to interpret reasonably independently of the treatment effect and this is true even for designs and models where the carry-over and treatment estimates may be assumed to be independent. For these reasons I never test for carry-over myself.

Another approach which is popular for more complex designs than the *AB*/*BA* design has been to include parameters for carry-over and estimate treatment and carry-over effects simultaneously: that is to say, estimate treatment in the presence of carry-over and vice versa. This approach suffers, however, from the fundamental flaw that it is necessary to make restrictive assumptions about the nature of the possible carry-over in order to model the phenomenon successfully. These assumptions are not at all reasonable (Fleiss, 1986b) and involve, for example, in a dose-finding trial assuming that the carry-over of effect from the highest to the lowest dose is the same as that from the highest to the next-highest. Furthermore, it has been shown (Senn, 1992; Senn and Lambrou, 1998) that if slightly more realistic forms of carry-over apply, then using these models and their associated designs can actually be worse than doing nothing at all about carry-over.

Again, although these models are considered briefly in Chapter 10, I must issue the following warnings. First, it must be clearly understood that these models cannot guarantee protection against realistic forms of carry-over adversely affecting treatment estimates. Second, I can think of no cases where the assumptions made under these models would even approximately apply unless the carry-over were so small as to be ignorable anyway. And third, that such models often impose a penalty in efficiency of estimates. For these reasons I never use them myself.

The third approach to dealing with carry-over is that of using a *wash-out period*. This may be defined as follows.

Definition A wash-out period is a period in a trial during which the effect of a treatment given previously is believed to disappear. If no treatment is given during the wash-out period then the wash-out is *passive*. If a treatment is given during the wash-out period then the wash-out is *active*.

When a wash-out period is employed it is assumed that all measurements taken after the wash-out are no longer affected by the previous treatment. If a passive wash-out is employed the patient is assumed to have returned to some natural background state before the next treatment is started. For example in the case of single-dose trial of beta-agonists in asthma it is generally believed that a wash-out period of a few days is more than long enough to eliminate all effects of previous treatment. In a multi-dose trial we might use a different approach. Patients might be given repeated doses of one therapy during a month after which they might be switched over to an alternative therapy for another month. As a precaution against carry-over we might limit the observation period to the second two weeks of each treatment period. Obviously this only makes sense if we wish to observe the steady-state behaviour of each treatment and believe that this will be reached after two weeks at the latest under each treatment regardless of what has happened before. Note, however, that a similar assumption would have to be made in a parallel group trial with the same objective.

The main difficulty with the wash-out approach to dealing with carry-over is that *we can never be certain that it has worked*. This would be a serious objection under one or both of two conditions. First, suppose it were the case in general (except for cross-over trials) that we could say that using the results from clinical trials did not require us to make assumptions we could not 'verify'. This is not, however, the case. All analyses of clinical trials depend on dozens of assumptions we choose to ignore because we are unable or unwilling to investigate them. For example we assume that trials carried out in patients who give consent yield results which are applicable to patients in general, including those who would refuse to enter a trial. In a modern democracy there is no way that this assumption could ever be examined using clinical trials and it might even plausibly be maintained that for certain mental diseases it is unlikely to be true. Second, it would be a serious objection if there were a realistic alternative. However, there is not. Even the most enthusiastic proponents of the modelling approach to carry-over concede that one has to assume that if carry-over occurs it has taken a particular form. Such an assumption is not only not verifiable but *a priori* unlikely to be true.

A further approach to the problem of carry-over is to recognize in general that the adequacy of assumptions made in clinical trials is tested by carrying out many studies with different designs. The isolated study in which all assumptions must be 'verified' (whatever that might mean) in order that the conclusion, which is then to stand for all scientific posterity, can be known to be true (or have reasonably been declared to be true using some probabilistic rule) is an unrealistic paradigm of research. Cross-over trials are carried out where diseases are not life-threatening. There is no reason why trials should not be repeated; there is every advantage in so doing in trying radically different designs. It is in this way that scientific knowledge is increased. As different trials with different designs come up with similar results it becomes increasingly difficult to maintain that some peculiar form of carry-over could be the common explanation.

Thus, in this book I shall be making the assumption that the practitioner will deal with carry-over as follows. First he will design his trials cautiously using what he believes to the best of his knowledge to be adequate wash-out periods (whether passive or active). Second he will accept that his findings will always be conditional on an assumption (amongst many!) that carry-over has not seriously distorted his results and that there is always the possibility that different trials with different designs may not repeat them.

1.9 OTHER ATTITUDES TO BE ADOPTED

The treatment of statistics in this book will be eclectic but largely restricted to frequentist methods. (A possible Bayesian analysis is included in Chapter 3 but this is the only such example.) That is to say, a variety of frequentist approaches which I personally consider to be useful will be employed. This does not imply any hostility on my part to the *Bayesian* programme (in fact I am only too happy to acknowledge the important contribution which Bayesians like Peter Freeman and Andy Grieve have made in correcting certain frequentist errors) but merely reflects a recognition that for the time being practicalities dictate that the majority of analyses of clinical trials in general and cross-over trials in particular will be frequentist. (Although it has been predicted that by the year 2020 things will be different!)

Heuristic arguments will be employed in developing analyses. There will be no formalism and little algebra.

Global tests of significance will not be covered. In my experience it is rarely of interest for the trialist to test the null hypothesis that all treatments are equal (where there are more than two treatments). Instead the testing and estimation of specific treatment contrasts with calculation of associated significance levels and confidence intervals will be covered.

A rough agreement between *modelling* and *randomization* will be maintained. (Although this is not always a simple or obvious matter in cross-over trials, it is very easy to put a foot wrong and I can give no guarantees that I will not do so.) For example for the AB/BA cross-over if patients are allocated completely at random to the two sequences I should regard it as being permissible to ignore any period effect in the model used for analysis: not so much because of any randomization argument *per se* but because this form of allocation is consistent with a belief that the period effect is negligible. In this case, however, I should also permit the fitting of a period effect because this form of allocation is also consistent with a belief that the period effect is important if it is known that it will be dealt with by analysis. On the other hand if the investigator had blocked the trial so as to balance the sequences and ensure that an equal number of patients were assigned to each, I would regard him as being bound to fit a period effect because his behaviour shows that he considers such effects to be important.

I shall extend the ban on pre-testing for carry-over to apply to all other forms of pre-testing as well. There will be no dropping of terms from models because they are not significant. Similarly, choices will not be made between parametric and non-parametric methods on the basis of tests of normality. Quite apart from any other reason for not performing such tests it is only the within patient errors anyway which need to be normally distributed for normal theory tests to be valid for most cross-over analyses. The 'correct' examination of this point requires the investigator to go so far down the road of fitting the parametric model that it is a charade for him to pretend he has not done so. The trialist should either determine on *a priori* grounds which form of analysis he favours or perform (and of course report) both. On the whole I have not found a great use for non-parametric methods in cross-over trials but I regard them, nevertheless, as useful on occasion and therefore worth covering.

Suitable approaches to analysis will be illustrated using the computer packages SAS® (version 8.02), GenStat® (fifth edition, release 4.2) and S-Plus® (version 6.0) as well as, on occasion, StatXact® (version 4.0.1) and the spreadsheet Excel 97®. However, this book does not provide a course in any of these packages. Books that I myself have found helpful in this respect are Cody and Smith (1997) and Der and Everitt (2002) for SAS®, Harding *et al.* (2000) and McConway *et al.* (1999) for GenStat®, Krause and Olson (2000) and Venables and Ripley (1999) for S-Plus and Berk and Carey (2000) for Excel. StatXact is provided with an extremely scholarly and remarkably readable manual (Mehta and Patel, 2000). (Other more specialist texts illustrating particular aspects of analysis with these packages are referred to subsequently.) Where possible, the analyses covered will also be illustrated using calculations done on a pocket calculator.

Finally, adjustments for repeated testing will not be covered. It will be assumed that the investigator will make a sensible choice of measures, think hard about what hypotheses are worth investigating, report the results of all analyses he makes (however disappointing) and do his best to make a cautious and sensible overview of his findings as a totality.

1.10 WHERE ELSE CAN ONE FIND OUT ABOUT CROSS-OVER TRIALS?

There is a book by Jones and Kenward (1989), with a rather more theoretical treatment than this one, as well as another by Ratkowsky *et al.* (1993). A web-based tutorial by the author (Senn, 2001a) provides an introduction to the subject. There are also encyclopaedia articles by Kenward and Jones (1998) and Senn (1998a, 2000b) which provide overviews of the field.

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

2

*Some Basic Considerations Concerning Estimation in Clinical Trials**

2.1 THE PURPOSE OF THIS CHAPTER

The purpose of this chapter is to review various basic statistical concepts regarding clinical trials which will either be referred to subsequently or assumed as general background knowledge. The chapter may be omitted altogether by readers who are already familiar with the general statistical approach to clinical trials or, alternatively, who are happy to learn how to analyse cross-over trials without concerning themselves over much about justifications for methods employed. It may also be passed over and reserved for later perusal (or simply referred to as necessary) by readers who are concerned to proceed directly to the study of cross-over trials. In particular I am anxious to make the point that the reader who dislikes algebra should not allow this chapter to put him off the rest of the book. It most definitely is *not* required reading.

2.2 ASSUMED BACKGROUND KNOWLEDGE

The reader is assumed to be familiar with various basic statistical ideas, to be able to calculate and understand standard descriptive statistics such as means, medians and standard deviations, to have acquired some elementary knowledge of statistical estimation and hypothesis testing, to be at ease with the concepts of random variables, estimators and standard errors, confidence intervals, significance levels, '*P* values', etc., and to have encountered and used *t* tests, *F* tests and analysis of variance. On the computational side some familiarity with studying computer output from statistical packages, in particular in connection with linear regression, will be assumed.

The book will not make a heavy reliance on such background knowledge. As has already been explained in the Preface and the Introduction, the emphasis is on heuristic justification and teaching through worked examples. The reader's memory will be jogged regarding background theory where this is appropriate. Equally well, however, this book does *not* include a general course in medical statistics. Further help in that direction will be found by consulting Armitage and Berry (1987), Altman (1991), Campbell and Machin (1990) or Fleiss (1986a). For the background knowledge regarding clinical trials Pocock (1983) is extremely useful.

Frequent reference (often implicitly rather than explicitly) will be made to the following basic statistical ideas.

2.2.1 Linear combinations

We frequently form estimators of treatment effects by weighted addition of individual observations. Such sums are referred to as *linear combinations*. We now state three rules regarding such combinations.

(2.1) If X_i is a random variable with expected value $E[X_i] = \mu_i$ and variance $\text{var}[X_i] = \sigma_i^2$ and a and b are two constants, then the expected value $E[a + bX_i]$ of $a + bX_i$ is $a + b\mu_i$ and its variance, $\text{var}[a + bX_i]$, is $b^2\sigma_i^2$.

Practical example Suppose X_i is a random variable standing for temperature measurements in Celsius of a population of patients. Suppose its expected value is 37°C and its variance is 0.25°C^2. Then, since the corresponding Fahrenheit readings are obtained by multiplying by 9/5 and adding 32, we may substitute 32 for a and 9/5 for b, from which the expected value in Farenheit is 98.6°F and the variance is 0.81°F^2.

(2.2) If X_i and X_j are two random variables, then $E[aX_i + bX_j] = a\mu_i + b\mu_j$ and $\text{var}[aX_i + bX_j] = a^2\sigma_i^2 + b^2\sigma_j^2 + 2ab\sigma_{ij}$, where $\sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)]$ is known as the *covariance* of X_i and X_j. If X_i and X_j are independent then $\sigma_{ij} = 0$.

Practical example Suppose that in a parallel group trial X_0 is a baseline measurement in forced expiratory volume in one second (FEV_1) and X_1 is the same measurement carried out at the end of the trial. We form a new variable, the difference from baseline, by substracting X_0 from X_1. This corresponds to substituting 0 for i, 1 for j, -1 for a and 1 for b in (2.2). Hence the expected value of this difference is $\mu_1 - \mu_0$ and the variance is $\sigma_0^2 + \sigma_1^2 - 2\sigma_{01}$.

Remark Difference from baseline is a measure which is commonly encountered in clinical trials. If we assume that baseline and outcome variances are the same, so that $\sigma_0^2 = \sigma_1^2 = \sigma^2$, then it will be seen that providing the covariance σ_{01} is greater than half the variance, σ^2, the difference from baseline measure, $X_1 - X_0$, has lower variance than the raw outcomes measure, X_1.

(2.3) If X_1, $X_2, \ldots, X_n$ are n independent random variables, with expectations μ_1, $\mu_2, \ldots, \mu_n$ and variances σ_1^2, $\sigma_2^2, \ldots, \sigma_n^2$, respectively then $\sum a_i X_i$ has expectation $\sum a_i \mu_i$ and variance $\sum a_i^2 \sigma_i^2$.

Practical example Suppose we take a random sample of size n from a given population with mean μ and variance σ^2, we then have a special case of (2.3) for which $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$ for all $i, i = 1$ to n. Let the ith observation be X_i: now since the sample mean, which may be calculated as $\bar{X} = \sum (1/n) X_i$, corresponds to a weighted sum of the form considered in (2.3) with weights $a_i = 1/n$, for all i, it has expected value μ and variance $\sigma^2 \sum (1/n^2) = \sigma^2/n$.

Remark The formula for the *standard error of the mean*, commonly encountered in articles reporting medical research, is simply the square root of the variance quoted above. Thus the standard error of the mean $= \sigma/\sqrt{n}$. It should be noted, however, that the validity of this formula depends on the assumption of independence of observations which justifies the application of (2.3). This point is frequently overlooked. For example in a multi-centre clinical trial the sample of patients cannot be described as being independent random observations on any population.

2.2.2 Expected value of a corrected sum of squares

(2.4) If X_1, $X_2, \ldots, X_n$ is a random sample of size n from a population with variance σ^2, then $\sum (X_i - \bar{X})^2$ is known as the *corrected sum of squares* and has expected value $(n-1)\sigma^2$.

Remark The factor $(n-1)$ associated with the expected value of the corrected sum of squares is known as the *degrees of freedom*. It follows from the statement above that we can construct an unbiased estimate of the population variance by dividing the corrected sum of squares by the degrees of freedom. Thus the common unbiased estimate of the population variance is

$$s^2 \text{ or } \hat{\sigma}^2 = \sum (X_i - \bar{X})^2/(n-1). \tag{2.5}$$

Further remark The loss of one degree of freedom occurs because we measure the variation amongst the n observations with respect to each other rather than to some objective external standard. If we knew, for example, the population mean, μ, we could substitute this for the sample mean, $\bar{X}$, and obtain a corrected sum of squares based on n rather than $n-1$ degrees of freedom. More generally, we use one degree of freedom for every parameter estimated. So, for example, if we measure the variance of a set of observations with respect to predicted values based on a simple straight-line fit to some concomitant values, then, since we must estimate both the slope and the intercept of the line, the sum of squares corrected by fitting the line will have $n - 2$ degrees of freedom.

2.2.3 Distribution of a corrected sum of squares

(2.6) If the population from which a random sample of n observations has been drawn is *Normal* (i.e. the values are Normally distributed) with variance σ^2 and a sum of squares has been corrected by fitting m constants which are themselves linear combinations of the observations, then the expected value of the corrected sum of squares is $(n - m)\sigma^2$ and the ratio of the corrected sum of squares to the variance, σ^2, has a *chi-square distribution* with *$n - m$ degrees of freedom*.

Remark Strictly speaking this statement requires further qualification to make it absolutely correct but it will do for our purposes here. The Normality assumption is necessary for the distributional statement but not required for the expectation. The practical import of the statement is this: if in an analysis of variance or a regression we fit for a number of factors, we must reduce the degrees of freedom for our estimates of error accordingly.

2.2.4 *t* statistics

When we have obtained an estimator as a linear combination of observations from a clinical trial we also frequently proceed to estimate its variance. In order to make use of the estimates and their variances to calculate confidence intervals or test hypotheses we need to make probability statements about their distribution. We most usually do this using the *t distribution*. Student's (1908) analysis of the Cushny and Peebles (1905) data referred to in Chapter 1 and which will be covered in Section 3.3 is the earliest example of such an application.

(2.7) If Z is a random variable which is Normally distributed with mean 0 and variance 1 and Y is independently distributed as a chi-square with ψ degrees of freedom, then $t = Z/\sqrt{(Y/\psi)}$ has a t distribution with ψ degrees of freedom.

Practical example The mean, $\bar{X}$, of a random sample of size n from a Normal population with mean μ and variance σ^2 has a variance of σ^2/n. Hence

$$Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$$

has a Normal distribution with mean 0 and variance 1. By (2.6) the ratio, Y, of the corrected sum of squares, $\sum(X_i - \bar{X})^2$, to the population variance, σ^2, has a chi-square distribution with $n - 1$ degrees of freedom. Thus

$$W = \frac{\sum(X_i - \bar{X})^2/\sigma^2}{n-1}$$

is a chi-square divided by its degrees of freedom. As Student (1908) surmised and Fisher (1925) was later to prove, in samples taken from a normal population the corrected sum of squares is independent of the mean. Hence $Z/W^{1/2}$ has a *t distribution* with $n-1$ *degrees of freedom*. A little algebra is sufficient to show that

$$Z/W^{1/2} = (\bar{X} - \mu)/(\hat{\sigma}/\sqrt{n}),$$

where

$$\hat{\sigma}^2 = \sum(X_i - \bar{X})^2/(n-1)$$

is the usual sample estimate of the population variance. Thus a t statistic may be defined in terms of a population mean, a sample mean and its estimated standard error.

Remark This is the simplest example of an application of the t distribution. In this case the estimate of the population variance and the estimate of the population mean are obtained from the same sample. This is not, however, required by (2.7). There are occasions when we obtain an estimate of the population variance either partly or entirely using values other than those used to calculate the sample mean: for example when we are studying a number of populations whose variances are equal but whose means may be different. It should be noted, however, that for strict validity in order to apply (2.7) we require observations drawn from a Normal distribution. In practice this never happens though an assumption of Normality may apply approximately. Under many circumstances t-statistics are, however, *robust* and this assumption is not too critical. One of the issues which affects robustness is whether or not the variance was estimated from the same observations that were used to estimate the mean.

2.2.5 Distribution of sums of independent chi-square variables

(2.8) If W_1 and W_2 are distributed independently according to the chi-square distribution with ψ_1 and ψ_2 degrees of freedom respectively, then $W_3 = W_1 + W_2$ has a chi-square distribution with $\psi_3 = \psi_1 + \psi_2$ degrees of freedom.

Practical example When comparing the means of two treatment groups in an experiment consisting of independent observations (as, say, in a parallel group clinical trial), a two-sample t test is commonly used. The assumption that the two variances are equal is commonly made. (This assumption is perfectly natural under the null hypothesis that the two treatments are identical.) Suppose that there are n_1 observations in the first group and n_2 in the second, and let the values in the first sample be denoted X_1 and those in the second X_2. Let

$$W_1 = \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2/\sigma^2, \; W_2 = \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2/\sigma^2,$$

where σ^2 is the common variance. If X_1 and X_2 are Normally distributed, then W_1 and W_2 are distributed as chi-square random variables with $n_1 - 1$ and $n_2 - 1$ degrees of freedom respectively and, since they are independent, $W_3 = W_1 + W_2$ has a chi-square distribution with $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ degrees of freedom. Now, since $\bar{X}_1$ has variance σ^2/n_1 and $\bar{X}_2$ has variance σ^2/n_2 and since they are independent, the variance of their difference is $\sigma^2/n_1 + \sigma^2/n_2$. Therefore, if the original observations are Normally distributed, then

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

has a standard Normal distribution. Furthermore, given that the original values are Normally distributed, W_3 is independent of Z. Hence, $t = Z/\sqrt{W_3/(n_1 + n_2 - 2)}$ has a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Remark This is, of course, the basis for the common two-sample (or independent) t test covered in elementary courses in statistics. Note, however, that the general principle is that we may construct sums of corrected sums of squares from a number of sources in order to estimate a common unknown variance. If this variance is also the unknown variance relevant for describing the precision of a treatment contrast, such as that of $\bar{X}_1 - \bar{X}_2$ above, then a t statistic may be constructed. Sometimes we have a choice of variance estimate. For example, in a three-group clinical trial when comparing two of the three groups, we commonly choose between a pooled estimate of variance from the two groups being compared or from one based on all three groups. The latter gains more degrees of freedom but makes stronger assumptions since the equality of *all* the three treatments is not part of the null hypothesis being tested, which simply refers to two of them.

2.2.6 Linear regression

(2.9) Suppose we have a data set consisting of $k + 1$ observations recorded on n individuals where Y_i is an outcome value measured on individual i and

X_{1i}, $X_{2i}, \ldots, X_{ki}$ are *k predictor variables* measured for the same individual. For example, in a trial in hypertension, we might have diastolic blood pressure after four weeks as our outcome, Y, and diastolic and systolic blood pressure at baseline, together with a treatment indicator, as our X variables. Let $\beta_0, \beta_1, \ldots, \beta_k$ be a series of (unknown) constants and ϵ_i be n (unobserved) independent 'disturbance terms' such that $E(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_\mathrm{i}) = \sigma^2$ for all i. Then if

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \epsilon_i,$$

this *linear model* is referred to as a *regression equation* and the terms $\beta_0, \beta_1, \ldots, \beta_k$ are referred to as *regression coefficients*.

The above equation may be written in matrix form as

$$\mathbf{Y} = \mathbf{X} \quad \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{k1} \\ 1 & X_{12} & \ldots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \ldots & X_{kn} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

and

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

The matrix **X** is commonly referred to as the *design matrix*. Provided **X** is of full rank, $\boldsymbol{\beta}$ may be estimated using ordinary least squares as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Remark The heuristic expression of the requirement that **X** be of full rank is that there should be no redundancy in the explanatory variables. The simplest case is where one column is a linear transformation of another, as (say) when we use body temperature in Celsius and Fahrenheit as predictor variables. Clearly these are just the same thing and no statistical estimation procedure could reasonably be expected to work out how much predictive power should be attached to each, since using the one only or the other only or some mixture of the two would be equally successful. A more complicated case is where one variable is a linear combination of more than one other. For example, in a trial with three treatments *A*, *B* and *C*, if we use three so-called dummy variables $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$, recording a 1 if a given patient has the treatment in question and a zero otherwise, then for *any* given patient *i*, we have $X_{1i} + X_{2i} + X_{3i} = 1$, since every patient has one and only one treatment and hence $\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 = \mathbf{1}$, where $\mathbf{1}$ is a vector in which every element is 1. It thus follows that, for example, $\mathbf{X}_3 = \mathbf{1} - \mathbf{X}_2 - \mathbf{X}_3$ so that $\mathbf{X}_3$ is redundant. Such redundancy means that the design matrix is not of full rank, and this in turns means that the inverse is not identifiable.

(2.10) The estimator $\hat{\boldsymbol{\beta}}$ has (generalized multivariate) mean and variances given by

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \qquad V(\hat{\boldsymbol{\beta}}) = \sigma^2\ (\mathbf{X}'\mathbf{X})^{-1}.$$

If the disturbance terms are Normally distributed, then $\hat{\beta}$ has a multivariate Normal distribution.

(2.11) Fitted or 'predicted' values for the outcomes *Y* may be obtained using the equation $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

(2.12) In order to make inferences about the unknown regression coefficients, it is necessary to estimate the unknown variance, σ^2, of the disturbance terms. An unbiased estimator for σ^2 is given by

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 / (n - k - 1).$$

If the disturbance terms are Normally distributed, then $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/\sigma^2$ has a chi-square distribution with $n - k - 1$ degrees of freedom.

Remark This is an instance of what was referred to in Section 2.2.3, where it was mentioned that if *m* constants are fitted the degrees of freedom must be reduced accordingly. Here we fit one parameter for each of *k* variables and one additional parameter for the intercept, so that $m = k + 1$.

(2.13) In order to make inferences about a given regression coefficient, for example, β_i, it is necessary to calculate its standard error. This requires identification of the appropriate element of $(\mathbf{X}'\mathbf{X})^{-1}$. The diagonal elements of this matrix are the multipliers for the variances of the estimated regression coefficients and the off-diagonal elements are the multipliers for the covariances. If we refer to the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ as $\mathbf{A}$ and label the rows and columns of the matrix from 0 to *k*, referring to particular elements as

$$\mathbf{A} = \begin{pmatrix} a_{00} & a_{01} & \dots & a_{0k} \\ a_{10} & a_{11} & \dots & a_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k0} & a_{k1} & \dots & a_{kk} \end{pmatrix},$$

then

$$\mathrm{var}(\hat{\beta}_i) = a_{ii}\sigma^2$$

from which, if the ϵ_i terms are Normally distributed, it can be shown using analogous arguments to those, for example, in (2.8) above, that

$$\frac{\hat{\beta}_i - \beta}{\hat{\sigma}\sqrt{a_{ii}}} \sim t_{n-k-1},$$

where t_ψ is a random variable with Student's t distribution with ψ degrees of freedom. Thus, Student's t distribution can be used to make inferences about elements of β.

Remark Note that $(\hat{\beta}_i)$ is the product of two factors. The first, a_{ii}, is determined by the design matrix $\mathbf{X}$. The second, σ^2, is the conditional variance of Y given $\mathbf{X}$ and hence depends on the model. Suppose that we are interested in a particular coefficient β_i. Then, for a given model, the structure of $\mathbf{X}$ will determine a_{ii} and hence the efficiency with which we estimate the coefficient in question. This fact forms the basis of the theory of optimal design of experiments. $\mathbf{X}$ is chosen in such a way so as to minimize a_{ii} (or, where more than one parameter is of interest, some other suitable function of $\mathbf{A}$). There is a considerable theory of optimal design of cross-over trials based on design choices that lead to advantageous forms of $\mathbf{X}$ given the model assumed and the target parameters of interest. Some aspects of this theory and its application are critically examined in Chapter 10.

2.2.7 Relationships between various common statistical distributions

In (2.7) above we defined a relationship between t, Normal and chi-square distributions. We now mention two other relationships between common distributions which we shall have occasion to use in order to calculate probabilities.

(2.14) If Z has a standard Normal distribution, then $Y = Z^2$ has a chi-square distribution with one degree of freedom.

Remark This relationship may be used to calculate probabilities for a chi-square with one degree of freedom. Suppose we desire to calculate the probability that a chi-square variable Y is greater than some value C. Then since $Y > C$ either where $Z > \sqrt{C}$ or where $Z < -\sqrt{C}$ then the one-tailed probability for Y in terms of a chi-square with one degree of freedom corresponds to a two-tailed probability for its square root in terms of a standard Normal. Thus for example the value which cuts off the upper 5% of the chi-square, 3.84, is the square of the value which cuts off 2.5% of the standard Normal, 1.96.

(2.15) If t has a t distribution with ψ degrees of freedom, then $F = t^2$ has an F distribution with 1 and ψ degrees of freedom.

Remark This relationship may be put to a very similar use to that defined in (2.14). For example the value of a t with nine degrees of freedom which cuts off an upper tail of 2.5% is 2.262. The value of an F with 1 and 9 degrees of freedom which cuts off 5% is $5.117 = 2.262^2$.

2.3 CONTROL IN CLINICAL TRIALS

It will be taken for granted that the basic purpose of a controlled clinical trial is causal, that is to say to establish and estimate the *effects* of treatments.

Such estimates of effect are always comparative. In a *placebo-controlled trial* of an active treatment (or *verum*) we loosely refer to the effect of the active treatment but it is actually the difference between the effect of verum and the effect of placebo which we are studying. The pure effects of verum and placebo are not separately identifiable. This point is worth making because a common error in reporting clinical trials is to attempt to make separate statements about individual treatments (Senn, 1989a; Senn and Auclair, 1990). Thus one encounters statements of the sort ‘systolic blood pressure was lowered by 10 mm ($P < 0.05$) under active treatment but only by 7 mm (not statistically significant) under placebo.’ Such statements betray an ignorance of the fundamental logic of clinical trials. They make only a poor and indirect use of the control. Note that in the example the ‘result’ for the verum can be interpreted only by considering that from placebo. A 10 mm reduction in systolic blood pressure has been observed under verum but how much of this is due to treatment, how much due to secular changes and how much due to other subtle biases is impossible to ascertain. The result with placebo is used indirectly to attempt to answer this question: in other words to ‘qualify’ the results for the verum. A better approach is to abandon any pretence of being able to identify separate effects of treatment and placebo but make a direct use of placebo to estimate the difference of the two effects. Thus a preferable formulation of the results from such a trial would be to say something on the following lines: ‘the difference between mean systolic blood pressure (placebo − verum) was 3 mm,' and note whether this difference was or was not statistically significant.

It may be noted in passing that, whereas estimates of the differences of effects may be calculated for balanced experiments from the simple means under each treatment, a corresponding result does not in general hold for associated inferential statistics such as standard errors, confidence limits, *P* values and status with respect to statistical significance. Thus in the example above the fact that the difference from baseline for active treatment was statistically significant whereas that for placebo was not does not show that there was a significant difference between the two.

Obviously much more useful information than just this may be imparted in discussing the results of a trial. The purpose of the example is not, however, to claim that the labels 'statistically significant' or 'not statistically significant' are the most useful that may be attached to the results from clinical trials but to stress that the discussion should proceed in terms of estimates of differences between the effects of treatments.

A further point to note is that we have little control in clinical trials over that which may be termed the *presenting process*. That is to say, we have little influence over who *will* enter a clinical trial although, of course, through inclusion criteria we can influence who *may* enter a clinical trial. This means that modes of inference that rely on random (or even representative) sampling are not appropriate. On the other hand, we do have precise control over what may be termed the *allocation algorithm*. Such algorithms, for example simple randomization, do not determine who will enter the trial but do determine the method by which those who do enter the trial will be allocated to treatment. In the frequentist mode of inference this provides a justification for making inferences about differences between treatments.

These distinctions are less important in the Bayesian theory of statistics. However, even here, where it is the characteristics of the patients actually chosen rather than the rule by which they were selected, and the characteristics of the patients in each group rather than the rule by which they were allocated that are more important, it is still true that the ability to make precise inferences about treatment means rather than treatment differences is much more restricted. Amongst other reasons, this is because one may plausibly believe that the extent to which the difference between treatments varies from patient to patient (the treatment by patient interaction) is less than the extent to which patients vary from each other (the main effect of patients). Since in a clinical trial, if we concentrate on treatment differences, the fact that the patients may be rather different from some target population has no 'main effect' on the treatment estimate since this is eliminated by differencing, then only the interactive effect remains, which may be expected to be less important. Hence, more precise inferences are possible about treatment contrasts than about treatment means.

2.4 TWO PURPOSES OF ESTIMATION

There are at least two distinct purposes to estimation in clinical trials. First of all we are interested in making an assessment of causality for the patients in the trial. We wish to know what the effect was on the patients' health of allocating them to treatment. This turns out to be a difficult thing to do. Even more problematic, however, is the second major purpose of a clinical trial, namely, to predict what the effect of treatment will be on other or future patients not included in the trial. Only under very restrictive assumptions may these two goals be achieved using the same estimation procedure.

As was discussed in the introduction, the patients assembled in a clinical trial are not perfectly representative of the population we should like to treat. As regards the causal purpose of investigating the effect of treatment this does not particularly matter. On the other hand for the results to be directly applicable to other patients not recruited in the trial we need to assume that these patients will react in the same way despite many fundamental differences between the patients in the trial and the target population. This sort of assumption is what is referred to in the statistical literature as *additivity*. We assume that the treatment adds the same constant effect (when measured on some suitable scale) irrespective of the characteristics or baseline condition of the patient.

On the whole, in this book it will be assumed that the first of the two purposes is the reason for undertaking the clinical trial and that no formal way of achieving the second objective is possible. Of course, to the extent that additivity holds, the second objective is automatically achieved by achieving the first. We have good evidence, however, that even in the physical sciences estimated effects vary from experiment to experiment in a way which it is difficult to predict from the internal evidence of a given experiment. (See Youden, 1972, for a fascinating discussion of this problem.)

The purpose of including these remarks here is simply to remind the reader of the due caution which must be exercised in interpreting estimates from clinical trials and their associated confidence intervals.

2.5 SOME FEATURES OF ESTIMATION

There are a number of basic features associated with estimators which have to be considered: first, we must consider the circumstances under which an estimator is reasonable; second, we must calculate the value itself, the point estimate; third, the variance of the estimator will be related by some formula derived under a model to the variance or variances of the individual errors; fourth, we need to estimate these error variances. For example, as discussed in Section 2.2, in the case of a simple random sample of size n drawn from a normal population with mean μ and variance σ^2, as a point estimate we may take the sample mean $\bar{X}$, its variance will be σ^2/n and as an estimate of σ^2 we may take the sample variance:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1).$$

In general the way in which we approach each of these aspects of estimation can affect the answers we obtain. In order to illustrate this point we shall now consider an example.

Consider a two-treatment parallel group trial in two centres each recruiting n patients. Suppose that patients were allocated in equal numbers to the two treatments but otherwise totally at random. (When an allocation procedure is used which ensures that the numbers within any centre are identical for all treatment groups the allocation is described as *blocked by centre*. Here we assume the trial is *not* blocked by centre.) Suppose, furthermore, that no other relevant information is considered to be available regarding any patients. Table 2.1 illustrates the position. Here the italicized Roman letters represent observed means, the bold Greek letters represent true underlying 'expected values' and the figures in brackets are the number of patients.

If we choose a given patient at random from one of the four groups thus defined, his expected reading is equal to the expected value of the 'cell' to which he belongs. Thus for the *i*th patient in cell *IIA* we have

$$E[X_{IIAi}] = \mu_{IIA}.$$

Note that this also implies that the expected value of the mean of all the patients in cell *IIA* is also μ_{IIA}.

We assume, furthermore, that the difference between a given observation and its *expected value* may be represented by a *disturbance term* with constant variance so that, for example,

Table 2.1 Results of a two-treatment parallel group trial in two centres. *Means*, **expected** values and (numbers).

	Treatment		
Centre	*A*	*B*	Both treatments
I	$\bar{X}_{IA.}$ $\boldsymbol{\mu_{IA}}$ (m)	$\bar{X}_{IB.}$ $\boldsymbol{\mu_{IB}}$ $(n-m)$	$\bar{X}_{I.}$ (n)
II	$\bar{X}_{IIA.}$ $\boldsymbol{\mu_{IIA}}$ $(n-m)$	$\bar{X}_{IIB.}$ $\boldsymbol{\mu_{IIB}}$ (m)	$\bar{X}_{II.}$ (n)
Both centres	$\bar{X}_{.A}$ (n)	$\bar{X}_{.B}$ (n)	$\bar{X}_{...}$ $(2n)$

$$X_{IIAi} = \mu_{IIA} + \varepsilon_{IIAi}, \tag{2.16}$$

where ε_{IIAi} is the disturbance term. We assume that its variance is σ^2 and that it is independent of all other disturbance terms. These conditions together define our basic model.

We now suppose for the moment, as a special case of our model, that in each centre the treatment has exactly the same effect, so that

$$\mu_{IB} - \mu_{IA} = \mu_{IIB} - \mu_{IIA} = \tau. \tag{2.17}$$

Now suppose we produce an estimate of the treatment effect, τ, simply by using the mean in each treatment group over both centres to produce

$$\hat{\tau}_1 = \bar{X}_{.B.} - \bar{X}_{.A.}. \tag{2.18}$$

The estimator may be considered to be unbiased in the following terms. First it is conditionally unbiased given the particular randomization observed if the following apply:

(2.19) There happens to be perfect balance between centres so that $m = n/2$

or

(2.20) There is no difference whatsoever between centres so that $\mu_{IA} = \mu_{IIA}$ and $\mu_{IB} = \mu_{IIB}$.

This follows because it may easily be seen that conditional upon the particular allocation of patients the expected value of this estimator is

$$E(\hat{\tau}_1) = \{(n-m)\mu_{IB} + m\mu_{IIB} - m\mu_{IA} - (n-m)\mu_{IIA}\}/n. \tag{2.21}$$

Substitution of (2.17) and either of (2.18) or (2.19) in (2.21) yields the expected value τ.

Second, we may regard the estimator as being unconditionally unbiased (irrespective of any differences between centres) over all randomizations because over all randomizations the expected value of m is $n/2$ and as we have already seen substitution of $n/2$ for m and (2.17) in (2.21) yields the result τ.

Remark Randomization is an extremely controversial issue in statistics. Many statisticians find the property of unconditional unbiasedness over all randomizations totally irrelevant. I consider, however, that a decision to randomize does not require any justification in terms of this expected unbiasedness. The decision to randomize may be justified in terms of a willingness to regard the difference between centres as irrelevant so that condition (2.20) is assumed to apply. Hence the estimator is both conditionally and unconditionally unbiased.

Let us suppose that this circumstance does in fact apply, then applying the rules for linear combinations in Section 2.2.1 above we see that the variance of $\hat{\tau}_1$ is simply $2\sigma^2/n$. If we wish to estimate σ^2, then since we assume that patients are homogeneous across centres but not across treatments, we estimate σ^2 by using the pooled corrected sums of squares from each treatment group dividing by the total degrees of freedom, which is $2n - 2$ since we have $n - 1$ from each treatment group. Hence

$$\hat{\sigma}_1^2 = \frac{\sum\left(X_{IAi} - \bar{X}_{.A.}\right)^2 + \sum\left(X_{IIAi} - \bar{X}_{.A.}\right)^2 + \sum\left(X_{IBi} - \bar{X}_{.B.}\right)^2 + \sum\left(X_{IIBi} - \bar{X}_{.B.}\right)^2}{2n-2}. \quad (2.22)$$

Given our assumption about the differences between centres being irrelevant it then follows that $E[\hat{\sigma}_1^2] = \sigma^2$.

Remark The use of randomization blocked simply for the trial as whole (but not further by centre), of (2.18) to estimate τ and of (2.22) to estimate the error variance constitutes a *consistent position*. It is consistent with the belief that centres are irrelevant.

In practice, however, we may fear that there is a difference between centres. Suppose that the difference is δ so that

$$\mu_{IIA} - \mu_{IA} = \mu_{IIB} - \mu_{IB} = \delta. \quad (2.28)$$

If we substitute (2.23) in (2.21) we see that our estimate, $\hat{\tau}_1$, is conditionally biased unless $m = n/2$, since its expected value is

$$\tau + (2m - n)\delta/n.$$

We can, however, construct an estimator which is unbiased. For example, if we calculate

$$\hat{\tau}_2 = \{(\bar{X}_{IB.} - \bar{X}_{IA.}) + (\bar{X}_{IIB.} - \bar{X}_{IIA.})\}/2, \quad (2.24)$$

then this is unbiased whatever the value of m. If we look at the variance of (2.24) then by applying the rules for linear combinations in Section 2.2.1 above, we see that this is equal to

$$\sigma^2\{1/m + 1/(n-m)\}/2. \quad (2.25)$$

If it so happens that the allocation is balanced and $m = n/2$, then (2.24) and (2.18) are identical and substituting $n/2$ for m in (2.25) will show, as is only logical, that the variances are identical and equal to $2\sigma^2/n$. When, however, m does not equal n, then the variance given by (2.25) exceeds $2\sigma^2/n$.

Remark This shows the proper value of blocking by centre. Blocking is not at all necessary to eliminate bias. This is done by *conditioning*. The estimator, $\hat{\tau}_2$, was constructed in such a way as to be conditionally unbiased, whatever the result of the randomization. The value of blocking is that the variance of any estimator which has been constructed so as to be conditionally unbiased will be a minimum when the groups are balanced.

It may be thought that since the simpler estimator (2.18) is also conditionally unbiased if the trial has been blocked by centre, the issue of conditioning may be avoided altogether. In fact this is frequently assumed but it is false. The reason is that we do not know σ^2 and we have to estimate it. If we believe that there are important differences between centres then (2.22) will overestimate the variance, σ^2, since the sums of squares are only corrected by the means in each treatment group and therefore must also reflect the differences, δ, between centres. Since in fact we have eliminated the influence of δ from our estimate of τ it would be inconsistent to leave it in our estimate of σ^2. Balancing the experiment does nothing to eliminate the influence of δ from our estimate of σ^2: in fact its influence is greatest for balanced experiments. If we wish to eliminate δ from our estimate of σ^2, then again we have to achieve this by conditioning.

Table 2.2 shows how this may be done. Each of the four cells contains three possible predictors of the response of a given patient falling into the category. The first corresponds to a model whereby we assume that only the treatment and not the centre is relevant. The second is appropriate where we assume both treatment and centre are relevant. For the third we allow for interactions as well in that we do not assume that the treatment effect is the same in each centre. Depending on which of these three models is appropriate we correct the sum of squares within each cell prior to pooling using the appropriate predictor. The degrees of freedom to be used in the first case are $2n - 2$, in the second are $2n - 3$ and in the third are $2n - 4$. The first case will be seen to correspond to

Table 2.2 Three possible schemes for correcting sums of squares: (1) treatment effects only; (2) treatments and centres; (3) treatments, centres and interactions.

Centre	(scheme)	Treatment	
		A	*B*
I	(1)	$\bar{X}_{.A.}$	$\bar{X}_{.B.}$
	(2)	$\bar{X}_{...} + (\bar{X}_{IA.} - \bar{X}_{IIB.})/2$	$\bar{X}_{...} + (\bar{X}_{IB.} - \bar{X}_{IIA.})/2$
	(3)	$\bar{X}_{IA.}$	$\bar{X}_{IA.}$
II	(1)	$\bar{X}_{.A.}$	$\bar{X}_{.B.}$
	(2)	$\bar{X}_{...} + (\bar{X}_{IIA.} - \bar{X}_{IB.})/2$	$\bar{X}_{...} + (\bar{X}_{IIB.} - \bar{X}_{IA.})/2$
	(3)	$\bar{X}_{IIA.}$	$\bar{X}_{IIB.}$

expression (2.22). The second case gives the minimum correction required if we consider centre effects are important.

Remark To produce conditionally correct inferences we condition on what we consider is relevant. Blocking by relevant features in order to ensure balance does not of itself lead to correct inferences. Balance has to do with efficiency of our estimates. It is inconsistent to balance an experiment for factors which are not reflected in the analysis.

2.6 PRACTICAL CONSEQUENCES FOR CROSS-OVER TRIALS

When we design cross-over trials we have the opportunity to block by various features in order to ensure balance. There is no point in balancing for such features unless we consider that they are important. If we consider they are important we must use an analysis which reflects them. Balancing for features which are important leads to more efficient estimates.

Balancing for features we do not consider important enough to take account of in our analysis should be avoided unless (as is occasionally the case) avoiding such balance is very inconvenient. If we force irrelevant balance we restrict unnecessarily the set of designs we consider capable of yielding adequate results. This reduces the strength of any blinding and may, by indicating inconsistency in our beliefs, make our results less credible to others. Where we see no value in balancing we should randomize instead.

There are some features which it is either extremely inconvenient or in fact impossible to balance for. If we can measure them we may nevertheless be able to find analyses which adjust for them.

In the remainder of the book we discuss approaches to planning and analysing cross-over trials. We shall begin with the simplest and most commonly encountered design. As we shall see, depending on which of various factors we consider to be important, our attitudes to design and analysis will differ.

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

3

The AB/BA Design with Normal Data

3.1 AN EXAMPLE

The simplest of all cross-over designs is the *AB*/*BA* cross-over. We introduce this design by considering an example.

Example 3.1 The data in Table 3.1 are taken from those which were reported in Senn and Auclair (1990, p. 1290). They are measurements of peak expiratory flow (PEF), a measure of lung function (National Asthma Education Program, 1991, pp. 6–9), made on 13 children aged 7 to 14 with moderate or severe asthma in a two-treatment two-period cross-over comparing the effects of a single inhaled dose of 200 μg salbutamol, a well-established bronchodilator, and 12 μg formoterol, a more recently developed bronchodilator (Faulds *et al.*, 1991). The children were randomized to one of two sequence groups. In one group they were given a dose of formoterol in the morning and observed for 8 h in the clinic. They then travelled home where they or their parents took further measurements 10, 11 and 12 h after treatment. (PEF is a measurement which patients may record themselves using a simple device.) On a subsequent occasion after a wash-out of at least one day they presented at the clinic again and were given a single dose of salbutamol. Measurements in the clinic followed as before and were again succeeded by measurements at home. For the second sequence group, the procedure was as for the first except that they received salbutamol on the first visit to the clinic and formoterol on the second visit to the clinic.

The structure of the trial may thus be represented as follows:

Sequence	Period 1	Wash-out	Period 2
for/sal	formoterol	no treatment	salbutamol
sal/for	salbutamol	no treatment	formoterol

Table 3.1 (Example 3.1) Peak expiratory flow (PEF) in litres per minute measured 8 hours after treatment.

		PEF			
Sequence	Patient number	Period 1	Period 2	Period difference	Patient total
for/sal	1	310	270	40	580
	4	310	260	50	570
	6	370	300	70	670
	7	410	390	20	800
	10	250	210	40	460
	11	380	350	30	730
	14	330	365	−35	695
sal/for	2	370	385	−15	755
	3	310	400	−90	710
	5	380	410	−30	790
	9	290	320	−30	610
	12	260	340	−80	600
	13	90	220	−130	310

for/sal = formoterol followed by salbutamol
sal/for = salbutamol followed by formoterol

The PEF values presented in Table 3.1 are those measured in the clinic 8 h after treatment. A graphical representation of the data is given in Figure 3.1. Further details regarding this trial are to be found in Graff-Lonnevig and Browaldh (1990).

If, in general, we label the treatments in a two-period two-treatment cross-over as *A* and *B* respectively (in Example 3.1, say, with formoterol as *A* and salbutamol *B*) then the patients in the sort of cross-over trial described above have been randomized to the two sequences *AB* and *BA*. This is why this design is sometimes referred to (as it is here) as the 'AB/BA cross-over'. It is also referred to quite commonly as the 'two-treatment, two-period cross-over', but this labelling is less precise since cross-over designs have been proposed in which patients are allocated at random to one of four sequences *AA*, *AB*, *BA*, *BB* (Balaam, 1968). Such designs are then also two-period two-treatment designs.

PEF is a continuous measure which may be taken to be approximately Normally distributed. (Strictly speaking the point of interest in such a cross-over trial is really whether *differences* in PEF calculated for a given patient are Normally distributed and this condition may be roughly satisfied even if the raw values themselves looked at over all patients are not Normally distributed.) In this chapter ways of analysing *AB*/*BA* cross-over designs with Normally distributed outcomes will be considered.

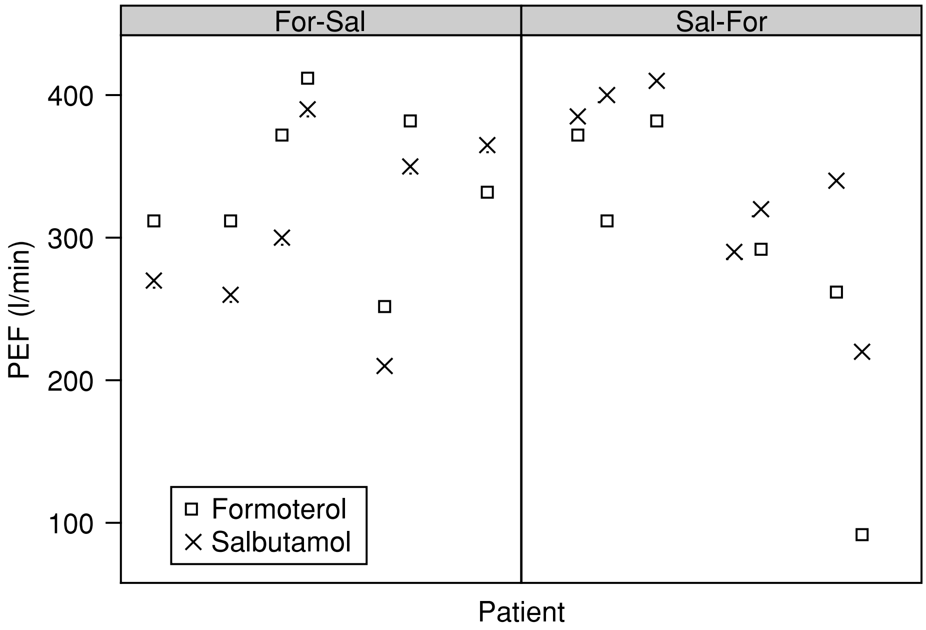


Figure 3.1 (Example 3.1) Peak expiratory flow for fourteen patients in an *AB*/*BA* cross-over.

3.2 A SIMPLE ANALYSIS IGNORING THE EFFECT OF PERIOD

Before proceeding to analyse the data in Table 3.1 two features are worth drawing attention to. First it will be seen that the patient numbers run from 1 to 14 but that there are no data from patient 8. This patient dropped out after the first visit (having been treated with salbutamol) and his values are not recorded. It will be assumed that this does not cause a problem for the analysis of this trial. It is an illustration, however, of a particular difficulty which may arise in cross-over trials and which was mentioned in the introduction. It will be assumed that the decision to leave the trial had nothing to do with the effect (or lack of effect) of the treatment given in the first period. This assumption is not always reasonable. Further discussion of this issue will be found in Chapter 9.

The second feature which is worth drawing attention to is the precision of the measurements. It seems that it is possible to measure PEF to the nearest 5 ℓ/min (judging from the values for patients 2 and 14 in period 2). Nevertheless there seems to be a strong digit preference or bias in favour of 0 as opposed to 5. This sort of thing is a very common feature of real data (Preece, 1981) and we shall encounter it in several further examples in this book. It suggests that a slightly greater attention to the business of measurement might have been made in the trial plan. Rounding to the nearest 10 ℓ/min, although probably practically perfectly realistic, is perhaps a little coarse for the purpose of some of the calculations we shall perform below (Preece, 1982). The main conclusions of the analysis will be unaffected. Nevertheless, in the analysis which follows we shall be producing statistics to many significant figures based on differences between two measurements made on a given patient, each of these measurements being subject to a maximum possible rounding error of either 2.5 or 5 ℓ/min. The reader should bear in mind that this precision of calculation is produced for the purposes of illustrating methods only.

The representation of the data in Table 3.1 is in a form commonly favoured in the statistical literature. The values are arranged in columns according to the period in which they were collected. A *period difference* is calculated for each patient by subtracting the result in period 2 from that in period 1. A total is also calculated for each patient. The use to which these figures may be put will be explained in due course.

The same data are represented again in Table 3.2, in a form more likely to be encountered in the medical literature. Here the data are grouped in columns according to the treatment received and the so-called *cross-over difference* (Koch, 1972), that is to say the difference between the two treatments (formoterol–salbutamol), has been calculated. It is noticeable that with the sole exception of patient 14 this difference is positive, or in favour of formoterol (since high values of PEF are desirable and the difference has been formed by subtracting the value under salbutamol from that under formoterol). A graphical representation of the data is given in Figure 3.2.

Table 3.2 (Example 3.1) Peak expiratory flow in litres per minute measured 8 hours after treatment.

		PEF		
Sequence	Patient number	Formoterol	Salbutamol	Cross-over difference
for/sal	1	310	270	40
	4	310	260	50
	6	370	300	70
	7	410	390	20
	10	250	210	40
	11	380	350	30
	14	330	365	−35
sal/for	2	385	370	15
	3	400	310	90
	5	410	380	30
	9	320	290	30
	12	340	260	80
	13	220	90	130

for/sal = formoterol followed by salbutamol
sal/for = salbutamol followed by formoterol

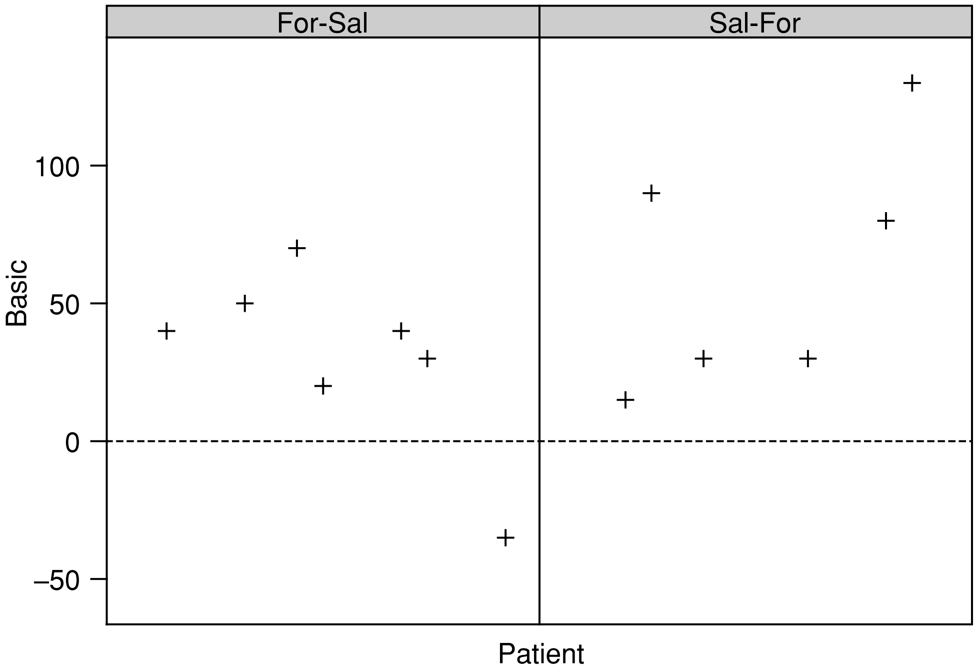


Figure 3.2 (Example 3.1) Basic estimators (formoterol–salbutamol) for peak expiratory flow for 13 patients.

A simple way of analysing data arranged in such a way is to perform a *matched-pairs t test*, sometimes also known as a *correlated t test*. (The first designation reflects the fact that the 26 PEF values may be matched in 13 pairs corresponding to the 13 patients. The second reflects the correlation that may be expected between values obtained for a given patient under one treatment and those obtained under another.) The analysis will first be illustrated numerically. We then consider a possible justification and some gentle exposition of the theory.

First we calculate the mean response, $\bar{X}$, over all 13 patients and the associated sample standard deviation, $\hat{\sigma}$, obtaining the following results:

$$\bar{X} = 45.385\,\ell/\text{min}, \quad \hat{\sigma} = 40.593\,\ell/\text{min}.$$

This calculation was performed on a scientific pocket calculator using the statistical facility. The 13 data values were entered and the mean read off using the $\bar{X}$ key. For the standard deviation the σ_{n-1} key rather than the σ_n key was used. (For some scientific calculators the symbol s is used instead of σ_{n-1}.) The formula for $\hat{\sigma}$ is thus $\sqrt{\{\sum(X-\bar{X})^2/(n-1)\}}$. In the rest of this book, except where specifically stated otherwise, wherever a sample standard deviation is calculated it will be calculated in this way using the divisor $n-1$. On the assumption that the 13 observations may be regarded as independent, then according to the results of Section 2.2 a standard error may be estimated by dividing the standard deviation by the square root of the number of patients (13 in this case). This gives the result:

$$\text{estimated SE}(\bar{X}) = 11.26\,\ell/\text{min}.$$

The ratio of the mean to its estimated standard error yields the t statistic:

$$t = 45.385/11.26 = 4.03.$$

The degrees of freedom for this statistic are the degrees of freedom available for calculating the standard deviation. One degree of freedom having been used for the calculation of the mean, this leaves 12. From tables of percentage points of Student's t distribution (Diem and Seldrup, 1982, p. 30; Lindley and Scott, 1984, p. 45) we may find that the critical value of t, two-tailed at the 5% level, is 2.179. The result is thus clearly significant. If we wish to calculate 95% confidence limits for the treatment effect, then we first form the product of the standard error and the critical value, finding $2.179 \times 11.26\,\ell/\text{min} = 24.5\,\ell/\text{min}$. This value must then be subtracted or added from the estimated treatment effect $45.4\,\ell/\text{min}$. If we use the symbol τ for the true *treatment effect*, which is defined as the difference between the expected response under formoterol and that under salbutamol, then we have the 95% confidence interval:

$$21 \leqslant \tau \leqslant 70,$$

to the nearest litre per minute.

Alternatively, from tables of the t distribution function (Lindley and Scott, 1984, p. 43) we may calculate the probability, under the null hypothesis of equality of effects for the two treatments, of observing a t statistic as extreme as that calculated (namely 4.03): the so-called P value. For this example we have $P = 0.0017$.

Implementation of various analyses in connection with the AB/BA design using a number of different computer packages will be covered at the end of the chapter. Here we limit ourselves to discussing how the matched-pairs t test can be carried out using the 'Data Analysis Tool' within Excel®. The data on PEF need to be prepared in two columns for the two treatments with one row for each patient, as is the case with columns 3 and 4 of Table 3.2. From the Tools menu 'Data Analysis' should be selected and, within that, the option 't-Test: Paired Two Sample for Means' chosen. A dialog box requests that two ranges of cell references should be input (one for each column) and that an 'alpha' value (a Type I error rate) should be given. If this is taken to be 0.05 (which is the default), the following analysis results:

	Formoterol	*Salbutamol*
Mean	341.1538	295.7692308
Variance	3558.974	6866.025641
Observations	13	13
Pearson Correlation	0.887796	
Hypothesized Mean Difference	0	
df	12	
t Stat	4.031195	
P(T<=t) one-tail	0.000833	
t Critical one-tail	1.782287	
P(T<=t) two-tail	0.001666	
t Critical two-tail	2.178813	

It will be seen that the degrees of freedom (df), t statistic (t Stat), critical values and P-values are as have been calculated above. Unfortunately, confidence limits are not provided. However, the excellent Excel® add-on package, StatPlus®, will perform the test and calculate confidence limits. This is available 'free' with the useful and very reasonably priced book by Berk and Carey (2000). The use of StatPlus® will be illustrated in connection with some of the analyses in Chapter 4. Alternatively, the regression approach within Excel® can be used to calculate confidence intervals. This will not be illustrated here, however, but in connection with an alternative analysis later in the chapter.

3.3 STUDENT'S APPROACH*

Since we have just used a matched-pairs t test and since Student (1908) in introducing the t statistic quoted the Cushny and Peebles (1905) data considered in the first example in this book (Example 1.1) it is worth noting as a matter of historical interest that Student used a similar procedure to that given above to analyse his example. We shall consider the comparison to control of L-hyoscine HBr (which Student incorrectly copied as L-hyosciamine). The relevant differences for this comparison (from one point of view analogous to the cross-over differences in an AB/BA trial) are given in the middle column of Table 1.2.

Student's table of the t distribution was originally in a different form to that which we use today and he calculated the statistic to use in connection with it as the ratio of the sample mean to the sample standard deviation (using the divisor n in calculating the latter rather than $n - 1$) obtaining, for the Cushny and Peebles data, the ratio 1.23 (Senn and Richardson, 1994). In terms of his own table this yielded a one-tailed P value of 0.0026. From this he calculated what he regarded as the odds that the true mean was positive as

$$(1 - 0.0026) \text{ to } 0.0026 = \text{nearly } 400 \text{ to1.}$$

The reader may wish to confirm that the t statistic (using the modern definition) is 3.68 with 9 degrees of freedom giving a one-sided P value which is the same as Student's.

3.4 ASSUMPTIONS IN THE MATCHED-PAIRS *t* APPROACH

In analysing the data from cross-over trials in this way we are effectively saying that we expect the cross-over differences to be distributed at random about the true treatment effect. For the t statistic to be an efficient way of examining uncertainty about the treatment effect we should also believe that the data are approximately Normally distributed, although this assumption is less important.

The question that then arises might be expressed 'are there any factors that might cause the crossover differences not to be distributed at random about the true treatment effect?' We shall consider five possible sorts of factors.

First of all there may be a *trend* affecting the experiment as whole: what is usually referred to as a *period effect*. Consider the cross-over trial of formoterol and salbutamol of Example 3.1. Suppose, for example, we estimate the treatment effect separately for each of the two sequences. We shall then find that the mean cross-over difference for the 7 patients in the formoterol/salbutamol sequence group is 30.7 ℓ/min. On the other hand the mean cross-over difference for the 6 patients in the other group is 62.5 ℓ/min. The difference between the two groups may have a number of causes. Whatever other causes are relevant, chance will undoubtedly have played its part. This will not necessarily be the whole story, however. It is possible, for example, that there is a general tendency, which would have been observed even if the patients had been given identical treatments in both periods, for values in the second treatment period to be higher than those in the first, thus attenuating the treatment difference in favour of formoterol in the formoterol/salbutamol group and accentuating it in the salbutamol/formoterol group.

There would be two consequences of such a simple trend effect. First, given any particular unbalanced allocation such as that observed in the trial above, it would tend to bias any estimate of the treatment effect which was based on a simple mean of all the cross-over differences. Of course, if the patients had been allocated at random to the two sequence groups then on average, over all allocations, the estimated treatment effect would be unbiased. This has been referred to (Senn, 1989b) as 'the consolation of the marathon experimenter'. It is a hotly debated philosophical point as to whether or not one can take any comfort from this fact. Second, if such a trend effect existed we should have ascribed to random variation a difference which was really systematic. Thus the variance of the cross-over differences would be inflated by the period effect, even (or rather especially) if we had a balanced cross-over, so that even if the treatment effect was in no danger of partial confounding with the period effect, we should nevertheless incorrectly assess the extent to which our estimate of the treatment effect might be subject to random variation. This problem is analogous to that discussed in Section 2.5 in which we showed that, for a trial in two centres, if there was a difference between centres this would inflate our estimate of variance unless we took specific steps in the *analysis* to deal with it. A way in which these two consequences of a period effect may be dealt with through analysis will be considered below.

For the moment, however, we return to the list of factors, apart from treatment and chance, which may affect the values of the cross-over difference. A more complex phenomenon than a simple period effect would be a *period by treatment interaction*. It might be the case, for example, that the effect of treatment varied according to the period in which it was given. Suppose we ran a cross-over trial in asthma in which all of the first visits were conducted in June and all of the second visits in October. The patients might be affected by hay fever at the first visit but not at the second. This in itself might lead to a period effect, but if one of the treatments were effective for asthma in general, except when complicated by or provoked by hay fever, this would also lead to a treatment by period interaction. That is to say the effect of a treatment would be modified according to the period in which it was given. Note that this would also be a problem for interpreting the results of a parallel group trial.

A third factor that could affect the results is *carry-over*. The trial itself seems to show that formoterol has a longer duration of action than salbutamol (a result which has been duplicated in many other trials with different designs, see Richardson and Bablok (1992)). It is possible to believe that patients who took salbutamol after formoterol might have their results in the second period affected to a greater extent than those who took formoterol after salbutamol. This would have the effect of making the cross-over differences for the for/sal group smaller than those for the sal/for group, which is in fact on average the case here. Obviously the reasonableness of this assumption depends on the washout employed. In fact in this trial, as is common in single-dose trials of bronchodilators, no fixed interval between the two treatment days was maintained. For most patients an interval of 48 hours was observed, for others 5 days and for one patient $3\frac{1}{2}$ months; two of the patients, patients 13 and 14, had a one-day interval only; one of these, patient 14, had salbutamol administered after formoterol. He is thus the patient who, on *a priori* grounds, is most likely to have had his results affected by carry-over and it just so happens that he is precisely that patient who has the lowest cross-over difference (in fact the only one that is negative). I am inclined, on the basis of my knowledge of other trials, to think that this is nothing but a coincidence but in general, as will be demonstrated when considering carry-over in more detail below, it is not possible to examine the question as to whether carry-over has or has not occurred by examining the cross-over differences alone from a given trial. If it has occurred, however, it will bias the estimate of the treatment effect.

A fourth factor that is worth considering is that of *patient by treatment interaction*. It may be the case that there is no such thing as a general treatment effect of a drug but that it varies from patient to patient. It could be, for example, that for patient 14, irrespective of the order in which we administered salbutamol and formoterol, we should almost always find a negative cross-over difference, whilst for other patients we should always find a positive one. This sort of discovery would certainly affect our interpretation of results. It would suggest that whatever opinion we had regarding the superiority of formoterol over salbutamol some patients would be better off under salbutamol. This would be an example of a qualitative interaction. If all patients are better off under a given treatment but to differing degrees, we have what is called a quantitative interaction. On the other hand it might be the case that the negative cross-over difference observed for patient 14 has no permanent association with him, is a matter of pure luck reflecting the general instability of lung-function from day to day in asthmatics, and might just as easily be observed in any other of the 13 patients were we to run the trial again.

In general, patient by treatment interactions cannot be investigated in a two-period treatment cross-over: designs are needed in which patients are given the same treatments a number of times. (It might be possible to investigate the phenomenon in the Cushny and Peebles trial of Example 1.1, were the original data available to us, since the patients were studied on a number of nights for each treatment. See Preece, 1982 for a discussion.) What is possible (at least in principle), however, is to study the differences in the way in which patients of a given type respond to treatment. For example, in the formoterol/salbutamol trial patients 2, 3, 6 and 11 are female, the rest being male. If we arrange the cross-over differences according to sex we get the position in Table 3.3.

We can then test the null hypothesis that the expected cross-over difference is identical for the two sexes using a *two-sample* (or *independent*) *t test*. The required calculations are also given in Table 3.3.

In Table 3.3 it will be seen that the t statistic is 0.33 (on 11 degrees of freedom) and is clearly not significant, so that there is no good reason provided by these data to believe that these drugs have different effects on PEF according to the sex of the child.

A fifth factor that might be operating could be *patient by period interaction*. This would arise if patients were subject to trend effects which were not the same for everyone. Again we use the example of hay fever to illustrate this.

Table 3.3 (Example 3.1) Cross-over differences for PEF in litres/minute (formoterol–salbutamol) arranged by sex of patient.

Sex			
Female		Male	
Patient	Cross-over difference	Patient	Cross-over difference
2	15	1	40
3	90	4	50
6	70	5	30
11	30	7	20
		9	30
		10	40
		12	80
		13	130
		14	−35
Mean	51.25		42.78
Corrected sums of squares	3618.75		15 955.56

Degrees of freedom $= (4 - 1) + (9 - 1) = 11$
Pooled estimate of variance $= (3618.75 + 15\,955.56)/11 = 1779.48$
Estimated standard error $= \surd\{1779.48(1/4 + 1/9)\} = 25.35$
t statistic $= (51.25 - 42.78)/25.35 = 0.33$

Suppose patients were recruited over a considerable period of time. It might be the case that for those patients recruited earlier during the year their first period of treatment was before the start of the hay-fever season but that the second period was during the hay-fever season. Such patients might be subject to a secular deterioration in PEF. On the other hand patients recruited somewhat later during the year might have their first period of observation during the hay-fever season and the second once the season was over. Such patients would be subject to secular improvement. There are many other causes which could produce patient by period interaction.

Of the five factors which may affect the *AB*/*BA* cross-over discussed above—period effects, period by treatment interaction, carry-over, patient by treatment interaction and patient by period interaction—the first can be simply dealt with and the last two not so much cause a problem as regards validity of analysis but add to the general variability of the results and may cause some difficulties with interpretation. As we discussed in Chapter 1, the remaining two factors, carry-over and more generally period by treatment interaction, have long been considered to be the major problems of cross-over designs. A standard proposal for dealing with them will be dealt with in some detail later in this chapter. First, however, a simple way of adjusting for period effects will be outlined.

3.5 ADJUSTING FOR A PERIOD EFFECT: TWO-SAMPLE *t* APPROACH

A very simple procedure allows us to adjust for period effects. Consider the period differences in Table 3.1. If a constant trend is present then this must affect each of the period differences identically. Thus any differences between them cannot be due to the period effect. Differences between period differences in the same sequence group can be regarded as being random. On the other hand differences between any two period differences in different sequences would also reflect treatment differences. (They would also reflect carry-over if it were present, but for the purpose of the current discussion this will be assumed not to be a problem.) Thus by comparing the means of the period differences for the two sequences we may examine the treatment effect (Jones and Kenward, 1989). This may be done by using a two sample t test for the period differences. The calculations are set out in Table 3.4.

It will be seen that whereas previously, not adjusting for the period effect, the t value was 4.0 on 12 degrees of freedom, it is now 4.3 on 11 degrees of freedom (one degree of freedom having been used for the period).

These results are very similar. The mean differences and standard errors produced in the calculation, however, are not directly comparable. This is because the mean period difference for the formoterol/salbutamol sequence is an estimate of the difference between formoterol and salbutamol and the difference between period 1 and period 2 whereas the mean period difference for the salbutamol/formoterol sequence is an estimate of the difference between salbutamol and formoterol and the difference between period 1 and 2. In eliminating the period difference by subtracting the second estimate from the first we end up with an estimate of *twice* the difference between formoterol and salbutamol. This can easily be adjusted by dividing the difference in means and its associated standard error by 2 (obviously this leaves the t statistic, which is the ratio of the two, unaffected).

Table 3.4 (Example 3.1) Calculations for testing the treatment effect in the presence of a period effect (two-sample t approach).

	Sequence	
	formoterol/salbutamol	salbutamol/formoterol
Mean period difference PEF (ℓ/min)	30.71	−62.5
Corrected sums squares PEF (ℓ/min)^2	6521.43	9987.5
Pooled estimate of variance (ℓ/min)^2 $= (6521.43 + 9987.5)/(6 + 7 - 2) = 1500.81$		
Difference in means (ℓ/min) $= 30.71 - (-62.5) = 93.21$		
Standard error of difference in means (ℓ/min) $= \surd\{1500.81(1/6 + 1/7)\} = 21.55$		
$t = 93.21/21.55 = 4.33$		
P value $= 0.0012$		

Performing these calculations we get an estimated treatment effect of 46.6 ℓ/min with a standard error of 10.78 ℓ/min. These results are very similar to those we obtained in Section 3.2, where the values were 45.4 ℓ/min and 11.3 ℓ/min respectively. (It may be noted in passing that for a balanced cross-over in which the number of patients in each sequence is the same, adjusting for any period effect has no influence at all on the treatment estimate. It does, however, affect the estimate of its standard error.) From tables of the *t* distribution (Diem and Seldrup, 1982; Lindley and Scott, 1984) we see that the value of the *t* distribution corresponding to a tail area of 2.5% for 11 degrees of freedom is 2.201. The product of this with the standard error is 23.7 ℓ/min. Adding and subtracting this to the treatment estimate yields a 95% confidence interval:

$$23 \leqslant \tau \leqslant 70$$

to the nearest litre per minute of PEF. This is extremely similar to the result obtained in Section 3.3.

The analysis of this section may also be performed using the Excel® Data Analysis tool. The most convenient approach is that provided by the 'Regression' option since this will also give confidence limits. The data should be prepared in two columns. The first should consist of the semi-period differences. That is to say, the values given in column 5 of Table 3.1 should be divided by two to give 20, 25,. . . , − 65 ℓ/min. The second column should be a dummy variable coded 1 for any patient in the first sequence group and 0 for any patient in the second treatment group. The cell references for the first column should be input as the 'Y-range' and those for the second as the 'X-range'. If this is done, the output will include the following:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	−31.25	7.91	−3.95	0.0023	−49	−14
X Variable 1	46.607	10.78	4.32	0.0012	23	70

The row associated with 'X Variable 1' will be found to contain the various statistics calculated by hand in this section.

3.6 ADJUSTING FOR A PERIOD EFFECT: THE HILLS–ARMITAGE APPROACH

An alternative approach for adjusting for the period effect (Hills and Armitage, 1979) leads to identical results in the case of an *AB*/*BA* cross-over but is more readily generalizable to more complex designs (Senn and Hildebrand, 1991). In order to discuss this method we first of all introduce a concept we shall make use of in subsequent chapters, namely that of the basic estimator, which we define as follows.

Definition A basic estimator of a given treatment contrast is the given contrast calculated for an individual patient.

For a simple AB/BA cross-over the only possible treatment contrast of interest is $A - B$ (or its converse $B - A$) and so the basic estimators are simply the cross-over differences. For Example 3.1 these are given in Table 3.2. For each patient we regard the cross-over difference as a *basic estimator* of the treatment effect. That is to say, *in the absence of any other knowledge either about period effects or treatment effects*, and were we only able to study one patient, we should regard his cross-over difference as the best estimate of the treatment effect available to us.

Now of course if, unbeknown to us, there is a trend over time, then this estimate will be confounded with a period difference. Note, however, that if we average two basic estimators, one from each of the two sequences, the resulting mean will not be confounded with a period effect.

This in general shows how we may form an estimator which is not biased by a period effect. What we need to do is first to average all the cross-over differences (i.e. the basic estimators) in one sequence and then average the resulting means. (This approach is available for a very wide range of cross-over designs as will be discussed in Chapter 5 below.) We now need to establish the variance of the resulting estimator. This may be done simply using the rule for linear combinations discussed in Section 2.2. Assuming that all the basic estimators have the same unknown variance, σ^2, then if there are n_1 patients in the first sequence and n_2 in the second, the means within each group of the basic estimators have variances equal to σ^2/n_1 and σ^2/n_2 respectively. The mean of these two means then has a variance of $\sigma^2(1/n_1 + 1/n_2)/4$. (Note that for a balanced cross-over with n patients in total $n_1 = n_2 = n/2$ and the formula for the variance reduces to σ^2/n, which is identical to the variance of the simple mean of all the cross-over differences. This is as it should be since the estimated treatment effect for a balanced cross-over will be the same whether or not we adjust for the period effect.)

The same does not, however, apply to the estimate of σ^2 itself. This we may obtain by calculating the corrected sums of squares for the basic estimators within each sequence group, adding the two results together and dividing by the residual degrees of freedom, $n_1 + n_2 - 2$. Using the results of Section 2.2 we substitute the estimated value of σ^2 in the formula for the variance of the mean and take the square root of the result to give us an estimate of the standard error of the estimated treatment effect. The ratio of the estimate and its estimated standard error is then a t statistic.

Using this procedure for the example of this chapter we start with the cross-over differences in Table 3.2 noting that $n_1 = 7$ and $n_2 = 6$. For the two sequence groups we obtain means of 30.71 and 62.50 ℓ/min and corrected sums of squares of 6521.43 and 9987.50 $(\ell/\text{min})^2$. The mean of the means is thus 46.61 ℓ/min and the pooled estimate of $\sigma^2 = 1500.81(\ell/\text{min})^2$ on $(6-1)+(7-1)=11$ degrees of freedom. The standard error we calculate as 10.777 ℓ/min and hence the t statistic is 4.33.

It will not have escaped the reader's attention that not only is the final result identical to that obtained using the two sample t approach but the calculation is virtually identical. This is because we have formed a treatment estimate by averaging two means of basic estimators (cross-over differences) whereas in the previous approach we obtained half the difference between two means of the period estimates. Since for the first group the period difference is the same as the basic estimator and since for the second the period difference is simply the basic estimator with the sign changed, the result is the same.

In general, however, for more complicated designs, this simple relationship between basic estimators and period differences does not exist. Consider, for example, a three-period cross-over design for three treatments, A, B and C, in which patients are allocated at random to one of the three sequences:

Seq I $\quad A\ B\ C$

Seq II $\quad C\ A\ B$

Seq III $\quad B\ C\ A$

If we are interested in the difference between the effects of A and B a basic estimator can be simply obtained for each patient but it corresponds to a different period difference in each sequence.

The particular analysis outlined in this section is the one which we recommend as the standard approach for the AB/BA cross-over. We shall refer to it on many subsequent occasions in the book and shall therefore dignify it with a particular name and, following Freeman (1989), refer to it as the *CROS analysis*.

We conclude this section by explaining how the basic estimator approach may be implemented using Excel®. Again the 'Regression' option is chosen. This time, however, the 'Y-range' column consists of the basic estimators (or cross-over differences), as given in the last column of Table 3.2. The 'X-range' column is again a column of dummy variables but the coding is now different, being 0.5 for patients in the first sequence and −0.5 for those in the second. When this analysis is performed, the output includes the following:

	Coefficients	*Standard Error*	*t Stat*	*P*-value	*Lower 95%*	*Upper 95%*
Intercept	46.61	10.78	4.32	0.0012	23	70
X Variable 1	−31.79	21.55	−1.47	0.1683	−79	16

It is now the intercept row that contains the required statistics. This reflects the relationship between the approaches of this and the previous section already discussed.

3.7 EXAMINING PERIOD EFFECTS*

A matter of minor interest in cross-over trials is to know whether there is a difference in the results from the two periods. As was the case for treatment effects carry-over can again affect the picture. As an extreme example of carry-over we may consider the case when a treatment (say *A*) cures the patient. In the first period half of all the patients, those being treated by *A*, will be cured, the other half being treated by *B* will still be ill. In the second period, all of the patients, even those treated by *B*, will be cured because all have received *A* at some time or other. This will not only have the effect of making *B* seem to be a better treatment than it really is but will also give the impression that the patients as a whole have undergone a secular improvement when comparing the second with the first period. By a period effect, however, we do not mean an improvement over time due to treatment but a change which would have occurred even in the absence of treatment.

Note also that period is not (except rarely) a calendar period (Senn, 1994). The exception would be when the subjects are recruited to start the trial simultaneously and the decision is made to have a fixed wash-out period. This is occasionally possible with a trial in healthy volunteers. However, patients will present when they fall ill and it will commonly be the case that some patients will have finished the trial before others have started, so that for some period 2 will be before period 1 for others. Furthermore, in some cases, especially where the cross-over trial is run as a series of individual days on which the patient is observed under treatment, as was the case with Example 3.1, a minimum wash-out period may be used to permit some flexibility in scheduling appointments. Hence, not only do different patients start the trial at different times but the interval between first and subsequent periods may differ from patient to patient. The net effect is that where period effects occur they may reflect a number of different influences: seasonal effects, changes in conditions of measurement, disease progression, habituation and so forth. These may be imperfectly reproduced in further trials so that this is yet another reason why period effects are of little interest. Nevertheless, for completeness, we now consider how they may be estimated. It will be assumed that carry-over has not occurred.

All we need to note is that the period differences have the same relationship to period effects that cross-over differences (basic estimators) have to treatment effects. We may thus use either of the two methods described above (Sections 3.5 and 3.6) for estimating treatment effects in the presence of period effects for estimating period effects in the presence of treatment effects providing we simply substitute cross-over differences for period differences and vice versa.

We have in fact already performed most of the necessary calculations. If we use the second of the two approaches we may note that for the formoterol/salbutamol sequence, the mean of the period differences is the mean of the cross-over differences and equals 30.71 ℓ/min, whereas the corresponding mean for the salbutamol/formoterol group is minus the mean of the cross-over differences and equals −62.50 ℓ/min. The estimate of the period difference (period 1 minus period 2) is thus $(30.71 - 62.5)/2 = -15.895$ ℓ/min. The standard error is obviously the same as for the treatment effect since we have simply taken the difference between two independent estimates we had previously added. Thus the t statistic is $t = -15.895/10.777 = -1.47$ and the period effect is not significant.

It may be noted that the t statistic is exactly the same as that given in the 'X Variable 1' row for the regression analysis of Section 3.6. This is no coincidence. Assuming there is no carry-over, just as the average of basic estimators may be used to estimate treatment effects, so the difference between them may be used to test for period effects. It is the intercept in this approach that corresponds to the former and the 'slope' that corresponds to the latter. The reason why the point estimates and confidence intervals for the period effect are not the same as before is that the difference in question corresponds to twice the period effect. If the values are halved they will be seen to be the same as before.

The fact that the period effect is not significant, however, does not constitute a reason for not adjusting for it and therefore for preferring the estimate for the treatment effect produced in Section 3.2 above to that produced in Sections 3.5 and 3.6. In the logic of the significance test there is no justification for asserting a hypothesis because we failed to reject it. The doubts which existed before it was tested remain. The decision to adjust or not for the period effect should be determined on *a priori* grounds (see Section 1.9).

3.8 TESTING FOR CARRY-OVER AND/OR TREATMENT BY PERIOD INTERACTION*

I wrote in the introductory chapter that I did not carry out tests for carry-over myself and did not advise the reader to do so. I nevertheless think that there is some use in understanding how such tests may be done since it helps to appreciate why they should *not* be done. The standard test for carry-over in the AB/BA cross-over will, therefore, now be described.

In the AB/BA cross-over the effects of carry-over and treatment by period interaction are not separately identifiable. This statement will be justified in due course. For the moment we shall accept that it is so, dispense with considering period by treatment interaction and consider what may be done to test for carry-over. We simply note for the moment that any test which had apparently detected carry-over might equally plausibly (in the absence of additional background information) be regarded as having detected period by treatment interaction.

In general no use can be made of cross-over differences (or for that matter of period differences) for examining carry-over. The reason is simply explained. First we may note that any difference between cross-over differences within a group will be due to either random variation, patient by treatment interaction or patient by period interaction (these last two may themselves be regarded as random if we regard the patients as randomly selected). Therefore comparing the differences within a given sequence group tells us nothing about carry-over. On the other hand we used half the difference between the means of the cross-over differences to investigate the period effect and the mean of their means (or, what comes to the same thing, half the difference between the means of the period differences) to estimate the treatment effect. Of course we did this assuming that carry-over was not a problem but whether or not carry-over is a problem these contrasts will partly measure the period or treatment effect. They cannot, therefore be used to provide a pure estimate of carry-over.

We may do this, however, using the patient totals given in the rightmost column of Table 3.1. These represent the average response in both periods for each patient. If we compare such totals for any two patients in different sequence groups we see they cannot differ by a treatment effect, for each patient has had both treatments, nor can they differ by any effect due solely to some secular trend, for each patient has been treated in both periods. If, however, the effects of treatment persist then in the second period the patient on the *AB* sequence will have a carry-over from *A* whereas the patient on the *BA* sequence will have a carry-over from *B*. Unless these two carry-over are identical the differences between the patient means will reflect carry-over.

They will also, of course, reflect any intrinsic difference between the patients. If, however, we have allocated the patients at random to the sequence groups then the *expected* intrinsic difference between patients in the two groups is zero. (This point is taken up again when discussing a model of the *AB*/*BA* cross-over in Section 3.9.) We can thus use a two-sample *t* test to examine the difference between the two sequences. The calculations for the analysis are set out in Table 3.5. We can see quite clearly that the result is not significant.

It is worth reminding oneself, however, of a basic feature of significance tests which was put by Fisher, writing in 1935, thus: 'it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation' (Fisher, 1990b, p. 16). The null hypothesis in this case is that there is no carry-over and this is a point which we should like to be able to assert. However, we know that whatever the result of the test we shall not be able to assert this and so the test is pointless.

It would be more revealing for this example to calculate a confidence interval. This we shall do in due course, but before we do that we shall need to decide on what scale to measure carry-over. The *t* statistic is dimensionless, so that in order to carry out a significance test we do not need to worry about scale.

Table 3.5 Calculations for testing for carry-over using patient totals.

	Sequence	
	formoterol/salbutamol	salbutamol/formoterol
Mean PEF (ℓ/min)	643.571	629.167
Corrected sums squares PEF (ℓ/min)^2	78 435.71	151 320.83
Pooled estimate of variance (ℓ/min)^2 $= (78\ 435.71 + 151\ 320.83)/(6 + 7 - 2) = 20\ 886.96$		
Difference in means (ℓ/min) $= 643.571 - 629.167 = 14.404$		
Standard error of difference in means (ℓ/min)$\sqrt{\{20\ 886.96(1/6 + 1/7)\}} = 80.41$		
$t = 14.404/80.41 = 0.18$		

As we saw when calculating the confidence interval for the treatment effect for the two-sample t procedure (Section 3.5) in order to make this comparable to that associated with the matched-pairs t we had to divide the difference between the period difference means by 2. Since it is really the relative importance of the true carry-over effect and the true treatment effect which is of concern we need to establish some common scale to examine them. We shall use the need to investigate this as an excuse for introducing a model which is commonly used in connection with the AB/BA cross-over.

3.9 A MODEL FOR THE *AB/BA* CROSS-OVER*

First of all we define the following symbols:

π the period effect: the expected secular difference between period 2 and period 1;
τ the treatment effect: the expected difference due to treatment between treatments A and B;
λ_A a carry-over effect due to A;
λ_B a carry-over effect due to B;
μ_i an ‘effect’ due to patient i: the response we should expect of patient i were we to treat him in period 1 with B.

We assume for the moment that there is no period by treatment interaction which is not due to carry-over. It should also be noted that the period effect in Section 3.7 was implicitly defined in terms of period 1 minus period 2 and in terms of this model is then $-\pi$. The μ_i term also requires some discussion. Its nature depends on how we look at it. It can be either fixed or random. If we work in terms of cross-over differences and basic estimators this distinction doesn’t matter since μ_i always disappears when forming these statistics. The distinction does matter for testing for carry-over, as we shall see below.

Table 3.6 Expected values of responses and associated statistics for two patients allocated to different sequences of an AB/BA cross-over.

	Expected response	
Patient	Period 1	Period 2
j	$\mu_j + \tau$ (treatment A)	$\mu_j + \pi + \lambda_A$ (treatment B)
k	μ_k (treatment B)	$\mu_k + \pi + \tau + \lambda_B$ (treatment A)

	Expected value of statistic		
Patient	Cross-over difference	Period difference	Patient total
j	$\tau - \pi - \lambda_A$	$\tau - \pi - \lambda_A$	$2\mu_j + \tau + \pi + \lambda_A$
k	$\tau + \pi + \lambda_B$	$-\tau - \pi - \lambda_B$	$2\mu_k + \tau + \pi + \lambda_B$

Now consider two patients, patient j and patient k, and suppose that patient j has been allocated at random to the sequence AB and patient k to the sequence BA. Table 3.6 gives their expected responses. (The scheme presented here is not by any means the only one possible. For example we can speak of separate treatment effects for A and B, τ_A and τ_B. If we do this, however, we shall find that it is only the difference between τ_A and τ_B which is identifiable and if we call this difference τ then it is equivalent to the representation here.)

If we represent the observation taken on patient i in period t (where $t = 1$ or 2) by Y_{it}, then we have

$$Y_{it} = E(Y_{it}) + \epsilon_{it} \tag{3.1}$$

where ϵ_{it} is a random (within patient) error or disturbance term and $E(Y_{it})$ may be established by referring to Table 3.6. By definition $E(\epsilon_{it}) = 0$. The ϵ_{it} may owe their existence to a variety of phenomena: random fluctuation in the patient's health, patient by treatment interaction, measurement error, etc. These sources are not separately identifiable and so need not concern us further here. If we assume further that the ϵ_{it} are independently Normally distributed with constant variance σ_w^2 (where the subscript w stands for *within*), then this assumption together with (3.1) and Table 3.6 constitutes a model for the AB/BA cross-over. Note that because a basic estimator (or cross-over difference) is constructed as a difference of the form $Y_{i1} - Y_{i2}$ or $Y_{i2} - Y_{i1}$, as the case may be, and because the ϵ_{it} are independent, its variance, σ^2, is equal to $2\sigma_w^2$. This model is essentially that proposed by Grizzle (1965) and for further details the reader should consult that paper, or Grieve (1987) for an extremely clear presentation of its features.

We see now if we form an estimate on the basis of the mean of cross-over differences the period effect disappears but the treatment estimate will not equal

τ as we should like but will be biased by an amount $(\lambda_B - \lambda_A)/2$, reflecting the difference in carry-over alluded to in Section 3.4 above. If $\lambda_B = \lambda_A$, then the estimator is unbiased. Undoubtedly the most reasonable circumstance (perhaps the only reasonable circumstance) under which we could be prepared to believe this is if we believed that $\lambda_A = \lambda_B = 0$, that is to say we believed that no carry-over at all took place.

Turning now to tests for carry-over, these are based on differences between patient totals as discussed in Section 3.8. Suppose that we are prepared to regard the μ_i terms as random, that is to say to regard differences between patients as random, using as justification the fact that which patients end up in which sequence group is a matter of indifference to us and was decided at random. Then apart from this random effect, the difference for patient totals between sequence groups simply estimates the difference between λ_B and λ_A. If we form the difference in the opposite manner to the way it was done for the test of significance in Section 3.8 and divide it by 2 we shall then have estimated $(\lambda_B - \lambda_A)/2$, this being exactly the term which we fear may bias our estimate of the treatment effect.

3.10 CARRY-OVER OR TREATMENT BY PERIOD INTERACTION?*

We have frequently used the terms 'carry-over' and 'treatment by period' interaction together. This is because in the *AB*/*BA* cross-over the two are not separately identifiable. We may adapt the representation of Table 3.6 to show this.

Suppose we designate the average of the μ_i over all patients as μ and then consider what the expected response would be for both sequence groups for both periods over all randomizations. We should then obtain the position in terms of expected response given in the top half of Table 3.7.

Table 3.7 Expected responses over all randomizations for an *AB*/*BA* cross-over.

	Original parameterization	
	Period	
Sequence	1	2
AB	$\mu + \tau$	$\mu + \pi + \lambda_A$
BA	μ	$\mu + \pi + \tau + \lambda_B$
	Alternative parameterization	
	Period	
Sequence	1	2
AB	$\mu + \tau$	$\mu + \pi^*$
BA	μ	$\mu + \pi^* + \tau + \lambda$

This already gives five parameters for four cell means for the table. We thus have more parameters than are separately identifiable and we have already seen that it is actually the difference between carry-overs, $\lambda_B - \lambda_A$, which we may estimate and it is in any case this difference which biases our estimates of τ. We may reparameterize by writing: $\lambda = (\lambda_B - \lambda_A)$, and $\pi^* = \pi + \lambda_A$ to obtain the position in the lower half of the table. Now if we had a period by treatment interaction so that the response to treatment in the second period was different from that in the first we could represent this by adding the symbol Δ to τ for the *BA* group in the second period, Δ representing the difference in response due to interaction. This fifth parameter would, however, not be separately identifiable from λ since each of them would appear once only and in the same cell mean. A very good discussion of this is to be found in Hills and Armitage (1979).

I have already given reasons for believing (Section 1.7) that the problem of period by treatment interaction is not unique to cross-over trials. I also gave examples of ways in which carry-over could affect parallel group trials. Nevertheless, I think it is worth making this distinction between the two. Carry-over is a problem which may much more plausibly be believed to cause difficulties in cross-over trials than for parallel group trials. Period by treatment interaction on the other hand is something which may equally plausibly affect any trial. It is just usually conveniently swept under the carpet. The only difference between cross-over trials and parallel group trials in this respect is that because we are forced to think about periods in the former we are naturally led on to thinking about interactions with periods. Time also marches on in parallel group trials. For all we know the treatment effect may be changing over time. Whether we can detect this depends on how many measurements we take and also on the length of time for which we study patients (Senn and Hildebrand, 1991). In short, although concern about carry-over is justified in cross-over trials any extra concern compared to parallel group trials with respect to period by treatment interaction is unjustified.

3.11 CONFIDENCE INTERVALS FOR CARRY-OVER*

We return now to the matter of calculating confidence intervals for carry-over. To do this we shall assume that no other period by treatment interaction takes place. Using the calculations already performed in Table 3.5 we reverse the sign of the calculated difference and divide it by 2 to obtain $-7.20\,\ell/\text{min}$. We likewise divide the standard error by 2 to obtain $40.20\,\ell/\text{min}$. The value of the t statistic on 11 degrees of freedom associated with a $2\frac{1}{2}\%$ tail probability we already know to be 2.201. The product of this with the standard error is $88.48\,\ell/\text{min}$ and adding and subtracting this from the point estimate we obtain the 95% confidence limits for carry-over as

$$-96\,\ell/\text{min} \leqslant (\lambda_B - \lambda_A)/2 \leqslant 81\,\ell/\text{min}.$$

Of course this confidence interval is neither reasonable nor useful except that it serves to point out the uselessness of the significance test for carry-over. We did not, as a result of that test, reject the null hypothesis of no carry-over but the confidence interval shows us that there are many other hypotheses regarding carry-over, including some for which the carry-over effect is quite important, which could also not be rejected on the basis of the sample data alone. Our failure to reject the original null hypothesis is no reason for asserting it. Other reasons must be found.

Part of the reason why the confidence interval is so large is because the standard error is a function of the variance between patients. If we look at Table 3.6 we see, as has already been noted, that the patient effect does not disappear from the patient totals. (As has already been remarked, the test for carry-over requires us to regard the patient effects as random. In terms of the model defined in Section 3.9, we would introduce the additional assumption that the μ_i were distributed independently of each other and the ϵ_{ij} with common mean μ and variance γ^2.) Since we use the patient totals as the basis for our calculation of carry-over the differences between patients which we hope to control for in running a cross-over trial are not eliminated from our estimate and affect adversely its variance and the power of any test associated with it (Brown, 1980). Thus, tests for carry-over are useless (Senn, 1988).

3.12 ARE UNBIASED ESTIMATORS OF THE TREATMENT EFFECT AVAILABLE?*

A very simple unbiased estimator of the treatment effect is available although it turns out not to be useful. Suppose we were to pay no attention whatsoever to the values in the second period. We could then compare the results between groups for the first period only, just as if they had been obtained from a parallel group trial. (Again we should have to treat patient effects as random.) Consideration of Table 3.6 makes it clear that such a comparison estimates τ without bias. A simple two-sample t approach enables us to estimate τ and its associated confidence limits as well as associated P values. It is left as an exercise for the reader to check that the result for the data in Table 3.1 are as follows:

- estimated value of τ: 53.81 ℓ/min;
- estimated standard error: 45.28 ℓ/min;
- 95% confidence interval: $-46\ \ell/\text{min} \leqslant \tau \leqslant 153\ \ell/\text{min}$;
- t statistic: 1.19;
- P value: 0.13.

These results show why this estimator is not useful. It is a between-patient estimator and this feature is reflected in its large standard error. The variance of the first period values is the sum of the variances of the μ_i and the ϵ_{it} terms, so that if we use ψ^2 for this variance we have $\psi^2 = \gamma^2 + \sigma_b^2$. There really is no point in carrying out a cross-over trial if we are only going to use the values from the first period.

3.13 CAN WE ADJUST FOR CARRY-OVER?*

In Section 3.11 we produced an unbiased estimate of the carry-over effect. Is it possible to use it to adjust the treatment effect? The answer is that to do so is both possible and useless.

In Section 3.6 we used the cross-over differences (or basic estimators) to estimate the treatment effect as 46.61 ℓ/min. To adjust this for carry-over we now need to subtract our estimated bias due to carry-over of -7.20 ℓ/min and this then gives the answer 53.81 ℓ/min which is identical to the result using the first period values only obtained in Section 3.12. This is no coincidence. The two approaches are absolutely equivalent, a fact which it is left as an exercise to the reader to prove. Consequently adjusting for carry-over is pointless: one might just as well calculate the first period difference directly.

3.14 THE TWO-STAGE ANALYSIS*

For many years one of the most popular approaches to analysing the *AB*/*BA* cross-over (and one which had acquired a sort of officially approved status) was the so-called two-stage approach first outlined by Grizzle (1965) and also clearly described by Hills and Armitage (1979). In a very important paper, Freeman (1989) carried out a thorough investigation of this approach and was able to show that this procedure was potentially misleading and unsatisfactory. The two-stage procedure is therefore of historical rather than scientific interest and can no longer be regarded as a serious option for analysis. Nevertheless there is a considerable literature on cross-over trials which takes it more or less for granted that this is the correct way to analyse the *AB*/*BA* design. It is therefore worth taking a brief look at this procedure in order to examine its problems.

The two-stage procedure is as follows. First the test for carry-over described in Section 3.8 is carried out. Because the power of the test is low, being based on between-patient differences, a high nominal level of significance (usually 10%) is used. If the result of this test is non-significant the standard within-patient test taking account of period effect is used as described in Section 3.6 (the CROS test in Freeman's designation). If the result is significant the first period test described in Section 3.12 is used (the PAR test in Freeman's designation). Many authors also considered that since the practitioner could never guarantee that he would not be forced into using the PAR test the cross-over trial should always be designed with enough patients to make the PAR test a credible option (Brown, 1980).

The logic of this procedure is superficially plausible. The CROS test for treatment is certainly the test we should use if we knew that carry-over were not a problem. On the other hand we can never be absolutely sure that it has not taken place so we carry out a test to check for its presence. If we 'find' carry-over we then feel unable to use the CROS test because we fear that it will be biased. We then fall back on the PAR test which is unaffected by carry-over.

A point that was overlooked by all commentators before Freeman (1989), however, was that the test for carry-over and the PAR test are highly correlated. (The test for carry-over uses the totals over both periods for each patient whereas the PAR test uses the first period values. But of course the first period values not only contribute to the total but are in any case not independent from the second period values. For Example 3.1, the estimated correlation between the first period values and the patient totals in Table 3.1 based on corrected sums of squares and products calculated within each sequence group is 0.975. Figure 3.3 provides a plot of totals against first period values which illustrates this correlation quite clearly.) A consequence of this is that although the PAR test is an unbiased test of treatment (although a very weak one) if carried out unconditionally, given that the result of the test for carry-over is significant it is highly biased. Under such circumstances a test which on the face of it has a type I error rate of 5% in fact has one which lies between 25% and 50% (Senn, 1991). The two-stage procedure thus gives results which are either identical to the CROS procedure or totally misleading.

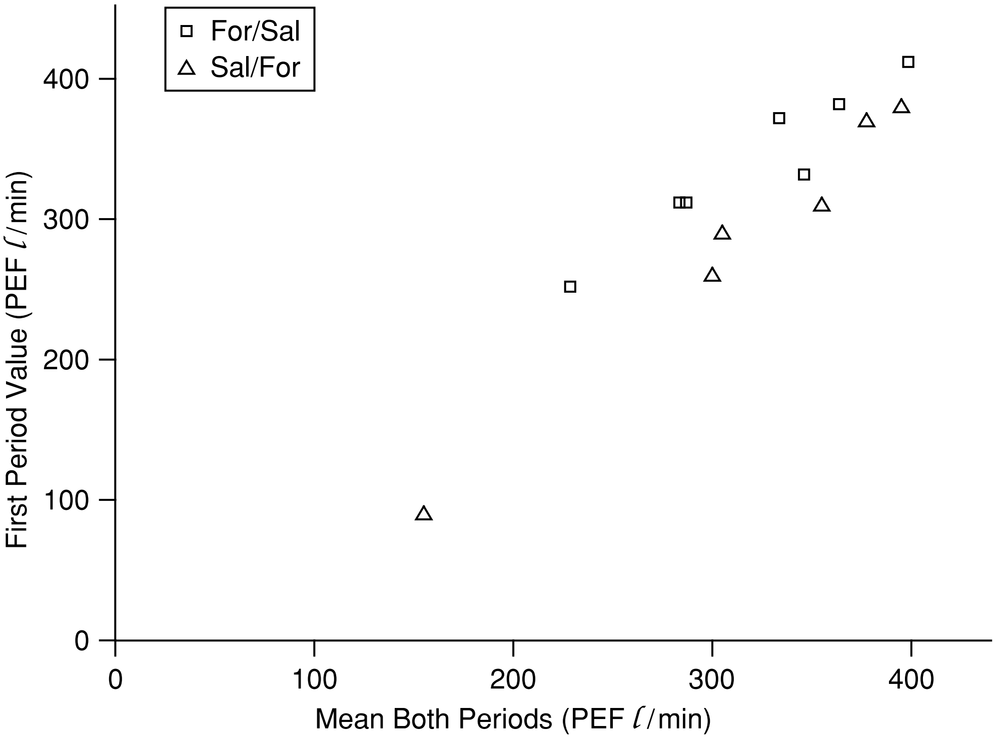


Figure 3.3 (Example 3.1) Illustration of correlation between test of carry-over and test of the treatment effect using first period values only.

Another way of understanding the problem with the two-stage analysis is as follows. Suppose we follow the advice of recruiting enough patients to perform the PAR analysis if necessary. Why not, in that case, perform the PAR test anyway? Carry-over will not then be a problem. We may then ask ourselves if any problems could effect the validity of the PAR test? One possibility would be if all the healthier patients had been recruited into one sequence group. A possible test of this would be to perform the between-patient test using the totals from each group. Of course, if this were significant we should then be reluctant to use the PAR test. One solution, however, would be to use the CROS test since, if each patient acts as his own control, it does not matter if the sequence groups are unbalanced with regard to status of health. The procedure just outlined, however, is also a two-stage analysis but one which uses the PAR test wherever the other uses the CROS test and vice versa. This then shows the problem with the conventional two-stage analysis: the more unbalanced the sequence groups are, the more likely it is to lead to the use of the PAR test, a test which (unlike the CROS test) is vulnerable to such imbalance.

Yet another argument against the two-stage procedure is in terms of the efficiency of the PAR and CROS estimators. CROS has a lower variance than PAR. Indeed, in many cross-over trials it will be much lower. Since the variances of these two estimators differ considerably, it is inevitable that the estimators themselves will often disagree (Senn, 1995b, 1996). Furthermore, since in clinical trials, for ethical and practical reasons, we rarely have high precision this disagreement will often be important in practical terms. However, unless where CROS and PAR disagree substantively we prefer CROS, there is no point in running the cross-over trial because that would imply that CROS was redundant and hence that the second period of study was unnecessary. But PAR, CROS and CARRY satisfy the relationship CROS $=$ PAR $-$ CARRY/2, so where PAR and CROS disagree CARRY will be large and possibly significant and quite plausibly PAR will be quite unreliable. Hence, testing for carry-over will lead us to use the least reliable of the two statistics where they disagree.

There is also some empirical evidence concerning carry-over that encourages the approach of not testing. In a remarkable paper, D'Angelo *et al.* (2001) analysed a series of 324 *AB*/*BA* designs in bioequivalence testing for carry-over for both area under the concentration curve (AUC) and concentration maximum ($C_{\max}$) at the 10% level. There was significant carry-over in only 37 trials for AUC and in 34 for $C_{\max}$, almost exactly what was expected under the null. For both measures the distribution of P values was remarkably even. Of course, bioequivalence is an application area in which it is easier than most to ensure adequate wash-out and guard against carry-over. Nevertheless, it is still common in some quarters to test for carry-over in this area, showing that far too much attention is being placed on inherently unreliable statistics and not enough on prior considerations.

Thus, in conclusion, the two-stage analysis should not be used.

3.15 CORRECTING THE TWO-STAGE PROCEDURE*

It is possible to correct the two-stage procedure so that the overall Type I error rate for testing the null hypothesis of no treatment effect does not exceed 5% (Senn, 1996, 1997a). A simple conservative procedure can be based on the following argument.

As is conventional in discussing probability, we write $P(A \cap B)$ for 'the probability of A and B' and $P(B|A)$ for 'the probability of A given B'. We let A stand for 'CARRY is significant under H_0' and B stand for 'PAR is significant under H_0'. Note that $P(B)$ is the *nominal* level of the PAR test for treatment, because it is *not* conditional on CARRY, whereas $P(B|A)$ is the *actual* level, conditional on CARRY being significant.

Now, for any two events A and B we have the identities $P(A \cap B) = P(A)P(B|A)$ and $P(A \cap B) = P(B)P(A|B)$. From the first of these we have $P(B|A) = P(A \cap B)/P(A)$, provided that $P(A) \neq 0$, and from the second we have $P(A \cap B) \leqslant P(B)$, since $P(A|B) \leqslant 1$. Hence, putting these two together, we may write

$$P(B|A) \leqslant P(B)/P(A). \tag{3.2}$$

Now we require that the probability under H_0 of a significant result for PAR given that CARRY is significant should be less than or equal to some specified Type I error rate, say α_2, where typically we have $\alpha_2 = 0.05$. However, in terms of our symbols, this is to require that $P(B|A) \leqslant \alpha_2$. But this is clearly satisfied if the right-hand side of (3.2) is set equal to α_2. Hence, we require

$$P(B)/P(A) = \alpha_2$$

or

$$P(B) = P(A)\alpha_2.$$

Now suppose that we have carried out our test for carry-over with a Type I error rate of α_1, where typically we have $\alpha_1 = 0.1$. Then we simply have to set $P(B) = \alpha_1\alpha_2$. In other words, provided that we set the nominal Type I error rate for the test for PAR equal to the product of the level for CARRY and the desired level for PAR we have a valid procedure. In practice, using the usual levels, this requires us to carry out PAR at the $0.1 \times 0.05 = 0.005$ level.

The greatest bias in the conventional two-stage procedure occurs when the correlation between CARRY and PAR is 1. Under such circumstances, under the null hypothesis of no treatment effect and no carry-over effect (the second hypothesis being plausibly implied by the first), whatever the P value for CARRY, the P value for PAR will be the same. Since in practice the level of significance for CARRY is higher than that for PAR, this simply means that if

PAR is significant CARRY will be. Hence, we have $P(A|B) = 1$ and thus $P(A \cap B) = P(B)$, in which case (3.2) becomes an equality and the procedure is not conservative. For other cases, the procedure is conservative. In other words, if we drop our nominal level of significance to a tenth of the usual value, to 0.5% rather than 5%, we have a test whose conditional size cannot be greater than 5%, this limit only being reached when the correlation is 1.

If one is going to perform a two-stage analysis, then something like this needs to be done. However, this procedure cannot be recommended. Its interest is purely theoretical. This interest arises because the *uncorrected* two-stage analysis has been compared by some researchers to the simple strategy of always using CROS. It has been claimed to have superior power for the sort of sample size often encountered in cross-over trials if the carry-over is at least half the size of the treatment effect (Jones and Lewis, 1995). However, this comparison is illegitimate since two procedures with unequal type I error rates are being compared (Senn, 1997a). If that is permissible, then the procedure of declaring the treatment significant without even analysing the data is superior still, since this procedure has 100% power. If one is going to make a comparison in terms of standard Neyman–Pearson theory, it is necessary to compare procedures which hold to the same Type I error rates. A legitimate comparison would be that of the *corrected* two-stage analysis described above and CROS. When this is done, the power advantage of the two-stage analysis disappears for all plausible scenarios. In short, it turns out that the only apparent advantage of the two-stage analysis was one it claimed unfairly: the power was increased by increasing the Type I error rate.

As already explained, my advice is *not* to use the two-stage procedure even if the necessary correction described above is made. However, those readers I have failed to convince may like to have a look at the papers already cited (Senn, 1996, 1997a) or the paper by Wang and Hung (1997).

3.16 USE OF BASELINE MEASUREMENTS

It will sometimes be possible to make baseline measurements in cross-over trials. Baselines are measurements made on the patient with the object of giving general or background information rather than direct information on treatment. Their use may, however, increase indirectly the precision of measurements made directly on treatment. In parallel group trials, for example, baseline measurements are frequently used in analysis of covariance to increase precision.

In the *AB*/*BA* cross-over trial we may distinguish three kinds of baseline: those taken before the start of the first treatment, those taken after the completion of the first treatment and before the start of the second treatment and those taken after completion of the second treatment. Strictly speaking only the first kind is a true baseline (Kenward and Jones, 1987a): there is always the possibility that due to carry-over the second or third kind may reflect the previous treatment.

We shall now consider the use which may be made of these measurements, taking the most important case first: that of two baselines, one before each treatment.

3.16.1 Two baselines

In what follows it will be assumed that the baselines are truly concomitant, that is to say they do not reflect treatments under investigation. Obviously this must be true of baselines before the first treatment but those taken after the first but before the second treatment could reflect the first treatment via carry-over. Some authors have used this fact to propose estimators adjusted on a within-patient basis for carry-over. In practice, however, unless the wash-out period is long compared to the treatment period, carry-over at the end of the second treatment period is unlikely to be what it was at the beginning and if there is a long wash-out there is unlikely to be any carry-over anyway. This is not, therefore, a realistic use of baselines. Thus, we now consider the case where the wash-out is regarded as adequate to eliminate carry-over and the baseline values are regarded as carrying no direct information on any of the treatments.

If the cross-over differences for the baselines in Table 3.8 are compared to the basic estimators there appears to be a correlation between the two. This is shown quite clearly by the plot of one against the other given in Figure 3.4.

Table 3.8 (Example 3.1) Peak expiratory flow in litres per minute: baselines and basic estimators for 8 hour values.

		PEF			
		Baselines			Basic estimators for 8 hours
Sequence	Patient number	Period 1	Period 2	Cross-over difference	
for/sal	1	290	270	20	40
	4	300	270	30	50
	6	250	210	40	70
	7	390	390	0	20
	10	250	240	10	40
	11	365	380	−15	30
	14	190	260	−70	−35
sal/for	2	350	345	−5	15
	3	350	370	20	90
	5	350	360	10	30
	9	280	290	10	30
	12	270	310	40	80
	13	220	220	0	130

for/sal = formoterol followed by salbutamol
sal/for = salbutamol followed by formoterol

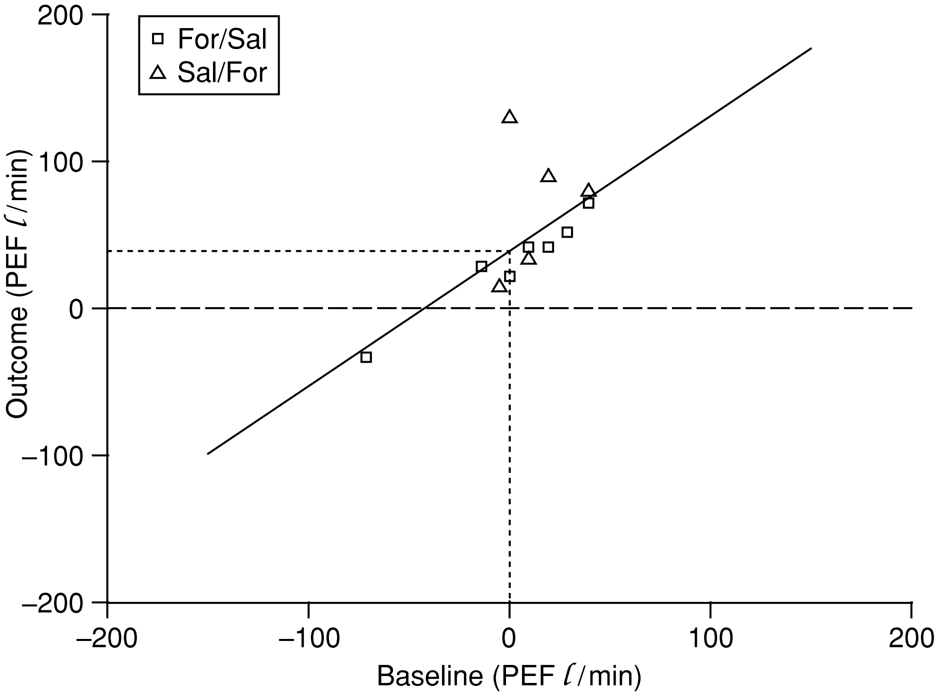


Figure 3.4 (Example 3.1) Illustration of analysis of covariance. Basic estimators for outcomes plotted against baseline differences.

This may be because there is a general trend effect affecting both baselines and outcomes. If we saw a period effect on outcomes we should then see a similar period effect on baselines. Even if we compare within sequence groups, however, it still seems to be the case that a higher cross-over difference for baselines is associated with a higher value of the baseline estimator. One possible explanation is that patients are subject to individual trend effects over the whole experiment (some are deteriorating whereas others are improving). Were this to be the case the baselines would contain useful information.

One simple way of using the information, which is appropriate if the baselines are believed to be strongly predictive of outcome, is simply to adjust the basic estimators by subtracting the cross-over differences for the baselines. (Obviously this is equivalent to subtracting the corresponding baseline from its associated outcome value and then forming the basic estimators from these differences.) A disadvantage with this procedure is, however, that if the baselines are not strongly predictive of outcome then the variability of the resulting estimator may well be increased. This seems not to be the case here. The reader may check for himself that if the baselines are used in this way the resulting treatment estimate is 39.3 ℓ/min with an estimated standard error of 8.4 ℓ/min. The baselines thus appear to have provided useful information in that they have led to a reduced standard error. This will not be the case generally, however, and the important point to note is that it is the strength of the relationship between baselines and outcomes once patient and period effects have been taken into account that matters. A similar result applies to that which we discussed in Section 2.2 when considering subtraction of baselines: if the partial correlation between baseline and outcome is not greater than 0.5 then simple adjustment by subtraction is counter-productive (Senn, 1989a).

A better procedure is to adjust the outcomes using *analysis of covariance*. (In practice baselines are the only covariates which may be used in this way in cross-over trials since they are the only covariates which change during the trial and are therefore not identical for each treatment.) We shall now describe how this may be done. It should be noted, however, that nowadays hardly anybody does this sort of calculation by hand and if they do they probably do not do it in the way illustrated below. The exercise will be gone through for its didactic value, not because it should be regarded as a practical option. In order to simplify the exposition the period effect will not be allowed for. The way that the analysis can be implemented on the computer (including allowing for the period effect) will be discussed later in the chapter.

The necessary calculations are given in Table 3.9. The baseline cross-over differences are regarded as independent explanatory variables and the regression of the basic estimators is calculated on these. The slope estimate $\hat{\beta}$ is 0.919 14 suggesting a very similar correction to simply subtracting the baseline cross-over difference. (The line of best fit is shown in Figure 3.4.) The regression coefficient is remarkably close to 1 and values this high are not in general to be expected. The column headed Y^* gives the basic estimators adjusted by subtraction of 0.91914× the corresponding baseline differences. The mean of the

Table 3.9 (Example 3.1) Analysis of covariance for the 8 hour PEF data.

Patient	Baseline cross-over X	Basic estimator Y	Products XY	Baseline-corrected basic estimators Y^*
1	20	40	800	21.62
4	30	50	1500	22.43
6	40	70	2800	33.23
7	0	20	0	20
10	10	40	400	30.81
11	−15	30	−450	43.79
14	−70	−35	2450	29.34
2	−5	15	−75	19.60
3	20	90	1800	71.62
5	10	30	300	20.81
9	10	30	300	20.81
12	40	80	3200	43.23
13	0	130	0	130

Statistics (PEF in ℓ/min or PEF^2 in ℓ^2/min^2)

$\sum X = 90$ $\bar{X} = 6.923$ $\sum Y = 590$ $\bar{Y} = 45.385$ $\sum X^2 = 10\,350$ $\sum XY = 13\,025$

$\sum(X - \bar{X})^2 = 9726.923$ $\sum(X - \bar{X})(Y - \bar{Y}) = 8940.385$

$\hat{\beta} = 8940.385/9726.923 = 0.91914$ $\hat{\tau} = \bar{Y}^* = 39.0$ $\sum(Y^* - \bar{Y}^*)^2 = 11\,555.3$

$\hat{\sigma}^2 = 11\,555.3/11 = 1050.5$

est.var($\hat{\tau}$) $= (1050.5 \times 10\,350)/(13 \times 9726.923) = 86.0$ est. $SE(\hat{\tau}) = 9.27$

adjusted values gives the estimated treatment effect $\hat{\tau}$. This might have been obtained directly as $\bar{Y} - 0.919\,14\bar{X}$, the intercept of the regression line. (Indeed the geometrical construction of this value, being the predicted outcome for a baseline difference of 0, is illustrated in Figure 3.4.) It is interesting to see, however, how the individual basic estimators after adjustment are all positive. Furthermore the corrected sum of squares of these adjusted values when divided by the degrees of freedom, which in this case are 11, being reduced from 13 by 2 (one for the overall mean and one for the regression), gives an estimate of the residual variance. This cannot, however, be directly divided by 13 to obtain the variance of the treatment effect as we did when using the matched-pairs t test in Section 3.2 because the 13 corrected estimates are not entirely independent. In general terms the required formula is

$$\mathrm{var}(\hat{\tau}) = \{\sum X^2/(n\sum(X - \bar{X})^2\}\sigma^2. \tag{3.3}$$

Only if $\sum X^2 = \sum(X - \bar{X})^2$ does this reduce to σ^2/n; otherwise it will exceed σ^2/n, but this condition is itself revealing since the uncorrected and corrected sums of squares are only identical if the mean is 0. The mean of the cross-over differences will be equal to zero, however, only if, on average, the baselines prior to treatment with formoterol are equal to those prior to treatment with salbutamol. If this is not the case it indicates that the treatments are unbalanced with respect to baselines (there is nothing we can do as experiments to ensure the treatments will have equal baselines except to control the conduct of the experiment to the best of our ability) so the ratio,

$$\sum X^2/\sum(X - \bar{X})^2, \tag{3.4}$$

which may also be expressed as

$$1 + Z, \tag{3.5}$$

where $Z = \bar{X}^2/\{\sum(X - \bar{X})^2/n\}$, the ratio of the square of the mean to the (uncorrected) variance for the baseline cross-over differences, represents a penalty we must pay for the observed lack of orthogonality or balance. On the other hand, the reduction achieved in $\hat{\sigma}^2$ by fitting baseline differences as a covariate is a reward for carrying out analysis of covariance.

A point, therefore, which applies generally to analysis of covariance using prognostic information in parallel group trials applies also to analysis of covariance using baselines in cross-over trials. If we fit covariates which are only very weakly prognostic we shall find that if they are badly imbalanced the estimated variance of our treatment estimate will be increased, possibly quite unnecessarily so, the reason being that if the covariate is heavily confounded with the treatment group there is great difficulty under the model of deciding whether any observed difference in outcome between the groups is due to treatment or covariate. We might in running a cross-over trial find that the baselines are badly imbalanced with respect to treatment. If we knew that a difference in baselines was of little use in predicting a difference in outcome this would not disturb us. If we carry out analysis of covariance, however, the model will do its best to separate the effects of baseline imbalance and treatment, and if these are heavily confounded, have difficulty in doing so and thus impose a penalty in increased variance. (Of course, if we do regard the baselines as important this increase in variance is absolutely right. Confounding is a genuine problem and the fact that estimates using analysis of covariance have high variances, due to confounding, is not a reason for not doing analysis of covariance. The justification for not using analysis of covariance in such circumstances must be external.)

In summary, therefore, it is recommended that the following questions be put before looking at the data.

- Is it reasonable to suppose that all effects of carry-over will have disappeared once the second baseline has been taken? If not, do not use the baselines.
- Is it felt that baselines will contain useful information not already accounted for by adjusting for patient and period effects (perhaps because patients may have individual trend effects)? If not, do not use the baselines.
- If, however, the answer to the above two questions is ‘yes’, perform an analysis of covariance using the baselines.

3.16.2 Case 2: measurements before first treatment only

In general, if there are n patients in an AB/BA cross-over (and if they all complete both treatment periods), there are $2n$ outcome results available. If we only take baseline measurements before the first treatment period, however, we shall have only n baseline measurements available. For each patient there will be two outcome measurements, one for each treatment, but only one baseline. This means that such baseline measurements cannot be used to tell us anything directly about treatment effects. Suppose, for example, we subtracted the patient’s single baseline measurement from each outcome measurement, we should find, when proceeding to calculate the cross-over differences, that the effect of the baselines would totally disappear and the cross-over differences would be just the same as if the baselines had never been used.

A use which is possible for such single baseline measurements is similar to the one made of information regarding sex in Section 3.4. In the same way that it was possible to examine differences in the treatment effect according to the sex of the patient it is possible to check whether the treatment effect varies according to the patient’s baseline state. This corresponds to an analysis of baseline by treatment interaction. In principle such an approach is possible for all covariates measured on the patient.

There is an alternative use, however, which may be made of the baseline measurement. Even if a baseline can contribute no information about the treatment effect it might, under certain fairly strong assumptions, be able to provide information about variability. Thus we leave our treatment estimate unchanged but use the extra measurements to provide an estimate of the within-patient variability based on $3n$ observations rather than $2n$.

These analyses will not be illustrated here but are covered in Section 3.18.

3.16.3 Case 3: measurements taken after first and second treatments

Suppose we have a trial with a fixed wash-out period. Suppose we allow the same fixed interval to elapse after the last treatment period and that we take a measurement once this second wash-out period has elapsed. We then have 'baseline' measurements taken after each treatment period. These could be analysed using the standard approach to analysing ordinary outcomes in order to examine the possibility of any carry-over effects persisting to the end of the wash-out.

The problem is, however, that, even so, interpretation will be extremely difficult. There is bound to be some (although it may be small) correlation with the within-patient treatment estimates. The test will have the power of a within-patient comparison but this does not mean that simply because carry-over is not found that it is not there. I am thus rather reluctant to recommend this procedure. It might be of some value in gaining background information for future trials. It is unlikely to be of much help for interpreting the trial in hand.

3.17 A BAYESIAN APPROACH*

A very simple Bayesian criticism of the two-stage procedure can be given. That is that it is an attempt to resolve an initial doubt as to whether carry-over has occurred or not by collecting an unrealistically small amount of information. This information is used to test for carry-over. Once the test has been performed, the trialist then either behaves as if it were known for sure that there was no carry-over and uses the CROS statistics, or as if nothing whatsoever were known and uses the PAR statistic (Grieve and Senn, 1998).

A possible Bayesian alternative is to base inferences on a weighted combination of the two approaches, where the weights depend on both prior belief and the evidence regarding carry-over. The approach we present here is a modification of that proposed by Grieve (1985; Grieve and Senn, 1998). The modification allows a considerable simplification in analysis and also has the effect of making it more robust (Senn 2000a). Under optimal assumptions it loses in efficiency compared to Grieve's approach.

The details of Grieve's original paper are too complex to summarize here. A broad overview will be given. The key to the approach is the carry-over parameter, λ. If this is known to be zero, then, given suitable 'uninformative' priors, inference about the treatment effect, τ, may be made using the CROS statistic and a conventional frequentist confidence interval based on the t distribution may be given a Bayesian interpretation. If nothing is known about the carry-over parameter, then, given uninformative priors for the other parameters also, a suitable point estimate of the treatment effect is provided by PAR. However, the standard t distribution does not provide a means to produce the relevant Bayesian credible interval. This is because, although the first period data carry all the information that there is about the effect τ, they do not provide all the information that there is about the variance ψ^2. If we assume that the variances of first period and second period values are identical, then even though the second period data carry no information about τ, they are informative about ψ^2. However, they do not provide independent information about ψ^2, since the first and second period data are correlated with correlation $\rho = \gamma^2/\psi^2 = \gamma^2/(\gamma^2 + \sigma_b^2)$. The influence of the nuisance parameter, ρ, means that we have an example of a Fisher–Behrens problem (Sprott and Farewell, 1993). Dealing with this problem is controversial in the frequentist framework. In a Bayesian framework it introduces more complexity in integration and does not produce a t distribution. The final part of Grieve's procedure is that inferences from these two approaches are mixed by using the posterior probability of carry-over as calculated from the so-called Bayes factor.

The approach outlined below will follow Grieve's with one exception (Senn, 2000a). The assumption that the first and second period variances are identical will be relaxed. This leads to a considerable simplification and permits use of the conventional t distribution for confidence/credible intervals for τ using PAR. We illustrate the approach using Example 3.1.

In addition to estimates of the various parameters, we shall need to calculate the so-called within-sum of squares, *SSE*. This can be calculated quite simply as half of the corrected sum squares of the cross-over differences that we calculated in Section 3.5. In other words, the mean of the two quantities given in Table 3.4. This quantity stands in relation to σ_w^2 as the corrected sum of squares in Section 3.5 does to $\sigma^2 = 2\sigma_w^2$. We thus calculate $(6521.43+9987.5)/2 = 8254.5$. We also need the between-patient sum of squares, *SSP*. This can be based upon the analysis of Section 3.8 and is equal to the corrected sum of squares based on the patient totals divided by 2. The figures can be obtained from Table 3.5 and the result is $(78\,435.71+151\,320.83)/2 = 114\,878.3$.

Define $q = 1/n_1 + 1/n_2$, $N = n_1 + n_2$ and let $T(t, \nu)$ be the probability density of a t distribution with ν degrees of freedom. We write SE_{CROS} for the estimated standard error of *CROS*, SE_{PAR} for the estimated standard error of *PAR* and SE_{CARRY} for the standard error of *CARRY*. We write P_0 for the probability that the model without carry-over is 'correct' and κ for the prior odds in favour of this model and against the model with carry-over. $(CARRY/SE_{CARRY})^2$ is an F

statistic that will be used in calculating the Bayes factor, and since this is the only F statistic we shall be calculating, we shall simply denote it by F. It is, of course, the square of the t statistic used for testing for carry-over. (See Chapter 2 for an explanation of relationships between random variables.) For Example 3.1, the value of F may be calculated by squaring the t statistic given in Table 3.5 and is 0.0321.

First, we obtain the posterior distribution under the assumption of no carry-over. Under these circumstances, *CROS* is an unbiased estimate of the treatment effect, τ, and has variance $\sigma^2(1/n_1 + 1/n_2)/4 = q\sigma^2/4 = q\sigma_w^2/2$. Given 'uninformative priors' for mean and variance, this leads to the standard result for the marginal posterior distribution of a mean, namely that it is given by a scaled t distribution. (Scaled in the sense that the probability density is a multiple of a t distribution if we wish to make inferences upon the original rather than the standardized scale of measurement.) In fact if $T(t, \nu)$ is the probability density of a random variable, t, that has Student's t distribution with ν degrees of freedom, we have

$$f_{CROS}(\tau) = \frac{1}{SE_{CROS}} T\left(\frac{\tau - CROS}{SE_{CROS}}, N - 2\right) \tag{3.6}$$

for the posterior probability density of τ based on *CROS*.

Where, on the other hand, we have an uninformative prior for carry-over, an appropriate estimate for t is *PAR*, and since we have relaxed the assumption of equality of variances, there is no information regarding variability from the second period data that is of relevance to inferences using the first. The analogous equation to (3.6) is now

$$f_{PAR}(\tau) = \frac{1}{SE_{PAR}} T\left(\frac{\tau - PAR}{SE_{PAR}}, N - 2\right). \tag{3.7}$$

To make use of these two distributions, we now calculate the so-called Bayes factor, B. When comparing two models, this is the ratio by which the prior odds in favour of one model and against the other must be multiplied to get the equivalent posterior odds (see O'Hagan, 1994, Chapter 7). Following Grieve (1985; Grieve and Senn, 1998) we calculate

$$B = \left(\frac{3}{2q}\right)^{1/2} \left(1 + \frac{F}{N - 2}\right)^{-N/2} \tag{3.8}$$

for the odds against the model with the carry-over. Here we have $F = 0.0321$, $N = 13$ and $q = 0.3095$. Hence, we calculate $B = 2.16$.

To use (3.8) to calculate the probability, P_0, of the model without carry-over being 'correct', we have to assume a value for the prior odds, κ. Given such a value, we may calculate

$$P_0 = \frac{\kappa B}{1+\kappa B}, \qquad 1 - P_0 = \frac{1}{1+\kappa B}. \tag{3.9}$$

For example, if we have prior odds of 1, corresponding to both models being equally likely, then we calculate $P_0 = 2.16/3.16 = 0.68$ and hence $1 - P_0 = 0.32$. These posterior probabilities may then be used as weights to mix the probability densities given in (3.6) and (3.7).

Figure 3.4 illustrates these two distributions and also the weighted average of them both using the value of 0.68 for P_0 given above. The 95% credible (confidence) limits are marked for the original distributions. The weighted distribution can be calculated quite easily, for example, using a spreadsheet. One simply prepares two columns corresponding to the densities given in (3.6) and (3.7) and then a third, which is the weighted average of the two. This Bayesian posterior is also calculated.

Calculating credible intervals from this distribution appears to be not so easy, involving as it does inversion of the weighted sum of the two t-integrals, but in fact they can easily found by trial and error using a spreadsheet. The key is that it is a fairly simple matter to calculate the posterior probability that the treatment effect exceeds any given amount. For this we simply have to work with the two t statistics treating the point estimate as known and the parameter as a random variable. In general, we can use probability statements of the form

$$P(\tau > \tau') = P\left(\frac{\tau - \hat{\tau}}{SE(t)} > \frac{\tau' - \hat{\tau}}{SE(t)}\right) = 1 - P\left(t < \frac{\hat{\tau} - \tau'}{SE(t)}\right)$$

Suppose, for example, that we wish to calculate the probability that the treatment effect exceeds 10. Then we can set out the calculations as given in the Excel® spreadsheet below. In this, the bold figures are input statistics that do not change. The italic figure 10 in the highlighted cell is the tentative parameter value, which we can change. The spreadsheet has been set up so that when it is changed in the column headed CROS, the corresponding value in the column headed PAR also changes to the same value. For a Bayesian who believes there can be no carry-over, inferences should be based on CROS and the requisite probability is $P(\tau > 10) = 0.997$. For the Bayesian who believes carry-over could be anything at all, PAR yields $P(\tau > 10) = 0.823$. For the Bayesian who had prior odds of evens regarding the models, these two models now have posterior probability of 0.68 and 0.32 respectively, so the answer is calculated as

$$P(\tau > 10) = 0.68 \times 0.997 + 0.32 \times 0.823 = 0.941.$$

This is the output from the spreadsheet and is in the cell in the lower right-hand corner.

	CROS	PAR	Combined
Parameter Value	*10*	10	
Point estimate	**46.61**	**53.81**	
SE	**10.78**	**45.28**	
t Statistic	3.40	0.97	
DF	11	11	
Probability	0.997	0.823	
Weight	**0.68**	**0.32**	
Weighted Probability	0.677974	0.263349	0.941

Now, by changing the input cell to try other values of the parameter apart from 10, we can establish the 95% credible intervals. We require the value that produces an answer of 0.975. This is found to be -15. Similarly, the other limit can be found to be 122. Hence, our 95% credible interval is -15 ℓ/min to 122 ℓ/min.

Remark It should not be assumed that this *result* is reasonable. This result, but not the method in general, depends entirely on the analyst accepting that prior odds of evens are appropriate. In practice, it will be necessary to think carefully as to what value of κ is indicated and calculate the posterior odds accordingly. This calls for very careful consideration and is not easy. The reason is as follows. There are very few circumstances under which it would be reasonable to believe that carry-over might be anything at all. Even if carry-over were believed to be a distinct possibility, circumstances under which one might believe it would be very large (larger than the sought-for treatment effect, for example) would be rare. On the other hand, the probability of it being literally zero might also be low. That being so, to use Grieve's approach (or its modification here) it is generally necessary to give more weight to the model in which carry-over is zero than one might actually believe would be the case (Grieve and Senn, 1998).

3.18 COMPUTER ANALYSIS

These days nobody but a masochist or a diehard traditionalist will regularly carry out the analysis of cross-over trials by hand. The reason so much emphasis has been placed on manual calculation so far in this chapter has been to illustrate what the various forms of analysis achieve. We shall now illustrate various approaches to computation using the SAS® computer package. (Analyses have been performed using version 8.1 but should also work in version 6.12.) Program for analyses in S-Plus® and GenStat® are included in appendices to this chapter. Various analyses with Excel® have been covered earlier in the chapter.

3.18.1 Fixed effects analysis of original data

Proc glm is a procedure within SAS® which may be used to analyse unbalanced experiments using *ordinary least squares* (*OLS*). It can handle a mixture of categorical variables (such as periods, patients and treatments) and continuous variables (such as baselines). In the first edition of this book *proc glm* was used exclusively for analysis. Since the book appeared, *proc mixed* has come into common use for analysis and, indeed, there is now an excellent book in the same series as this one giving extensive advice on its use for analysing mixed models (Brown and Prescott, 1999). The *proc mixed* procedure will be used for random effect analysis below. In this section we use *proc glm*. We start by illustrating one very simple approach.

First we assume that the outcome measurements on the n patients have been gathered togethered as $2n$ observations under one variable called *OUTCOME*. We also have a variable called *PATIENT* which has n values each repeated twice, a variable called *PERIOD* which has 2 values each repeated n times and a third called *TREAT* with two values each repeated n times. For some of the other analyses considered in this chapter we should also have to record the *SEX*, the sequence *GROUP* and the *BASE*line values. For the moment we do not concern ourselves with these. If we print the values out in columns under their variable headings then each row records data which apply to a given patient in a given period. Thus, for example, for the PEF data considered in this chapter, for patient 4 in period 2 we have:

OUTCOME	*PATIENT*	*PERIOD*	*TREAT*
260	4	2	SAL

showing that a PEF reading of 260 ℓ/min was obtained on patient 4 in period 2, at which time he was being treated with salbutamol.

The analyses can then be carried out using three simple statements: the *class, model* and *estimate* statements within *proc glm* of SAS®. We need to use the *class* statement to specify which variables in the analysis are categorical, the *model* statement to specify which variable is being 'explained' and which variables do the 'explaining' and the *estimate* statement to specify particular contrasts of interest. For example, the code

```
proc glm;
  class PATIENT TREAT;
  model OUTCOME = PATIENT TREAT;
  estimate "for-sal" TREAT 1 - 1;
run;
```

produces an analysis along the lines of the matched-pairs t described in Section 3.2 which includes in its output the lines:

Parameter	Estimate	T for H_o: Parameter = 0	Pr > $\|T\|$	Std Error of Estimate
for-sal	45.384 615 38	4.03	0.0017	11.252 835 21

Apart from the additional number of significant figures provided these results are the same as those found earlier in Section 3.2. The uses of the *class* and *model* statements are obvious and require no explanation. The estimate statement requires three pieces of information: first, a label for the estimated contrast (*for-sal* in this example), second, the factor whose levels are being used in the contrast (*TREAT* in this example) and third, the weights to be used on those levels in estimating the contrast (1 for formoterol and −1 for salbutamol in this example).

If *PERIOD* and/or *BASE* are to be included in the analysis they should be included in the *model* statement above. *PERIOD* should also be included in the *class* statement. When both of these factors are fitted in the model the output includes the lines:

Parameter	Estimate	T for H_o: Parameter = 0	Pr > $\|T\|$	Std Error of Estimate
for-sal	40.459 261 37	4.46	0.0012	9.067 548 98

This is not an analysis which was attempted by hand. The similarity of the results to those obtained in Section 3.16.1 (where the *PERIOD* effect was not fitted) is evident.

If we wish to calculate 95% confidence intervals using this output we need to check how many degrees of freedom for error our analysis has left us. Another part of the SAS® output itself informs us how many degrees of freedom for error there are. Alternatively a table on the lines of Table 3.10 may be constructed to account for them.

The degrees of freedom for error are thus 10. From statistical tables we find that the value of the t associated with 10 degrees of freedom for a $2\frac{1}{2}\%$ tail probability is 2.228. The product of this with the estimated standard error is 20.2 and so the 95% confidence interval for the treatment effect is

$$20\,\ell/\mathrm{min} \leqslant \tau \leqslant 61\,\ell/\mathrm{min}.$$

3.18.2 Random effects analysis of original data

The representation of patient effects in Table 3.10 is the one I usually use. This is because I am not interested in testing for carry-over. The reader who does not

Table 3.10 Accounting for degrees of freedom in an analysis with baselines.

Source	Degrees of freedom In general	 This example
Mean	1	1
Patients	$n-1$	12
Periods	1	1
Treatments	1	1
Baselines	1	1
Error	$n-3$	10
Total	$2n$	26

share my opinion on this subject will require a different representation of the patient effect; he will need to split this into two sources: *GROUP* and *PATIENTS* within *GROUP*. (It will be necessary to include a variable which identifies for each of the $2n$ observations to which sequence *GROUP* the particular patient on whom that observation was made belongs.)

We will now illustrate the use of *proc mixed*. For the analysis fitting *PERIOD* effects in addition to *TREAT*ment the SAS code is then as follows:

```
proc mixed;
  class GROUP PATIENT PERIOD TREAT;
  model PEF= GROUP PERIOD TREAT;
  random PATIENT(GROUP);
  estimate "treatment" TREAT 1 −1;
  estimate "carry-over" GROUP 1 −1;
run;
```

The syntax here is scarcely different from that using *proc glm*. The *model* statement has been amended to split the variation into two sources: that between groups (or sequences), *GROUP*, and the variation between patients within groups, *PATIENT (GROUP)*. This split is made necessary because we wish to use the difference between sequence groups to say something about carry-over. In addition, the *random* statement has been used to identify *PATIENT*s as a random effect.

As part of its output SAS® now includes the following:

Tests of fixed effects

Source	NDF	DDF	Type III F	Pr > F
GROUP	1	11	0.03	0.8611
PERIOD	1	11	2.17	0.1683
TREAT	1	11	18.70	0.0012

Here NDF stands for numerator degrees of freedom and DDF for denominator degrees of freedom. The last column, Pr > F, gives the P-values for the three tests. The corresponding analysis with *proc glm* fitting patient, period and treatment effects would include in its output an analogous table of results as follows:

Source	*DF*	Type III SS	Mean Square	*F* Value	Pr > *F*
PATIENT	12	115 213.461 538 46	9601.121 794 87	12.79	0.0001
PERIOD	1	1632.074 175 82	1632.074 175 82	2.17	0.1683
TREAT	1	14 035.920 329 67	14 035.920 329 67	18.70	0.0012

It will be noted that for *PERIOD* and *TREAT* the *F* statistics (referred to as Type III F in the one case and *F* value in the other) are identical for these two approaches, as are the *P* values. Indeed, these *P* values are the same as those associated with the *t* tests we carried out for the treatment effect adjusted for the period effect, or for the period effect when performing the calculations by hand earlier in this chapter (see Sections 3.5 or 3.6 and 3.7). The squares of the *t* statistics we calculated there, 4.33 (treatment) and −1.47 (period), are 18.7 and 2.2. The difference between the two approaches comes in the handling of the patient effect. Patients within groups are treated as random in the analysis with *proc mixed*. This permits a test for carry-over, the results of which are given by the *P* value associated with *GROUP*. This is found to be 0.8611. The *F* statistic of 0.03 is the square of the *t* statistic for carry-over of 0.18 given in Section 3.8.

Remark As already explained, my advice is not to test for carry-over, the reason being that the trialist is likely to overreact to the apparent discovery that carry-over has occurred.

If carry-over is not tested for, then there is usually no advantage in treating patient effects as random and using *proc mixed*, compared to treating patient effects as fixed and using *proc glm*. An exception is if there are very many patients with data missing in the second period. The data available may then conceptually be split into two parts, data for all patients who completed both periods and data for all patients who completed the first period only, the former having the structure of a cross-over trial. If patients have dropped out at random, then it is likely that there will be patients on both sequences who have dropped out, in which case the latter data will have the structure of a parallel group trial. We will thus be able to construct two estimates of the treatment effect of very varying precision and hence, from these two, a third weighted estimate. This will, in fact be done automatically by *proc mixed* if we declare patient effects to be random. However, if we declare patient effects to be fixed, then the difference between sequences, which in those patients who have received one treatment only is the difference between treatments, cannot be estimated as it is confounded with the patient effects. Consequently, it makes no difference whatsoever when employing a fixed effects model whether or not those patients who contribute data from one period are included or excluded.

Remark In the trial reported in Example 3.1 patient 8 contributed data from the first period only, as was noted in Section 3.2. The data from the first period were excluded from further consideration. As explained above, that has made no difference to the estimate of the treatment effect where patient effects have been treated as fixed. In fact, since there is only one such patient, it also makes no difference to the estimate of the treatment effect where patient effects are treated as random. For other cases it could make a difference. However, in all the many cross-over trials in which I have been involved, recovering such information has made a negligible difference to the result.

3.18.3 Recovering degrees of freedom for error from baselines

All of the analyses we have produced with SAS® so far have produced degrees of freedom which are inferior to the number of patients. In Section 3.16.2 we mentioned that given certain fairly strong assumptions baseline values could be used to generate further degrees of freedom for estimating the within-patient variance. We shall first explain how the analysis may proceed, then present the results and finally discuss the method.

We use a single baseline before treatment and no longer regard this as a concomitant value associated with the first outcome measure but as an extra ‘outcome’ in its own right. Thus, for Example 3.1, where we have 13 patients we now have $3 \times 13 = 39$ values of the variable *OUTCOME* as opposed to the 26 we had previously. Corresponding to each outcome value we record the *PATIENT* it was measured for and the *PERIOD* in which it was measured as before, only now *PERIOD* is variable with three values 0 (for the baseline period) and 1 and 2 (for the treatment periods). *PATIENT* and *PERIOD*, as before, are to be declared as *class* variables. We now have as our *TREAT*ment variable, however, an ordinary numerical variable with three values 1 (for formoterol) -1 (for salbutamol) and 0 (for baseline). The effect of this is that the baseline itself contributes nothing to the treatment estimate.

The following code is then used:

```
proc glm;
  class PATIENT PERIOD;
  model OUTCOME = PATIENT PERIOD TREAT;
  estimate ‘treatment effect’ TREAT 2;
run;
```

Of course, an equivalent analysis using *proc mixed* is possible. Note the change in the form of the *estimate* statement. Because *TREAT* is a numerical variable *estimate* provides a slope. To move from -1 for salbutamol to $+1$ for formoterol is a change of 2 units. This is why the '2' is included in the *estimate* statement. This produces an estimate of the treatment effect, 46.61, which is identical to that obtained by our calculations in Sections 3.5 and 3.6. The same value would result using the 26 genuine *OUTCOME*s only and fitting *PATIENT PERIOD* and *TREAT* as class variables. What changes is the standard error. This is now 14.59, based on 23 degrees of freedom, as opposed to 10.78, based on 11 degrees of freedom. The extra 12 degrees of freedom have arisen because we have used 13 further *OUTCOME*s at the cost of only one further parameter being fitted for *PERIOD*.

It is interesting to trace the origin of the increase in the standard error. The SAS® output for these two approaches shows that the estimated error variance for 11 degrees of freedom is 750.41, that for 23 degrees of freedom is 1375.25. The ratio of the square of the standard errors to the error variance will be found to be the same in each case: $10.78^2/750.41 = 14.59^2/1375.25 = 0.155$. This is as it should be. Both methods will always produce identical estimates, hence the variance of the two methods must be the same. This variance is, in fact, $\sigma_{\mathrm{w}}^2(1/6 + 1/7)/2 = 0.155\sigma_{\mathrm{w}}^2$, where σ_{w}^2 is the within-patient variance and is half the variance of a basic estimator as defined in Section 3.6. What has changed from one case to the other is not the estimate itself, nor the variance of the estimate, but the estimate of the variance of the estimate. We have used the baseline values in addition for this purpose. Their effect has been to increase our estimate of σ_{w}^2 considerably. In fact if we isolate these 12 degrees of freedom we find that the estimated error variance associated with them alone is 1948.02 compared to the 750.41 for the other 11.

There are many possible explanations for this difference, not least among them chance. One possible explanation is that patients respond to different degrees to beta-agonists as a class of drug. Because each patient acts as his own control, in comparing two beta-agonists this source of patient heterogeneity is eliminated. Comparing to baseline it is not eliminated.

Of the two approaches, I prefer the one which does not use the baselines and estimates the error variance using 11 degrees of freedom rather than 23. This is not because it leads to a lower estimated standard error. I should still prefer it if the reverse were the case. I prefer it because it is based on the direct comparison of the two treatments. Further discussion of this general issue is given in Chapter 5.

3.18.4 Basic estimator analysis using *proc glm*

The first step for this sort of analysis is to calculate the basic estimator for each patient. The reduced data on outcomes then consist of n values, one for each patient, representing, for this example, the difference between outcomes when treated with formoterol and salbutamol. Any additional information we have on patients must also be reduced to a single value for each patient. So if we use baseline differences we need to calculate first the cross-over difference for baselines. For patient no. 4 a data line might look like this:

PATIENT	BASICEST	SEX	GROUP	BASE1	BASEDIF
4	50	M	FS	300	30

Here, *BASICEST* is the basic estimator for the patient, the PEF value at 8 hours under formoterol minus that under salbutamol, the patient's *SEX* is male, his sequence group is formoterol/salbutamol, *BASE1* is the baseline PEF value observed on treatment day 1 and *BASEDIF* is the cross-over difference for baseline PEF (for which the value before salbutamol has been subtracted from that prior to formoterol).

If the following code is then used:

```
proc glm;
  class GROUP;
  model BASICEST = GROUP;
  estimate "for-sal" intercept 1;
run;
```

an analysis corresponding to that is Section 3.6 will be produced. It should be noted that because the treatment estimate is produced by averaging and not by differencing, the necessary contrast is obtained from the intercept of the model. The *estimate* statement has been adjusted accordingly. The *GROUP* term was fitted to the model to account for a period effect since the difference between basic estimators from different sequence groups reflects the period effect.

The investigation in Section 3.4 of the interaction between sex and treatment could have been performed by using

```
proc glm;
  class SEX;
  model BASICEST = SEX;
  estimate "for-sal" INTERCEPT 1;
  estimate "sex treat" SEX −1 1;
run;
```

Among the lines of output will be found the following:

Parameter	Estimate	T for H_0: Parameter = 0	Pr > $\|T\|$	Std Error of Estimate
for-sal	47.013888 89	3.71	0.0034	12.674 6878
sex treat	8.472 222 22	0.33	0.7445	25.349 3755

Obviously the period effect could also have been fitted in the model here. It has been omitted to make the results comparable to those of Section 3.4. The results for the sex by treatment interaction agree with those obtained before. The estimate of the treatment effect requires some explanation and is not uncontroversial. In the model we have effectively fitted a sex by treatment interaction. Sex itself is not confounded with treatment in the same way that period was. In period 1 seven patients received formoterol whereas six received salbutamol. Thus period and treatment are (very slightly) confounded. Consequently if we were worried that a period effect might influence our results it would make sense to adjust the treatment effect for the period effect. On the other hand, sex is not at all confounded with treatment. Any general tendency, for example, for males to have higher PEF readings than females would not disturb the validity of a treatment estimate calculated as a straightforward average of the basic estimators. Such sex effects are already removed by calculating the basic estimators. It is only the sex by treatment interaction which is confounded with treatment effects, if, for example, formoterol works better for females than for males.

SAS® has calculated the mean treatment effect within each sex stratum and then taken the mean of these two values. Thus if we look at Table 3.3 we shall see that the mean estimate for females was 51.25 and for males was 42.78. When averaged these produce the value of 47.01 given by SAS®. This may be regarded as the average of the male treatment effect and the female treatment effect. Its standard error is rather large, reflecting the imbalance between the sexes, the error variance having been multiplied by $(1/9 + 1/4)/4 = 0.0903$ rather than $1/13 = 0.0769$. (In fact it can easily be verified that the standard error is simply half that of the sex by treatment interaction, having been obtained as the average of the two values for which this is the difference.)

In fact there is a choice of a number of possible estimates we might wish to report here: the average of all the basic estimators, the average of the average for each sex or the two averages for each sex. This is a general problem which affects all models in which interactions are considered. Various authorities have extremely strong opinions on the subject as to which it is that ought to be reported and some make the choice dependent on whether or not an interaction is detected by a significance test. I can give little help here apart from advising that if interactions are fitted it should clearly be stated what has been done. My own preference is not to fit interactive effects at all in cross-over trials when reporting treatment effects, and only to investigate interactions (if at all) at a later stage with the possible objective of generating new hypotheses to be the subject of planned investigation in further trials. This is consistent with the general advice of the International Conference on Harmonisation guideline on statistical analysis (International Conference on Harmonisation, 1999). Further discussion of the interpretation of main effects in the presence of interactions will be found in Senn (2000c).

Fitting first period baselines in a model is also an example of fitting an interactive effect. Taking the advice of the previous paragraph we assume that it is only the possibility of a baseline by treatment interaction which is of interest here. The thorny question of estimating the treatment effect for this model will not be addressed. Assuming we also fit a period effect, the code here is

```
proc glm;
  class GROUP;
  model BASICEST = GROUP BASE1;
run;
```

The output from SAS® permits the construction of the following ANOVA table.

Source	*DF*	Sum of squares	Mean square	*F*	pr > *F*
BASE1	1	401.09	401.09	0.25	0.6286
Error	10	16 107.84	1610.78		

Clearly there is no evidence of significant treatment by baseline interaction.

3.19 FURTHER READING

There is an extensive literature on the *AB*/*BA* design and before concluding with our recommendations we give some suggestions for further reading.

The papers by Hills and Armitage (1979), Armitage and Hills (1982) and Clayton and Hills (1987) give excellent introductions to the *AB*/*BA* design with good discussions of many of the problems and excellent advice regarding analysis (except as regards carry-over). The first, in particular, is a classic and highly recommended. A more mathematical introduction to the field is provided by Grizzle (1965, corrigenda: Grizzle, 1974; Grieve, 1982). Senn (1994) reviews the *AB*/*BA* design in detail as do the encyclopaedia articles by Kenward and Jones (1998) and Senn (1998a, 2000b). The paper by Koch (1992) is an excellent review of the use of baselines. Other useful references are Brown (1980), Dubey (1986) and the books by Jones and Kenward (1989), Fleiss (1986a, pp. 263–80), Matthews (1990a) and Lehmacher (1987, pp. 77–148). The latter two suffer from the disadvantage of being relatively difficult to access in that the first is a private publication and the second is in German.

For approaches to analysis not covered in this chapter alternative suggestions will be found in the following papers. Zimmerman and Rahlfs (1980) cover the modelling of the *AB/BA* cross-over in a multivariate framework. Grieve (1985) provides a Bayesian analysis. Willan and Pater (1986a) and Willan (1988) develop approaches based on assumptions regarding the relationship between carry-over and the direct treatments effect. Patel (1983), Willan and Pater (1986b) and Kenward and Jones (1987a) consider the use of baseline data. The book by Ratkowsky *et al.* (1993) also provides an alternative approach to analysing *AB/BA* designs with three baselines. However, the recommendation for the estimation of the treatment effect boils down to comparing the differences between the outcomes and baselines for the first period (the first period 'change scores'), as in any parallel group trial, and makes the use of a cross-over trial, irrelevant. It thus cannot be recommended. Lehmacher (1991) proposes a closed testing procedure based on a multivariate model of outcomes. Patel (1986) considers the problem of incomplete data. Wallenstein and Fisher (1977) consider the analysis of designs with repeated measures within periods.

Much of the advice regarding carry-over in many of the papers above, however, has been outdated by Freeman's (1989) detailed investigation of the two-stage procedure. The reading of this is mandatory for anyone who wishes to make a serious study of the problem with carry-over. An introduction to its contents will be found in Senn (1991) and it is also covered in detail in the encyclopaedia articles already cited (Kenward and Jones, 1998; Senn, 1998a, 2000b). A heuristic account, suitable for non-statisticians, is given in Senn (1995a). The paper is also reviewed in Matthews (1990a). Jones and Kenward (1989) also give an interesting discussion but in my opinion underrate its practical significance.

3.20 RECOMMENDATIONS

In conclusion the following recommendations are given regarding design and analysis.

The first concerns *carry-over*. No help regarding this problem is to be expected from the data. The solution lies entirely in design. The trialist must only use cross-over trials in appropriate indications and he must allow for adequate wash-out. If a wash-out period is not possible he should use an active washout as described in Chapter 1 and limit measurements to the latter part of the observation period. Alternatively he should consider running a parallel group trial.

A cautious approach will be to consider that *period effects* may be important and to allocate patients in equal numbers to each sequence. (In doing this I prefer to impose no further restrictions on allocation and use a block size equal to the trial size.) If this has been done, an analysis which fits a period effect should be performed. It may be considered desirable to adjust for period effects even though patients have been allocated completely at random to sequences.

If the investigator is unconcerned about period effects he may allocate the patients completely at random and perform a matched pairs analysis. By doing this he gains one degree of freedom for error. This is unlikely to be an important consideration except for very small trials. In general, therefore, I recommend adjusting for the period effect except where the following three criteria are satisfied. First, the investigator is unconcerned about period effects. Second, he has demonstrated this lack of concern by allocating patients completely at random. Third, there are few patients (say fewer than 12 patients).

If the *wash-out* period is long compared to the treatment period (at least equal to it), there may be considerable value in using *baseline* values. This should be done using analysis of covariance. Under other circumstances baseline values should not be used.

Unless the investigation of *interactive* effects has been specified as a major goal of the trial, I do not recommend fitting them when reporting treatment effects.

Finally, all of the above discussion regarding estimation has been on the assumption that we are dealing with outcome measures which are approximately Normally distributed. In the next chapter other sorts of outcomes will be considered.

APPENDIX 3.1 ANALYSIS WITH GENSTAT®

This section gives code suitable for carrying out with GenStat® most of the analyses outlined in this chapter. The code is free-standing, which is to say that if the code below were used exactly as given and a GenStat® session run, it ought to produce the analyses of this chapter. For that reason the data have been input as part of the program. In practice, reference might be made to an external file. However, this book is not about GenStat® and it is assumed that the reader who is interested will already be familiar with that program. Further advice regarding GenStat® will be found in Harding *et al.* (2000) and McConway *et al.* (1999). Therefore, apart from the comment statements enclosed in double quotes, only the briefest explanation is offered in the remarks that follow.

In analysing linear models, computer packages have to deal with the fact that one fewer parameter value is needed than the number of levels of a factor, otherwise the model will be over-parameterized. Different packages use different constraints for dealing with this. For example, the first parameter value may be arbitrarily set to zero, or the last may be set to zero or the constraint may be imposed that they sum to zero. GenStat® and SAS® differ in this respect: the former sets the first level to zero, the latter sets the average to zero. Advantages can be claimed for each of these approaches. No position is taken here on which is better. Usually, provided contrasts between the levels of a factor (say, differences between treatments) are calculated, this makes no difference. However, the intercept term in models will be different since this is not a contrast. The following adaptations to analysis have been made in applying GenStat®.

1. In coding the treatment factor, *Treat*, formoterol, which has label *for*, has been set to be the second level and salbutamol, which has label *sal*, has been set to the first level in order to ensure that the estimate of the parameter associated with this factor corresponds to the contrast formoterol – salbutamol rather than vice versa (Fit1–Fit4 below).
2. The basic estimator approach uses the intercept. To produce this approach in GenStat® it is necessary to code the sequence as a variate with the values -1 and 1 (Fit6 below). This ensures that the intercept returns the required value. In the program this coding is represented by *Seq2*.
3. Since the same coding has been used for the analysis of period differences (Fit5 below), *PEFdiff*, and since the difference between 1 and -1 is 2, it has not been necessary to use the semi-period differences as in the analysis using Excel®. *Seq2* is treated as a continuous predictor so that GenStat® returns the effect associated with a change of 1 rather than 2, and this corresponds to the analysis of semi-period differences.

```
"Example 3.1 from Cross-over Trials in Clinical Research"
"8 hour PEF data of Graff-Lonevig and Browaldh

"Input data values"
FACTOR[NVALUES=26;LEVELS=!(1...7,9...14)] Patient
"Patient 8 missing"
READ Patient; frepresentation=ordinal
1 4 6 7 9 10 13 2 3 5 8 11 12 1 4 6 7 9 10 13 2 3 5 8 11 12:
FACTOR[VALUES=13(1,2);LEVELS=2] Period
FACTOR[VALUES=7(1),13(2),6(1);LEVELS = 2;\
LABELS=!t('for','sal')] Treat
"Derive factor representing the sequence"
FACTOR[LABELS=!t(forsal,salfor)] Seq
CALC Seq = 1+ (Treat/=Period)
VARIATE[NVALUES=26] Base, PEF
READ Base
290 300 250 390 250 365 190 350 350 350 280 270 220 270 270 210
390 240 380 260 345 370 360 290 310 220:
READ PEF
310 310 370 410 250 380 330 370 310 380 290 260 90 270 260 300 390
210 350 365 385 400 410 320 340 220 :
```

```
"Demonstrate various regression approaches using patient as"
"a fixed effect"
MODEL PEF
"Fit1: Just fitting patient and treat"
FIT[PRINT=model,estimates; TPROB=yes] Patient+Treat
ADD[TPROB=yes] Period "Fit2: Including period effect"
ADD[TPROB=yes] Base "Fit3: Including baselines"

"Fit4: As Fit2 but treating patient effect as random"
VCOMPONENTS[FIXED=Seq+Period+Treat] RANDOM=Patient
REML[PRINT=effects; PSE=differences] PEF

"Begin calculation of period differences, basic estimators"
"& re-coding of Seq to Seq2"
CALC n = NLEVEL(Patient)
"To be used to set length of data vectors"
"Create two columns (for two periods) of PEF values"
"and treatment indicators"
VARIATE[NVALUES=n] PEFPat[1,2]
FACTOR[NVALUES=n; LEVELS=2] TreatPat[1,2]
EQUATE OLD=PEF,Treat; NEW=PEFPat,TreatPat
CALC PEFdiff = PEFPat[1] -PEFPat[2] "Period difference"
& Seq2 = TreatPat[1] -TreatPat[2]
"Recodes sequence as −1 and 1"
& Basicest = PEFdiff* Seq2
"Basic estimator: Seq2 reverses sign of PEFdiff"

"Fit5: period differences"
MODEL PEFdiff
FIT[PRINT=model,estimates; TPROB=yes] Seq2

"Fit6: analysis using basic estimators"
"Treatment effect is associated with intercept"
MODEL Basicest
FIT[PRINT=model,estimates; TPROB=yes] Seq2
```

APPENDIX 3.2 ANALYSIS WITH S-Plus®

This section gives code suitable for using S-Plus® for most of the analyses outlined in this chapter. The code is free-standing, which is to say that if the code below were used exactly as given and an S-Plus® session run, it ought to produce the analyses of this chapter. For that reason the data have been input as part of the program. In practice, reference might be made to an external file rather than including the data as part of the program. However, this book is not about S-Plus® and it is assumed that the reader who is interested will already be familiar with that program. An excellent introduction is given in the book by Krause and Olson (2000), a more advanced treatment is given in Venables and Ripley (1999), modelling in general is covered in Harrell (2001) and random effect models are covered in great detail in Pinheiro and Bates (2000). Therefore, apart from the comment statements preceded by the hash sign, #, in the program itself, only the briefest explanation is offered in the remarks that follow.

To obtain the analyses given in this chapter, a contrast for the treatment effect needs to be requested. As explained in the appendix on GenStat® analyses above, different packages use different conventions for dealing with coding and parameterization of linear models. S-Plus® also offers a number of different options for defining contrasts. The default contrasts are Helmert contrasts, which for two levels of a factor will return treatment estimates with half the value we would like. In general, for S-Plus®, I recommend setting the default treatment contrasts as follows:

```
options(contrasts=c(factor="contr.treatment",
ordered="contr.poly")).
```

For factors that are unordered, contrasts will then compare all treatments to the first, and for ordered factors, orthogonal polynomial contrasts will be used. For the analysis of original data given in this chapter, this setting produces the desired result. However, for the analysis of period differences, or that of basic estimators, the following setting should be used:

```
options(contrasts=c(factor="contr.sum",
ordered="contr.poly"))
```

In order to make sure in all these analyses that the treatment effect has a sign that is consistent with the difference formoterol−salbutamol, it has been necessary to make sure that formoterol is coded as the second level of the treatment factor and that the sequence formoterol/salbutamol is coded as the second level of the sequence factor.

```
#Example 3.1 from Cross-over Trials in Clinical Research
#8 hour PEF data of Graff-Lonevig and Browaldh

#Input data
n1< -7 #number of patients first sequence
n2< -6 #number of patients second sequence
n<-n1+n2
```

```
seqn<-factor((c(rep(1,n1),rep(2,n2),rep(1,n1),
rep(2,n2))), labels=c("forsal","salfor")) #sequences
patient<-factor(rep(c("1","4","6","7","0","11",
"14","2","3","5","9","12","13"),2))
period<-factor(c(rep("1",n),rep("2",n)))
treat<-factor((c(rep(2,n1),rep(1,n2),rep(1,n1),
rep(2,n2))), labels=c("salbutamol","formoterol"))
#Note: "formoterol" is coded second level of factor
pef<-c(310,310,370,410,250,380,330,370,310,380,290,260,
90,270,260,300,390,210,350,365,385,400,410,320,340,220)
base<-c(290,300,250,390,250,365,190,350,350,350,280,270,
220,270,270,210,390,240,380,260,345,370,360,290,310,220)

#Define contrasts
options(contrasts=c(factor="contr.treatment",
ordered="contr.poly"))

#Demonstrate various approaches using patient as fixed effect
#Analysis just fitting patient and treat
fit1<-lm(pef~patient+treat)
summary(fit1, corr=F) #The corr=F option suppresses the
#printing out of the correlation matrix
#Fitting period
fit2<-lm(pef~patient+period+treat)
summary(fit2,corr=F)
#Fitting baselines as well
fit3<-lm(pef~patient+base+period+treat)
summary(fit3,corr=F)

#Now illustrate use of random effects model
fit4<-lme(pef~seqn+period+treat, random=~ 1| patient)
summary(fit4)

#Calculate period difference, basic estimators,
#re-code seqn to seqn2
#Initialise variables
pefdiff<-numeric(n)
basicest<-numeric(n)
seqn2<-c("forsal","salfor")
#Note use of "ifelse" statement to check which sequence
# patient is in
# forsal sequence coded "B";, salfor "A"
period1 <- 1:n
```

```
period2 <- (n+1):(2*n)
pefdiff <- pef[period2] - pef[period1]
basicest <- ifelse(seqn[period1]=="forsal", pef[period1]
- pef[period2], pef[period2] - pef[period1])
seqn2 <- ifelse(seqn[ period1]=="forsal","B","A")

#Reset definition of contrasts
options(contrasts=c(factor="contr.sum",
ordered="contr.poly"))

# Illustrate analysis using period differences
#(equivalent to fit2 and fit4)
# The regression coefficient for seqn2 will give the treatment
# effect
fit5<-lm(pefdiff~seqn2)
summary(fit5, corr=F)

#Illustrate analysis using basic estimators
#(equivalent to fit2, fit4 and fit5)
# The regression coefficient for intercept gives treatment
# effect
fit6<-lm(basicest~seqn2)
summary(fit6,corr=F)
```

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

4

Other Outcomes and the AB/BA Design

4.1 INTRODUCTION

In the previous chapter the *AB/BA* design with Normal outcomes was covered in great detail. The time spent on the topic may be justified in the following terms. First, because many important outcomes measured in clinical trials are continuous. Even though few of these outcomes are Normally distributed, methods based on the Normal distribution are fairly robust and may appropriately be applied in many cases. Second, because the range of techniques and possibilities for analysis is greatest where methods based on Normal outcomes are employed. Third, because interpretation is easiest where such methods are used. Fourth, because some other types of outcome, for example survival, which are extremely important in clinical trials in general, have limited application to cross-over trials. Fifth, because considerable gains in precision are often possible by employing cross-over designs where Normal outcomes are concerned and, finally, because using a cross-over design increases the probability that the relevant error terms under the model will be Normally distributed since within patient errors may be so distributed where between patient errors are not. Nevertheless, even in the case of cross-over trials, suitable outcomes for measurement are not limited to those which are approximately Normally distributed. In this chapter appropriate analyses for the 'other' cases will be outlined.

Techniques for dealing with non-Normal outcomes will be considered under six major headings: transformations, non-parametric methods, methods for binary outcomes, methods for ordered categorical data, analysis of frequency data and survival analysis. (The latter is perhaps a rather surprising topic in a book on cross-over trials since 'you only live once', but there are some limited applications involving titrated challenges in which the concept of survival may be applied to cross-over trials.)

4.2 TRANSFORMATIONS

The analyses undertaken in Chapter 3 for Example 3.1 all had one thing in common: the effects were estimated in terms of the original units of PEF. In the medical literature, however, it is probably more usual to see results expressed in terms of PEF as a percentage of baseline (Senn, 1989a), where the baseline value is the value obtained just prior to dosing on the given day. Leaving aside for the moment the issue as to whether this is a wise use of the baseline values, the fact that a percentage measure is used is revealing. Such a measure is appropriate if it is suspected that the treatment effect will be multiplicative—if the improvement which a given treatment will provide is proportionate to the level which would apply without treatment. Under such circumstances a logarithmic transformation is useful since multiplicative effects in terms of the original data become additive ones in terms of logs. Standard statistical models then apply to the transformed data.

4.2.1 Logarithmic transformations

We now apply such an analysis to the PEF data of Example 3.1, transforming them first to logs, as presented in Table 4.1, and then using the method of Section 3.6 (i.e. correcting for any period effect). It makes no difference to any of the inferential statistics, such as t statistics, P values etc., to which base we take logs since, if a and b are two bases and Y is an outcome, these satisfy the relationship $\log_b Y = \log_b a \ \log_a Y$ (Diem and Seldrup, 1982, pp. 170–1) so that, $\log_b a$ being a constant, one log is a simple multiple of another. Thus logs to base e may be obtained by multiplying logs to base 10 by $\log_e 10 \simeq 2.3026$ and logs to base 10 by multiplying logs to base e by $\log_{10} e \simeq 0.4343$. All scientific pocket calculators provide natural logarithms so that the practical computational advantage of logs to base 10 is no longer relevant. Logs to base e also have the following advantage in terms of interpretation of results from clinical trials. The antilog to base e of $\log_e Y$, which is by definition Y, satisfies the relationship:

$$Y = 1 + \log_e Y + (\log_e Y)^2/2! + (\log_e Y)^3/3! + \ldots.$$

For values of $\log_e Y$ close to zero (to which our estimates of treatment effects often reduce) the first two terms on the right-hand side alone often provide a reasonable approximation so that $Y \simeq 1 + \log_e Y$. For example if we note that our treatment estimate in logs to base e is 0.1, then we can note straight away that its antilog is approximately 1.1 and hence the average effect of treatment appears to be to increase values by 10%. Of course, in practice we should always antilog such results properly and in this case the antilog to four figures of decimals is in fact 1.1052. We shall only use logarithms to base e therefore and will simply use log to designate them.

Using the standard CROS analysis on the values in the table we find that the estimated treatment effect is 0.188 log (ℓ/min) and the standard error is 0.06372 log (ℓ/min), yielding a t ratio of 2.95 on 11 degrees of freedom. (The details of the calculation have been omitted since, once logs are taken, they are identical to those in Section 3.6.) The critical value for a two-sided test at the 5% level we already know to be 2.201, so the result is clearly significant.

Alternatively from Table 9 of Lindley and Scott (1984) we obtain a *P* value of 0.013. Use of the Excel function TDIST (2.95,11,2) yields the same result. The confidence limits may be obtained by forming the product of the critical value and the standard error and substracting this from or adding it to the estimate. Here, the lower limit is 0.048 log (ℓ/min) and the upper limit is 0.328 log (ℓ/min).

The estimate of the treatment effect is not usefully left in terms of logs but should be antilogged to be expressed as a ratio. Doing this we find that the estimated treatment effect is $e^{0.188} = 1.21$. In other words using the log transformation our treatment estimate suggests that formoterol increased PEF by 21% compared to salbutamol. We could have achieved this estimate directly by calculating a basic estimator for each patient in terms of ratios of PEF. This has in fact been done in the last column of Table 4.1. For example, for patient 1 the basic estimator is $310/270 = 1.148$. Note that the log of this ratio is equivalent to the cross-over difference in logs. We now need to obtain the mean of these basic estimators within each group. The fact that we have taken ratios, however, implies that we need to take a geometric mean (which, for *n* numbers, is the *n*th root of their product) rather than the more usual arithmetic mean. If the geometric means of the basic estimators are calculated within each sequence group and the geometric means of the two means is then obtained, the result 1.21 is reached just as it was using logs.

Table 4.1 Log transformation of data on PEF from the *AB*/*BA* cross-over considered in Example 3.1.

		log PEF			PEF
Sequence	Patient number	Formoterol	Salbutamol	Cross-over difference	Cross-over ratio
for/sal	1	5.7366	5.5984	0.138	1.1481
	4	5.7366	5.5607	0.176	1.1923
	6	5.9135	5.7038	0.210	1.2333
	7	6.0162	5.9661	0.050	1.0513
	10	5.5215	5.3471	0.174	1.1905
	11	5.9402	5.8579	0.082	1.0857
	14	5.7991	5.8999	−0.100	0.9041
sal/for	2	5.9532	5.9135	0.040	1.0405
	3	5.9915	5.7366	0.255	1.2903
	5	6.0162	5.9402	0.076	1.0789
	9	5.7683	5.6699	0.098	1.1034
	12	5.8289	5.5607	0.268	1.3077
	13	5.3936	4.4998	0.894	2.4444

for/sal = formoterol followed by salbutamol
sal/for = salbutamol followed by formoterol

The confidence intervals are also better expressed in terms of ratios and these should be antilogged in the same way as the treatment estimate. This yields the limits 1.05 and 1.39.

If the use of baseline information is contemplated, exactly the same considerations apply as did in Chapter 3. Any of the analyses outlined in that chapter may be employed. All that it is necessary to do is to take logs of all the measurements, including the baselines, and then perform the calculations in the ways illustrated. At the end treatment estimates and confidence limits may be antilogged to express them as ratios.

In this particular example, the log transformation does not appear to be particularly successful. Patient 13 was something of an odd man out as regards response measured in terms of ℓ/min of PEF but this eccentricity has been accentuated by the log transformation. His results thus have a disproportionate effect on the treatment estimate and an even less desirable effect on the estimate of its standard error. This example illustrates effectively the properties of the t test with respect to robustness. Although the results from the atypical patient increase the treatment estimate they actually depress the t statistic. Thus if an analysis is carried out on log PEF excluding patient 13, the treatment estimate is 0.1258 log, considerably less than the value of 0.188 log achieved before. The standard error of 0.031 13 log is proportionately lower, however, than the previous values of 0.063 72 log, and this despite the fact that it is based on fewer patients (12 as opposed to 13) who show a greater imbalance between sequences (a 7–5 as opposed to 7–6 split). As a result the t statistic increases from 2.95 (on 11 degrees of freedom) to 4.04 (on 10 degrees of freedom) and this, despite the loss of a degree of freedom, yields a more impressive P value (two-sided) of 0.0024. Alternatively, looking at the matter in terms of ratios of PEF, the analysis excluding patient 13 yields a treatment estimate of 1.134, which is considerably lower than the value of 1.207 reached before. On the other hand the lower confidence interval is actually greater at 1.06 than the previous figure of 1.05. The general lesson for using the t test is this: rogue values are unlikely to cause an investigator to conclude that a useless treatment is effective; the danger is rather that he may wrongly conclude that an effective treatment is useless.

4.2.2 The logit transformation

Other transformations apart from logs may be indicated on the basis of *a priori* considerations. For example a commonly employed measurement technique in clinical trials is to ask patients to assess outcomes with the help of a visual analogue scale (VAS) (Aitken, 1969). The patient may be presented with a 10 cm line for which one extreme represents ‘good’ and the other ‘bad’ and asked to indicate the point on the scale to which his assessment of the efficacy of the treatment corresponds. The distance in mm from one end of the scale then gives a number between 0 and 100 which may be used as a rating of efficacy.

The constraint these limits impose, however, can cause difficulties for analysis and interpretation. Suppose that on the basis of performing an analysis as outlined in Section 3.6 we conclude that the VAS score is reduced on average by 20 mm for treatment A compared to treatment B. How would we interpret its effect on a patient who scores 15 mm for B? Considerations such as this may lead one to consider a transformation which produces a score which lies between $-\infty$ and $+\infty$. If the extreme values 0 and 100 may be excluded, such a score may be achieved by performing the transformation:

$$\log\{(\mathrm{VAS})/(100 - \mathrm{VAS})\}. \tag{4.1}$$

(Note that a very similar transformation, as a theoretical construct, is extremely important in general in probability and statistics. If P is the probability of an event then $P/(1 - P)$ is the *odds*, and $\log\{P/(1 - P)\}$, the *log odds*, has the property that it lies between $-\infty$ and $+\infty$. Such a transformation is referred to as a *logit* (Diem and Seldrup, 1982, p. 70). By analogy and for convenience values transformed by (4.1) will be referred to as *logit scores*.)

Example 4.1 The data in Table 4.2 report patients' opinion of efficacy for a trial in asthma. The treatments compared in the trial were a single dose of formoterol solution aerosol 12 μg and a single dose of salbutamol aerosol 200 μg. Patients were allocated at random in equal numbers to one of two treatment sequences:

Table 4.2 (Example 4.1) Patients' judgement of efficacy: visual analogue scale scores in millimetre (0 good, 100 bad) for a two-period two-treatment cross-over in asthma.

		VAS score		
Sequence	Patient number	Formoterol	Salbutamol	Basic estimator
for/sal	1	63	82	−19
	3	24	84	−60
	5	32	66	−34
	6	33	51	−18
	10	24	75	−51
	12	30	38	−8
sal/for	2	40	68	−28
	4	5	4	1
	7	21	57	−36
	8	5	23	−18
	9	32	56	−24
	11	4	53	−49

for/sal = formoterol followed by salbutamol
sal/for = salbutamol followed by formoterol

formoterol followed after a wash-out of seven days by salbutamol or salbutamol followed after a wash-out of seven days by formoterol. The main purpose to the trial was to study the protective effect of these bronchodilators on methacholine-induced bronchoconstriction, and during each treatment day the patients were given increasing doses of methacholine until a given response was observed. (This point will be discussed in due course below.) Patients were asked to give their opinion on efficacy using a VAS scale with 'bad' on the right and 'good' on the left. Thus high scores are bad and low scores are good.

If the standard CROS analysis of Section 3.6 is used on the original VAS measurements then the estimated treatment effect is −28.67 mm with a standard error of 5.415 mm yielding a highly significant t statistic of − 5.29 and a confidence interval for the effect of − 41 mm to − 16 mm.

Alternatively, if we work in logit scores, we need to use the values given in columns 3, 4 and 5 of Table 4.3. Using the transformed values we get an estimated treatment effect of − 1.4276 with a standard error of 0.3033 and a t statistic of − 4.71. Again the result is highly significant. The confidence interval for the effect is − 2.10 to − 0.75.

A VAS score is at the best of times not easy to interpret. When transformed in this way the task appears even more daunting. This is not necessarily a reason for preferring the original scale of measurement. It is the difficulty of the problem that imposes the penalty of interpretation. One way of trying to understand such a transformation is to investigate its implications. Associated with the transformation is an implicit model. This model assumes that treatment effects

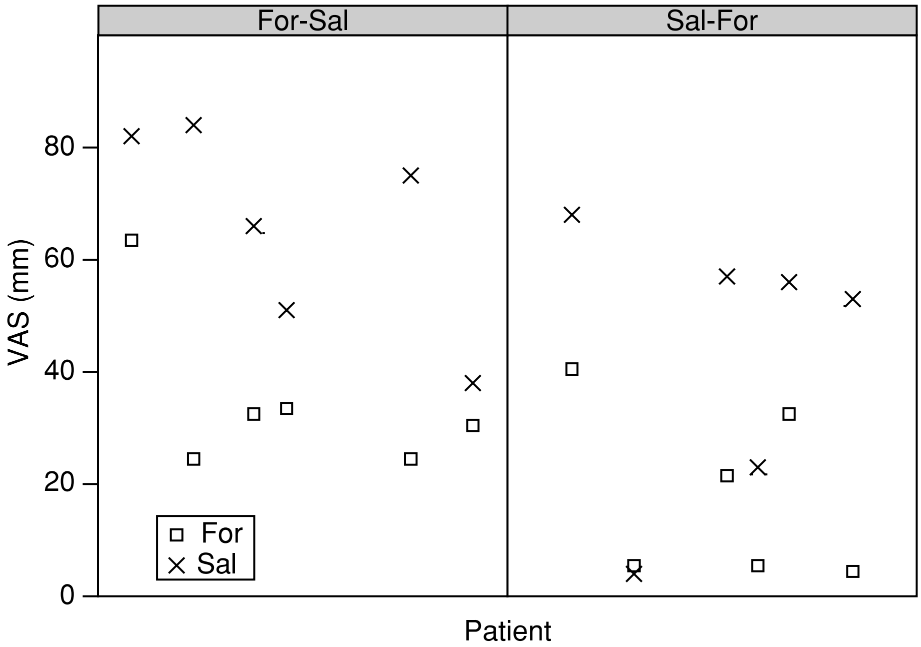


Figure 4.1 (Example 4.1) Visual analogue score (VAS) for patients' opinion of efficacy for an *AB*/*BA* cross-over (0 = 'good', 100 = 'bad').

are additive providing the VAS is expressed as a logit score. Thus the model states that if we had a patient whose VAS score given treatment with salbutamol was 60 mm, which corresponds to a logit score of 0.4055, our best estimate of what his logit score would have been under formoterol is $0.4055 + -1.4276 = -1.0221$. This itself can be transformed using the *inverse logit score transformation*:

$$100\{\mathrm{e}^{\mathrm{logit}}/(1 + \mathrm{e}^{\mathrm{logit}})\}. \tag{4.2}$$

In this case, applying the inverse transformation (4.2) yields an expected VAS score under formoterol for such a patient of 26 mm. If necessary in order to familiarize oneself with the implications of the model a table of such possible results can be constructed. Alternatively, since the logit score for 50 mm is 0, (4.2) may be applied directly to the estimate and confidence intervals and 50 mm may be subtracted from the result. In this example we then obtain − 31 mm for the estimate with confidence limits of − 39 mm and − 18 mm, for the expected treatment effect, formoterol–salbutamol, of a patient whose true response under salbutamol is 50 mm.

An alternative transformation which is sometimes considered for VAS scales is the *arc sine transformation* (Fleiss, 1986a, p. 62):

$$\sin^{-1}(\mathrm{VAS}/100)^{1/2}. \tag{4.3}$$

The values of this transformation are also given in Table 4.3. The reader may check for himself that if the standard CROS analysis of Section 3.6 is applied to these data a t statistic of 5.14 results.

One last point may be made with this example to remind the reader that the interpretation of a clinical trial is an issue that is much wider than the relatively simple one of transformation of data. This trial was a methacholine challenge. The patients were first treated with the drugs and then given doubling doses of methacholine until a lung function measurement, forced expiratory volume in one second (FEV_1), showed a 20% drop. This raises huge difficulties in interpretation. First the 'baseline' measurements with respect to which the 20% drop was determined were taken after treatment—see Senn (1989a, 1993c) for a critical discussion of the difficulties to which this leads. Second, this being a titration, the meaning of all other measurements is disturbed. The amount of methacholine received becomes the variable outcome of interest but the treatments are thereby confounded with the doses of methacholine. How can we be sure that when the patients recorded their impression of the efficacy of the treatment they were capable of disentangling the effect of treatment and challenge?

Table 4.3 (Example 4.1) Patients' judgement of efficacy: visual analogue scale logit scores ($-\infty$ good, ∞ bad) and arc sine scores for a two-period two-treatment cross-over in asthma.

		VAS logit score			VAS arc sine score		
Sequence	Patient number	Formoterol	Salbutamol	Basic estimator	Formoterol	Salbutamol	Basic estimator
for/sal	1	0.5322	1.5163	−0.984	0.91691	1.13265	−0.2157
	3	−1.1527	1.6582	−2.811	0.51197	1.15928	−0.6473
	5	−0.7538	0.6633	−1.417	0.60126	0.94826	−0.3470
	6	−0.7082	0.0400	−0.748	0.61194	0.79540	−0.1835
	10	−1.1527	1.0986	−2.251	0.51197	1.04720	−0.5352
	12	−0.8473	−0.4895	−0.358	0.57964	0.66422	−0.0846
sal/for	2	−0.4055	0.7538	−1.159	0.68472	0.96953	−0.2848
	4	−2.9444	−3.1781	0.234	0.22551	0.20136	0.0242
	7	−1.3249	0.2819	−1.607	0.47603	0.85563	−0.3796
	8	−2.9444	−1.2083	−1.736	0.22551	0.50018	−0.2747
	9	−0.7538	0.2412	−0.995	0.60126	0.84554	−0.2443
	11	−3.1781	0.1201	−3.298	0.20136	0.81542	−0.6141

for/sal = formoterol followed by salbutamol
sal/for = salbutamol followed by formoterol

4.2.3 Other transformations

The use of three common transformations has been illustrated but these are by no means the only possibilities. We now consider briefly some alternatives.

For certain types of measure *power transformations*, that is to say raising the measurements to whole or fractional positive or negative powers, may be appropriate. In asthma for example the reciprocal transformation of measurements (which is the power transformation with exponent − 1) is so common that some of the results have designated names and are regarded as measurements in their own right. Thus the reciprocal of specific airways resistance, *sRaw*, is known as specific conductance, *sGaw*. One sometimes even sees separate analyses of the two in one trial and the presentation of the results as if different things were being measured, rather than exactly the same thing on a different scale!

It is not uncommon to encounter frequency data in trials. For example we may record the number of occasions on which a patient suffered severe headache in a trial of migraine or the number of times in which rescue medication was used in a trial of rheumatism, or, in a trial of asthma, the number of asthma attacks suffered. For such data, it is reasonable to expect, once all systematic sources of variation have been eliminated, that the data will then have at least the variability of a *Poisson distribution* for which the variance is equal to its expected value (or mean). Thus if we knew what the expected responses were for a given patient in a cross-over trial we might then suppose that his basic estimator would have a variance greater than or equal to the sum of the expected values. Hence, if Y_{iA} and Y_{iB} are the observed responses for patient i when treated with A and B respectively and μ_{iA} and μ_{iB} are the two expected responses which reflect his personal level of health, the effect of treatment and the effect of the particular period in which the result was observed, then

$$\mathrm{var}(Y_{iA} - Y_{iB}) \geqslant \mu_{iA} + \mu_{iB}. \tag{4.4}$$

This means that in such a trial we should be rather less impressed by given differences observed for patients for whom a large number of events on average were recorded. For example to reduce the number of attacks of migraine from 3 to 1 may be considered to be more indicative of success than reducing them from 9 to 7.

Under such circumstances a possible transformation is to divide the cross-over differences by the square root of the sum of the events in the two periods to get the expression in (4.5) below:

$$(Y_{iA} - Y_{iB})\surd(Y_{iA} + Y_{iB}). \tag{4.5}$$

This transforms a difference of 2 constructed from 3 − 1 to 1 and a difference of 2 constructed from 9 − 7 to 0.5. Thus as regards influence on our estimate of the treatment effect we consider two patients with a reduction from 9 to 7 as having the same importance as one with a reduction from 3 to 1 or one with a reduction from 10 to 6. Obviously (4.5) cannot be used for patients for whom no event is recorded in either period, but such patients may be regarded as uninformative anyway and removed from analysis. Alternatively we may adopt a solution with a slightly *Bayesian* flavour and imagine that each patient starts the trial with a score of one for A and one for B to which his trial experience adds. We thus modify (4.5) to become:

$$(Y_{iA} - Y_{iB})/\sqrt{(Y_{iA} + Y_{iB} + 2)}. \tag{4.6}$$

An alternative transformation which is almost equivalent is to take the square root of one plus the individual frequencies prior to forming the cross-over difference.

Frequency data may, of course, be analysed using Poisson regression. Such an approach would model the expected number of events as a function of patient, period and treatment effects and assume that the actual number of events followed a Poisson distribution given the expected value. Such models are widely used in medical statistics and are special cases of generalized linear models (Nelder and Wedderburn, 1972). Discussions of Poisson regression will be found in Dobson (1983), Krzanowski (1998), Lindsey (1996) and McCullagh and Nelder (1989). An application of Poisson regression to the cross-over trial will be considered later in this chapter.

4.3 NON-PARAMETRIC METHODS

There are few topics in statistics which divide opinion so strongly as non-parametric methods. Statisticians tend to be either strongly for them or against them. On the one hand non-parametric methods are praised for their independence from distributional assumptions, on the other they are attacked for their dependence on randomization arguments and their irrelevance to inference. In certain fields they seem to have achieved an overwhelming position in the analysis of univariate data without having diminished the interest in parametric multivariate methods: the implication is that in such fields most multivariate data are Normally distributed and few univariate data are!

Here I shall adopt an intermediate position between the two extremes. On the one hand I am of the opinion that a rather dishonest use of non-parametric methods is frequently made in medical statistics. For example, a non-parametric test is performed and then a parametric interpretation is given to the treatment effect, as is the case when a rank test is carried out and then a difference between mean responses is used to describe the treatment effect. On the other hand I think the objections made to them from certain quarters are a bit ‘nice’. Such methods are not restricted to testing, as is often claimed, but embrace estimation as well, even if the confidence interval methods associated with them have a justification via hypothesis testing which some find objectionable. They undeniably do something and seem to provide adequate estimates for a wide range of parametric models, so they can be regarded as useful robust alternatives when there is extreme doubt about the best parametric approach to take.

Furthermore, methods which explicitly reflect the distribution of the test statistic over all randomizations undeniably have their place in experiments for which blinding is considered important. Consider the example of a so-called '*n* of 1' trial (to be discussed in Chapter 7) in asthma in which a single patient is assigned in a blind manner on three occasions to a bronchodilator and on three occasions to a placebo, the order of assignment being perfectly random and adequate washout being provided between treatments. Suppose we regard the period effect as unimportant and record the patient's peak expiratory flow 8 h after dosing, as was done in Example 3.1. Suppose we obtain the following results for PEF in ℓ/min:

Bronchodilator:	425, 450, 430
Placebo:	265, 255, 290.

If we use the two-sample *t* test to compare the two treatments we shall obtain a *t* statistic for the difference between treatments of 12.8 on 4 degrees of freedom. The critical value for a 0.0005 significance level (one-sided) is 8.6. The result is thus extremely impressive.

Suppose, however, that the patient had an extremely strong prejudice that the bronchodilator worked and was psychologically very suggestible. (After all, we 'blind' patients because we fear this sort of thing.) He might then have guessed on which day he was being given a bronchodilator. The probability of his guessing correctly is then equal to the probability of dividing 6 test days correctly into two groups of 3 or $6!/(3!3!) = 1/20$. Having guessed correctly his prejudice might lead him to produce the sort of values we observed. In that case 1/20 is a more reasonable *P* value for the observed *t* statistic. But 1/20 is exactly the probability we should get were we to perform a *Wilcoxon rank sum test* on the observed results. Thus the non-parametric procedure reflects the degree of protection to extreme prejudice offered by blinding whereas the parametric *t* test goes beyond the protection offered by blinding in an attempt to extract further information.

I do not consider this to be an overwhelming argument for preferring randomization tests or non-parametric methods but I think it provides a warning that extremely small *P* values (or for that matter impressive likelihoods) from very small samples for experiments in which blinding is considered important must be treated with caution: a sort of dual to the Bayesian viewpoint that moderate *P* values are more indicative of the falsity of the null hypothesis with small samples than with large ones. (See Lindley and Scott, 1984 p. 3 and Royall, 1986). Such extremely small *P* values can only be obtained for small samples using parametric methods.

The place of non-parametric tests in the analysis of the *AB*/*BA* cross-over trial will be considered again from a practical point of view in Section 4.3.13. For the moment we shall outline some techniques. These may be divided into two groups. Those to be used ignoring the effect of period and those allowing for it. We now consider these in turn.

4.3.1 Ignoring the effect of period

If the investigator believes that the possible effect of period on his results is negligible he will be prepared to allocate his patients completely at random to the two sequence groups. If he is sincere about his beliefs there will actually be an advantage in any double-blind experiment in doing so since it increases the number of possible allocations of patients to sequences and hence the degree to which he may be blinded himself. When patients have been allocated at random in this way various approaches are possible.

4.3.2 Methods based on the sign test

Consider the VAS data given in Example 4.1 and suppose simply for the sake of illustration that the patients had in fact been allocated completely at random to the two sequence groups. VAS scores being in any case very difficult to interpret and having a rather awkward and not easily determined distribution, we might argue that the most we ought to expect to do is simply note for each patient under which of the two treatments he was better. This may be done quite simply by noting the sign of the basic estimator in Table 4.2. If it is positive, he rated salbutamol more highly than formoterol whereas if it is negative, he rated formoterol more highly. For eleven of the patients the sign of this difference is negative and for one (patient 4) it is positive.

We may now carry out a form of significance test known as the sign test (Sprent and Smeeton, 2001, pp. 17, 57–60). There are two possible lines of justification. First, we might regard the patients as a random sample of a possible population of patients. We argue that under the null hypothesis of equality of the treatments, in such a population the probability of observing a plus should be equal to that of observing a minus. If we ignore the possibility of ties, this probability equals 1/2 and we may substitute this value in the general formula for a binomial. If n is the number of results and X is the number of minuses, the probability of obtaining exactly r is then given by:

$$P(X = r) = \frac{n!}{r!(n-r)!}\left(\frac{1}{2}\right)^{n}. \tag{4.7}$$

We are now interested in calculating the probability of observing 11 or more minuses in 12 results, which is $P(X \geqslant r)$ rather than $P(X = r)$ as given in (4.7). We thus substitute 12 for n in (4.7) and first 11 and then 12 for r, adding the two probabilities together. The probability may thus be calculated as $12/2^{12} + 1/2^{12} = 0.0032$ or read off from tables of the binomial distribution. For a two-tailed value we double this probability to obtain 0.0064.

Since this is the first occasion in the book in which we have calculated a P value for a discrete test statistic we take the opportunity of noting that an alternative to calculating the P value is to calculate the *mid P* (Lancaster, 1961; Barnard, 1990). This is defined as

$$\text{mid } P = \{P(X \geqslant r) + P(X \geqslant r + 1)\}/2,$$

for a one-sided value. If the significance test is regarded as providing weight of evidence against the null hypothesis (albeit in crude form) then for cases where the test statistic is discrete, mid P has some advantages over the standard P value (Barnard, 1990). (For the continuous case they are identical.) For this example the mid $P = (13/2^{12} + 1/2^{12})/2 = 0.0017$, one-sided, or 0.0034 two-sided. Mid P suffers from one important disadvantage, however, and that is that it is rarely used. Having raised this interesting topic, therefore, we shall revert to convention and from now on we shall only calculate conventional P values.

Returning to the sign test, an alternative argument uses the randomization performed. We argue that if the two drugs are equivalent in the group of patients studied we should observe exactly the same differences when comparing period 1 with period 2. If we calculate a basic estimator, however, we sometimes subtract the period 1 value from the period 2 and vice versa. This depends on the treatment sequence to which the patients are allocated. Under this argument in the absence of a treatment effect, the consequence of allocating a patient to the *BA* sequence rather than the *AB* sequence is simply to reverse the sign of his basic estimator. There are 2^{12} possible allocations of 12 patients to two sequences if we do this completely at random. (It is not, in fact, likely that patients were allocated completely at random in this trial.) Thirteen of these sequences would produce 11 or more minuses. The probability of observing a result as remarkable as this, therefore, under the null hypothesis is $13/2^{12} = 0.0032$ as before.

Note that exactly the same allocation of signs occurs if we use the logit score differences, or the arc sine score differences in Table 4.3 rather than the VAS score difference itself. This illustrates the strong degree of invariance to the original scale of measurement observed by the sign test. This invariance forms both its strength and its weakness. In order to carry out the test very little has to be assumed about the underlying measurements. On the other hand if we try to reverse the process and use the test as a basis for making statements about effects on a given scale of measurement we have difficulties.

Suppose we wish to construct a confidence interval for the treatment effect. We shall have to commit ourselves to a scale of measurement. We now illustrate the argument using the original VAS scale as follows. Suppose we take any given possible true difference between the treatments regarding the VAS score. Let us for argument's sake take the value of -15 mm. We now find that 10 out of the 12 basic estimators are less than -15 mm. This is equivalent to saying that if we substract -15 mm from the basic estimators, then 10 of the results are negative and 2 positive. We may now use the sign test on these results. We thus calculate that the probability of observing results as remarkable as these is the probability of observing 12, 11 or 10 minuses or $(1 + 12 + 66)/2^{12} = 0.0193$. For a two-tailed test we double this value to 0.0386. Thus, at the 5% level of significance, we would reject the null hypothesis that the treatment effect is -15 mm.

We could establish by trial and error all the values of the VAS score which when adopted as a null hypothesis for the treatment effect would be rejected at the 5% level. By definition all of those which would not be rejected at the 5% level would form a 95% confidence interval for the true treatment effect. Such a procedure is not, however, necessary. We may note from the tables of Lindley and Scott (1984), that the one-tailed probability associated with 9 successes is 0.0730 and thus exceeds the 0.025 limit required for a two-sided 5% significance level. On the other hand the value associated with two successes, as noted above, is 0.0193. If we now take the basic estimators and rank them in descending order of magnitude we get the following results, in millimetres:

$$1, -8, -18, -18, -19, -24, -28, -34, -36, -49, -51, -60.$$

If we subtract any value less than -18 mm from this list we shall be left with fewer than 9 minuses. If we make sure that the value we subtract is greater than -49 mm we shall not have more than 9 pluses. Thus the range -49 mm to -18 mm includes all possible values of the null hypothesis for the treatment effect which would not be rejected at the 5% level. Thus

$$-49 \text{ mm} \leqslant \tau \leqslant -18 \text{ mm}$$

is a 95% confidence interval for the treatment effect. If we want a point estimate for an effect, then a natural one to choose in association with this technique is the median basic estimator, namely -26 mm. These results may be compared to the value of -29 mm with confidence limits of -41 mm and -16 mm obtained in Section 4.2.2.

Note that although the sign test is invariant when used to test the null hypothesis of exact equality of treatments, the confidence interval associated with it is not. This is because any monotonic transformation of the original values preserves the sign of the basic estimator but not the ranking of those differences, and it is not possible to obtain, for example, the basic estimators in logit scores directly from the basic estimators in VAS scores.

4.3.3 Methods based on the Wilcoxon signed rank test

When looking at the basic estimators for the VAS scores, it is noticeable that the one positive value is also the smallest in absolute terms. We may consider this fact to be informative in itself and argue that of all the 12 possible ways in which we could have obtained one positive value we happen to have got the one which is most consistent with formoterol being superior to salbutamol. Thus the size in absolute terms of the difference is a further aid to ranking the possible outcomes from the trial.

A test which makes use of such features is the *Wilcoxon signed rank* test (Sprent and Smeeton, 2001, pp. 43–57). What we do is rank the outcomes in terms of their absolute magnitude, ignoring sign, and then attach signs to the ranks once these have been established. The sum of the signed ranks is then calculated. For Example 4.1 all of the ranks 2 to 12 are negative, rank 1 is positive and the result of performing this calculation is $-$ 76. We now argue as follows. There are $2^{12} = 4096$ possible ways of allocating negative or positive signs to the 12 ranks but only 2 of them (the case where all are negative or all except the first are negative) produce values as low as $-$ 76. Thus the probability of observing a result as low as this is $2/4096 = 0.0005$. Doubling this to obtain the two-tailed P value we have $P = 0.001$.

In practice we do not have to obtain the sum of all the ranks to define the value of the statistic. This is because in general if n is the number of observations, the sum of the absolute ranks is $n(n + 1)/2$. In this example $n = 12$ and the sum of the absolute ranks is 78. As soon, therefore, as we are informed that the sum of the positive ranks is 1 we know that the sum of the negative ranks is -77 and hence that their combined sum is -76. The test statistic may thus equally well be defined in terms of the sum of the positive or negative ranks (whichever is most convenient) as in terms of the rank sum. This fact is made use of in tables of critical values of the signed rank (Lindley and Scott, 1984; Diem and Seldrup, 1982) from which we may read that the critical value of the sum of the positive ranks for a sample size of 12 for a one-tailed test at the 2.5% level is 13.

We may use this critical value to establish confidence limits in a similar way to the way we did for the sign test. Consider again as we did in Section 4.3.2 above the null hypothesis that the treatment effect is -15 mm. We express the basic estimators as differences from -15 mm. The values are listed below in terms of their original ranking with their new signed ranks noted below:

differences (mm)

$$16,\ 7,\ -3,\ -3,\ -4,\ -9,\ -13,\ -19,\ -21,\ -34,\ -36,\ -45$$

ranks

$$7, 4, -1, -2, -3, -5, -6, -8, -9, -10, -11, -12.$$

The sum of the positive signed ranks is thus 11 and, since the critical value is 13, this is significant at the 5% level. Hence we reject the null hypothesis that the treatment effect is − 15 mm. By trial and error we may attempt various values for the null hypothesis and see how the rank sum is affected. We seem to be near the required value. Using a null hypothesis of − 17 mm, for example, produces a sum of positive ranks of 13, whereas − 18 mm gives a value of at least 15 (there are some ties, a point which will be dealt with in due course). Thus the lower confidence limit lies between − 17 mm and − 18 mm.

There is, however, a more systematic way of going about this. Note that if we subtract a given value from each of the basic estimators the rank sum can change only if one of the following happens:

- one or more basic estimators thus adjusted changes sign;
- two or more basic estimators thus adjusted change in order of absolute magnitude.

Now, to adjust a positive basic estimator to change sign we must subtract from it a greater value than itself, whereas to adjust two basic estimators so that they have different ranks in absolute magnitude we must subtract from them a value greater than their pairwise mean. (To see that this is so it is best to experiment with some pairs of values.) Thus if we form all pairwise averages of the basic estimators including the averages of the differences with themselves we shall have established all of the possible pivotal values at which the rank sum may change. For n basic estimators there are $n(n-1)/2$ possible pairwise means of the basic estimators and a further n formed as means with themselves. There are thus $n(n+1)/2$ pivotal values, and this also happens to be the maximum which it is possible for the sum of the positive ranks to achieve. If we rank these $n(n+1)/2$ means in order of magnitude we shall then have obtained all of the pivotal values corresponding to a change in the sum of the positive ranks as it moves from 0 to $n(n+1)/2$.

The pairwise means for this example have been set out in a Table 4.4. The figures below them in brackets are their ranks. Where two or more pairwise means are tied the relevant ranks have been assigned arbitrarily amongst them. Since the critical value of the positive rank sum is 13 the 14th counting from the beginning and the 14th counting from the end give the relevant confidence limits. The confidence interval for the treatment effect is thus

$$-41.5\,\mathrm{mm} \leqslant \tau \leqslant -17.5\,\mathrm{mm}$$

Table 4.4 Pairwise means in mm VAS and associated (ranks) of basic estimators given in Table 4.2.

	1	−8	−18	−18	−19	−24	−28	−34	−36	−49	−51	−60
1	1 (1)	−3.5 (2)	−8.5 (4)	−8.5 (5)	−9 (6)	−11.5 (7)	−13.5 (10)	−16.5 (13)	−17.5 (14)	−24 (30)	−25 (32)	−29.5 (43)
−8		−8 (3)	−13 (8)	−13 (9)	−13.5 (11)	−16 (12)	−18 (15)	−21 (22)	−22 (26)	−28.5 (41)	−29.5 (44)	−34 (50)
−18			−18 (16)	−18 (17)	−18.5 (19)	−21 (23)	−23 (27)	−26 (33)	−27 (37)	−33.5 (48)	−34.5 (53)	−39 (61)
−18				−18 (18)	−18.5 (20)	−21 (24)	−23 (28)	−26 (34)	−27 (38)	−33.5 (49)	−34.5 (54)	−39 (62)
−19					−19 (21)	−21.5 (25)	−23.5 (29)	−26.5 (36)	−27.5 (39)	−34 (51)	−35 (55)	−39.5 (63)
−24						−24 (31)	−26 (35)	−29 (42)	−30 (45)	−36.5 (58)	−37.5 (59)	−42 (66)
−28							−28 (40)	−31 (46)	−32 (47)	−38.5 (60)	−39.5 (64)	−44 (70)
−34								−34 (52)	−35 (56)	−41.5 (65)	−42.5 (67)	−47 (71)
−36									−36 (57)	−42.5 (68)	−43.5 (69)	−48 (72)
−49										−49 (73)	−50 (74)	−54.5 (76)
−51											−51 (75)	−55.5 (77)
−60												−60 (78)

The median of these pairwise means is known as the *Hodges–Lehmann estimator* and provides a suitable point estimate. The 39th value is −27.5 mm and the 40th is −28 mm, so the Hodges–Lehmann estimator is 27.75 mm.

Completing Table 4.4 by hand in its entirety is extremely tedious and for cross-over trials with many patients the labour becomes prohibitive. It is not, however necessary to complete the whole of the table to obtain the confidence limits and the median. The argument may proceed as follows.

First we may note that if we sort the original twelve basic estimators in descending order as was done in the table then the highest of any of the pairwise means will be the first member in a row. We fill in the main diagonal, which is simply the same as the original basic estimators, but assign no ranks as yet to these means. We then start with the first member of the first row and proceed to fill out the row with pairwise means assigning the ranks (1), (2) etc. until we get to a value which is lower than the first member of the second row. We record the mean in its correct place but do not assign it a rank. We then proceed with the second row, filling out means and assigning ranks until either we reach a mean which is lower than the mean with the unassigned rank in the first row or lower than the first mean in the third row. We record the last mean calculated in the second row but do not assign it a rank and proceed with our search. At any stage we always check the last value in the row we are working with against the last completed value in all the rows. We do this until we have assigned ranks to 14 means. The same may be done at the other end of the table. Obviously for sample sizes other than 12 we shall have different critical values for the sum of positive ranks. These critical values may be obtained from Table 20 of Lindley and Scott (1984) or from Diem and Seldrup, (1982, p. 163).

To obtain the Hodges–Lehmann estimator we first guess its value. A convenient guess is the median of the original basic estimators, −26 mm. Any pairwise sum greater than −52 mm will produce a pairwise mean greater than this. We can now quickly establish for each row what the last member of the row is with a pairwise mean greater than −26 mm. We write these down in the correct position in the table noting that from row 1 to 6 they are −25 mm, −22 mm, −23 mm, −23 mm, −23.5 mm, −24 mm. By counting for each row how many pairwise means should precede it we obtain $10 + 7 + 4 + 3 + 2 = 26$ as the number of values less than −26 mm apart from the 6 we have noted. We thus know that −25 mm is the 32nd value. We therefore carry on our search in this region until we have found the 39th and 40th values and so the Hodges–Lehmann estimator.

Useful discussions of these topics will be found in Sprent and Smeeton (2001) and, at a more advanced level, in Lehmann (1975).

4.3.4 A permutation test

The Wilcoxon signed rank test has two aspects. First, the replacement of the actual difference by a score which is the signed rank. Second, the calculation of how unusual the sum of the scores so produced is when compared to all possible sums based on all possible assignments of positive and negative values to these scores (that is to say, ranks).

An alternative test, the permutation test, simply ignores the first step. The values are not replaced by their ranks. Instead they are added up as they are. The 'scores' are thus the original differences rather than their ranks. Thus, the sum of the values we obtain in this instance is -344. This is then compared to all other possible sums that we could obtain. As we have already established, there are $2^{12} = 4096$ possible ways of assigning positive and negative signs to the ranks and clearly there are the same number of ways of assigning these to the original differences. Thus, counting tied sums separately, we have 4096 possible values of this statistic.

It is not hard to see in this instance that the most extreme possible value that could have been obtained for this statistic is –346. This would apply if the one positive value, 1, had been negative instead. It is also obvious that for the split actually observed we have the second most extreme value of the statistic. Therefore the probability of observing a result as extreme as or more extreme than that observed is 2/4096. To obtain the two-tailed P value we double this and obtain $4/4096 = 1/1029 = 0.001$ as when using the signed rank test.

Remark. For this example the signed rank and permutation tests agree. Obviously, methodologically there is a close similarity between the two procedures. Nevertheless, they do not have to agree. They agree in the sense that the number of possible values (distinguishing between ties) that the two sums of scores (ranks in the one case and differences in the other) can achieve are identical. In the most extreme case they will agree with each other and, indeed, with the sign test. In general, however, they do not have to agree and it should be noted that the rank test is more robust (in the sense of less influenced by extreme outliers). For example, here the minimum value that a difference could in theory obtain is -100 mm and the maximum is 100 mm, and such differences could be considerably different from those that typically obtain. On the other hand, the ranks must range from 1 to 12.

4.3.5 Computer analysis ignoring the period effect

Both the sign test and the Wilcoxon signed rank test may be carried out quite simply using SAS®. It is necessary, of course, to calculate the basic estimators first within SAS®. If, as in Chapter 3, we give these the variable name, *BASICEST*, then the following code:

```
proc univariate;
  var BASICEST;
run;
```

calculates a number of statistics, including the median of the basic estimators, and the P values for the sign test and the Wilcoxon signed rank test. For obtaining confidence intervals for the sign test *proc sort* may be used on the basic estimators and the relevant values thus identified using tables of the binomial distribution. A check may be carried out using *proc univariate*. For example, if the postulated lower 95% confidence interval is subtracted from the basic estimators and proc univariate is run, the resulting P value for the sign test should not be less than 5%, whereas if the next lowest basic estimator is substracted it should be. Similarly by subtracting the upper confidence interval and the next highest value a check may be performed.

The gold standard for non-parametric analysis is provided by StatXact®. This is also available through SAS® (for an additional fee) as ProcXact®. StatXact® will take data in the form of a cross-tabulation (which it refers to as TableData) or as a series of variates indicating outcomes and also group membership (which it refers to as CaseData). Since we shall refer to StatXact® again later in this chapter, we reproduce all the data we shall be using in CaseData form in Table 4.5. *PERDIFF*, *SEMIDIFF* and *SEQUENCE* are not needed here and will be referred to subsequently. *BASICEST* is as already defined.

The commands in StatXact® for performing the sign test, Wilcoxon signed rank test, the permutation test and for calculating the 95% confidence limits for the Hodges–Lehmann estimator are

```
PS SI/EX /PD=BASICEST
PS WI/EX /PD=BASICEST
PS PE/EX /PD=BASICEST
PS HL/EX /PD=BASICEST
```

respectively. In fact the analyses are also extremely easily done via menus. Here *PS* stands for 'paired sample', *SI* for 'sign', *WI* for 'Wilcoxon', *PE* for 'permutation' and *HL* for 'Hodges–Lehmann'. *EX* specifies that an exact result is required,

Table 4.5 (Example 4.1) CaseData for StatXact® analyses

BASICEST	PERDIFF	SEMIDIFF	SEQUENCE
−19	−19	−9.5	1
−60	−60	−30	1
−34	−34	−17	1
−18	−18	−9	1
−51	−51	−25.5	1
−8	−8	−4	1
−28	28	14	2
1	−1	−0.5	2
−36	36	18	2
−18	18	9	2
−24	24	12	2
−49	49	24.5	2

although the output also includes asymptotic approximations. *PD = BASICEST* shows that the paired difference being analysed is *BASICEST*, the basic estimator.

The results are as we have obtained here by hand with one exception. StatXact® calculates the 95% confidence limits as −39.5 mm and −18 mm. These are, in fact, the lowest (−39.5 mm) and highest (−18 mm) values, which, when subtracted from the individual basic estimators, would *not* lead to a rejection at the 2.5% level using the Wilcoxon signed rank test. Although these are conservative in terms of hypothesis testing, they lead to narrower confidence intervals than we have chosen above and so, from one point of view, are anti-conservative. The limits illustrated in the hand calculations above are, instead, the values that will just lead to rejection. In practice, the difference is likely to be unimportant.

An alternative computer package for performing this and other non-parametric analyses is the simple (and cheap) add-on to Excel®, StatPlus®, provided by Berk and Carey (2000). This has a sign test module and a Wilcoxon signed rank module. Data can either be put in as two matched columns or as a single column of differences. The confidence level can be specified by the user. In addition to performing the test, for the sign test, the module will calculate three kinds of confidence interval: conservative, which has a confidence of at least $1 - \alpha$; liberal, which has a confidence of at most $1 - \alpha$; as well as an approximate interval with confidence α. Which of the three is calculated is specified via a pull–down menu with the options 'at least', 'at most' and 'approximately', respectively. However, the resulting output is dynamic and the user can change his mind afterwards. If this module is run using the 'at least' option and requesting 95% confidence, then the same results as we have calculated more laboriously by hand in Section 4.3.2 are produced. The Wilcoxon signed rank module does not have these options but produces the same 95% confidence limits for this example as we calculated in Section 4.3.3.

Analyses using GenStat® and S-Plus® will be illustrated in appendices to this chapter.

4.3.6 Allowing for the effect of period

The investigator may fear the presence of a period effect. There are then two possible undesirable consequences of allocating patients at random to the two sequences and using the methods of Section 4.3.1 above. First, there may be a loss of power because a possible source of variation, variation between periods, has not been eliminated. Second, a particular allocation may result in an imbalance of patients between sequences. The investigator may then fear that his treatment estimates and tests are biased. Although he knows that over all randomizations they would be unbiased he fears that conditionally, given an observed imbalance between sequences, they may be biased.

These two problems may be dealt with by using the methods outlined in this section. These methods specifically allow for the period effect. In so doing they do not require patients to have been allocated in equal numbers to the two sequences. A gain in power, however, will result from such an allocation.

4.3.7 A period adjusted sign test

We may adapt the sign test to test for treatment effects whilst allowing for any period effect. Consider the basic estimators for Example 4.1 in Table 4.2. If there is a period effect then the differences between groups for the basic estimators will not just reflect random variation but also the period effect. For example, other things being equal, if patients tend to give higher VAS scores in the second period than in the first we shall find lower basic estimators from the first sequence group than from the second. We should thus be more likely to see a positive basic estimator in the second sequence group than in the first. If there were no treatment effect, however, whatever the trend, on average we should expect to find as many positive period differences in one sequence group as in another. This will not be the case if there is a treatment effect. Thus comparing the two sequences for the signs of the period effects is a way of testing the null hypothesis of no treatment effect.

The basic estimators in Table 4.2 were calculated by subtracting the results under salbutamol from those under formoterol. If we form the period differences by taking the results in period 2 from those in period 1, then for the first sequence group the period difference is the same as the basic estimator and for the second sequence group we may obtain the period differences by reversing the signs of the basic estimators. If all we are interested in is the sign of these period differences then we may gather the result in a *four-fold table* as has been done in Table 4.5. Here it may be seen that for all of the patients in the first sequence group the period difference is negative, whereas for the second sequence group, with the single exception of patient 4, the period difference is positive.

We may now use a statistical test conditioning on the margins to see whether there is any association between group and sign, such an association being indicative of a treatment effect. The appropriate test is *Fisher's exact test* (Fisher, 1990a, p. 96; Sprent and Smeeton, 2001, pp. 322–4). For this test we accept that we are evaluating Table 4.6 for its unusualness under the null hypothesis of no treatment effect by comparing it to all tables with the same margins. That is to say comparing it to all tables with 6 patients per sequence showing 7 negative and 5 positive period differences. The probability argument will not be rehearsed here. The reader is referred to the texts above or, indeed, to any elementary statistical text book.

Table 4.6 (Example 4.1) Four-fold table classifying patients by sign of period difference and sequence of treatment.

	Sign of period difference		
Sequence group	−	+	
for/sal	6 (a)	0 (b)	6 $(a+b)$
sal/for	1 (c)	5 (d)	6 $(c+d)$
	7 $(a+c)$	5 $(b+d)$	12 $(a+b+c+d)$

In general for any table with fixed margins the probability of a given configuration may be calculated as

$$\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!\; a!\; b!\; c!\; d!}. \tag{4.8}$$

In this particular example substitution of 6 for a, 0 for b, 1 for c and 5 for d in (4.8) yields a probability of 0.0076. In general, in order to calculate the P value we should also evaluate all more extreme tables. That is all those which showed even more evidence for an association between sequence group and the sign of the period difference. In this case we have a most extreme table. It is impossible to increase the number of minuses in the sal/for sequence. Thus 6 is the maximum value which a may obtain. There are various suggestions, however, as to how one may obtain the two-sided P value. A simple doubling has been recommended by Yates (1984). This gives a P value of 0.0152.

4.3.8 An adaptation of the Brown–Mood median test

If the trial is strongly affected by a trend it may be the case that a high proportion of the period differences will turn out to be of one sign. Under such circumstances the test described above may lack power. A more sensitive classification of the period differences may be made with respect to the median of all period differences. Instead of classifying period differences as ‘positive’ or ‘negative’ we may then classify them as ‘high’ or ‘low’ and proceed to test for association between group and value of the period difference as before. Such a test is then an adaptation for cross-over trials of a test used more generally to compare two samples for location: the *Brown–Mood median test* (Gibbons, 1982).

In Example 4.1 the median of all period differences is −4.5. All of the period differences in the formoterol/salbutamol group are below this value. All of the period differences in the salbutamol/formoterol group are above this value. We thus have the classification in Table 4.7. This is very similar to that seen in Table 4.6 but one patient, patient 4, has moved from contributing to the cell total *c* and now contributes to *d*. This is the only patient in the salbutamol/ formoterol sequence whose period difference is negative, the point being that this period difference, although below zero, is nevertheless ‘high’ when judged by the trend established for the trial as a whole.

Once the classification has been made the analysis proceeds as in Section 4.3.6, only now we may substitute 0 for *c* and 6 for *d* in (4.8) with *a* and *b* having the values 6 and 0 as before. This yields a one-tailed *P* value of 0.0011 which may be doubled to give 0.0022. Alternatively such probabilities may be obtained from suitable statistical tables, for example Table 26 of Lindley and Scott (1984).

For large samples, a *chi-square* with *Yates’ correction* provides a useful approximation to Fisher’s exact test. This we may calculate as

$$\frac{(|ad - bc| - n/2)^2 n}{(a+b)(a+c)(b+d)(c+d)}, \tag{4.9}$$

where $n = a + b + c + d$. The calculated value must then be referred to a table of the chi-square with one degree of freedom.

Substitution of the relevant values from Table 4.7 yields a chi-square value of 8.33 with an associated two-tailed *P* value of 0.0039. The corresponding approximation for Table 4.6 yields a chi-square value of 5.49 with a two-tailed *P* value of 0.019. These are quite reasonable approximations to the exact probabilities.

If we compare the two tests, the adapted Brown–Mood median test is more sensitive than the period corrected sign test but also slightly less robust. Any monotonic transformation of the original VAS scores will preserve the sign of

Table 4.7 (Example 4.1) Four-fold table classifying patients by magnitude of period difference and sequence of treatment.

	Period difference		
Sequence group	low	high	
for/sal	6 (*a*)	0 (*b*)	6 (*a* + *b*)
sal/for	0 (*c*)	6 (*d*)	6 (*c* + *d*)
	6 (*a* + *c*)	6 (*b* + *d*)	12 (*a* + *b* + *c* + *d*)

the period differences. The same is not true of the ranking of these differences and it is always possible that a transformation of original scores might lead to a different result for the median test. The period corrected sign test is thus more strongly *invariant* than the median test. In practice, however, the median test provides the more attractive of the two options. Since by definition half of the values must be above and half below the median if patients are allocated in equal numbers to the two sequences it has the added advantage of leading to a contingency table in which all margins are fixed and equal.

Nevertheless it is a rather pessimistic use of continuous data to reduce within-patient differences to 'low' and 'high' values and a yet more attractive alternative which we now consider is to use information on the ranking of these differences.

4.3.9 Koch's adaptation of the Wilcoxon–Mann–Whitney rank sum test

In an important paper on cross-over trials Koch (1972) proposed a number of non-parametric procedures for performing various hypothesis tests in connection with Grizzle's (1965) model of cross-over trials. Amongst the various tests proposed was a rank test for treatment effects in the presence of period effects. Like the procedures outlined in Sections 4.3.6 and 4.3.7 it is based on period differences, and like them and the parametric within-patient tests discussed in Chapter 3 and under Section 4.2 it assumes that there is no carry-over.

Koch's procedure consists of ranking the period differences for all of the patients in the trial and then using the *Wilcoxon–Mann–Whitney test* for differences between the two sequence groups. We shall illustrate Koch's test on Example 4.1 using the *Mann–Whitney U statistic* (Sprent and Smeeton, 2001, pp. 147–55).

A simple inspection of Table 4.2 shows that if all the patients are ranked in ascending order in terms of the period differences, then the ranks for the for/sal group are, in order of patient number

$$4, 1, 3, 5, 2, 6$$

.

The ranks for the sal/for group are

$$10, 7, 11, 8, 9, 12.$$

For the two groups the sums of the ranks are 21 and 57 respectively. To calculate the Mann–Whitney U statistic we subtract the lowest possible score from the lower of the two rank sums. Since we clearly have an extreme partition of the ranks here the lowest possible score is obviously 21 and hence the Mann–Whitney U statistic equals 0. (In general if we have n patients in a group the lowest possible rank sum is simply the sum of the integers 1 to n and hence equals $n(n+1)/2$.) Consulting Table 21 of Lindley and Scott (1984) we find that the appropriate critical value for a two-sided test at the 5% level is 5, whereas the critical value for a 1% test is 2. Hence the result is significant at the 1% level.

The calculation of the P value is particularly easy in this example. There are $12!(6!6!) = 924$ ways of partitioning 12 ranks into two groups of 6, only one of which produces a value of U as low as this. (In general we have to count all the ways in which we could partition the ranks to obtain values of U less than or equal to the value observed.) The probability of observing such a partition, therefore, is $1/\{12!/(6!6!)\} = 0.0011$. Doubling this yields a two-sided P value of 0.0022. It should be noted that this is exactly the same as that yielded by the Brown–Mood median test above. This is because we have observed the most extreme partition possible. When this occurs the two tests yield identical results.

The test we have just performed is a test of the null hypothesis that there is no treatment effect. On the assumption that any treatment effect will be constant for all patients any given hypothesis about its value may be tested. Suppose, as in Section 4.3.2, that we wish to test the null hypothesis that the treatment effect is − 15 mm. We now need to measure the period differences with respect to the value this treatment effect would produce. Consequently we must subtract − 15 mm from the period differences in the formoterol/salbutamol sequence and add − 15 mm to the difference in the salbutamol/formoterol sequence. Doing this we obtain the following differences and (ranks):

formoterol/salbutamol	− 4,	− 45,	− 19,	− 3,	− 36,	7
	(5)	(1)	(3)	(6)	(2)	(8)
salbutamol/formoterol	13,	− 16,	21,	3,	9,	34
	(10)	(4)	(11)	(7)	(9)	(12).

The sum of the ranks in the group with the lowest sum is thus 25 and the Mann–Whitney U statistic is the $25 - 21 = 4$. Since we have already established that the critical value at the 5% level two-sided is 5 we reject the null hypothesis that the treatment effect is − 15 mm.

We could by trial and error establish all values of the treatment effect which, when adopted as a null hypothesis, would not lead to its rejection at the 5% level. The set of all such values would then constitute a 95% confidence interval for the true treatment effect. As for the sign test and the signed ranks test, however, a more systematic approach is possible (Sprent and Smeeton, 2001, pp. 153–5; Jones and Kenward, 1989, pp. 58–9). We note, first of all, that the U statistic may change only if a period difference within one sequence group changes ranks with a period difference in another sequence group. (In any case since period differences in the same treatment sequence are subject to the same treatment effect they never change their ranking with respect to each other.) Consider now an example as to how the *U* statistic may change. Consider the two patients from the two sequences closest to each other in terms of period difference, namely patient 12 for whom the value is − 8 and patient 4 for whom the value is − 1. Any value of the treatment effect less (i.e. more strongly negative) than − 3.5 would, when adopted as a null hypothesis, lead to adjusted period differences, which produced changed ranks for these patients. But this value is simply half the difference between the two period differences, or equivalently the mean of the basic estimators. We may thus limit our search to the pivotal values produced by obtaining all pairwise means of the basic estimators where these means are formed by obtaining one value from each of the two sequence groups. This has been done in Table 4.8. The Mann–Whitney *U* statistic can in fact be calculated directly as the number of positive pairwise means in this table. There are no positive means and so $U = 0$. (This example, constituting an extreme case, is not particularly illuminating as the basis for a discussion of the equivalence of the two different formulations of the Mann–Whitney *U* and so we defer it until later. For the moment the reader is simply asked to accept that they are equivalent.)

The Hodges–Lehmann estimator is the median of the values in Table 4.8 and may be calculated as the mean of the 18th and 19th value and so is $(-27.5\,\text{mm} + -28.5\,\text{mm})/2 = -28\,\text{mm}$. Note that each of the pairwise means is a mean of two basic estimators each of which exhibits the opposite period effect. On the assumption that the period effect is additive, averaging the two eliminates it. It is interesting to note that if we obtain the mean of all the pairwise means, then the estimate is 28.67 mm, which is identical to that observed using the standard Hills–Armitage (1979) approach of Section 3.6 on the original VAS scores of Example 4.1 (see Section 4.2.2). This must in

Table 4.8 (Example 4.1) Means of two basic estimators (one from each sequence) in mm VAS and associated (ranks).

		formoterol/salbutamol sequence					
		−60	−51	−34	−19	−18	−8
salbutamol/ formaterol sequence	−49	−54.5 (1)	−50 (2)	−41.5 (7)	−34 (13)	−33.5 (14)	−28.5 (18)
	−36	−48 (3)	−43.5 (5)	−35 (11)	−27.5 (19)	−27 (20)	−22 (25)
	−28	−44 (4)	−39.5 (8)	−31 (15)	−23.5 (23)	−23 (24)	−18 (29)
	−24	−42 (6)	−37.5 (10)	−29 (17)	−21.5 (26)	−21 (27)	−16 (32)
	−18	−39 (9)	−34.5 (12)	−26 (21)	−18.5 (28)	−18 (30)	−13 (33)
	1	−29.5 (16)	−25 (22)	−16.5 (31)	−9 (34)	−8.5 (35)	−3.5 (36)

general be the case, even for trials in which there are differing numbers of patients per sequence group. This may be simply shown as follows.

Suppose we have m patients in the first sequence group and n in the second. Let the ith basic estimator in the first group be $X_i, i = 1$ to m and the jth in the second group be Y_j, $j = 1$ to n. Then the Hills–Armitage (1979) approach obtains the CROS estimate of the treatment effect as

$$\hat{\tau} = \{\sum X_i/m + \sum Y_i/n\}/2,$$

which may equally well be written as

$$\hat{\tau} = (n\sum X_i + m\sum Y_j)/(2mn). \qquad (4.10)$$

On the other hand each pairwise mean consists of a statistic of the form

$$(X_i + Y_j)/2.$$

There will be mn such pairwise means. Since each X_i will be paired with n Y_j values and each Y_j will be paired with m X_i, the sum of all of the pairwise means will involve each X_i n times and each Y_j m times. Hence it may be written as

$$(n\sum X_i + m\sum Y_j)/2.$$

On dividing this by the number of pairwise means we therefore obtain (4.10), proving the result.

This equivalence provides a useful check upon the values in the table but is also of interest in bringing out the connection between the Hodges–Lehmann estimator and the CROS estimator obtained from the Hills–Armitage approach: the first is the median of the pairwise means, the second may be obtained as the mean of these pairwise means. It turns out that calculating basic estimators averaging these within sequences and then averaging the means from the sequences provides parametric treatment estimators for a wide variety of cross-over designs (Senn and Hildebrand, 1991). On the other hand, by calculating all possible means obtained by combining one basic estimator from each sequence and then obtaining the median of these means, robust estimators similar to the Hodges–Lehmann estimator may be obtained. These matters will be discussed in Chapters 5 and 6.

Returning now to the consideration of Table 4.8 we may use this to establish confidence intervals for the treatment effect. Since the critical value of the Mann–Whitney U at the 5% level (two-sided) is 5 for this example we can see that were we to subtract any value less than $-$ 16.5 mm from the table we should be left with 6 or more positive values and the result would no longer be significant, but if we subtract a higher value than this we shall be left with 5 or fewer positive value. Hence − 16.5 mm is the maximum value of the treatment effect which, when adopted as a null hypothesis, would not lead to its rejection. Similarly, if we subtract any number less than − 42 mm from the table we shall be left with 5 or fewer negative values. Hence a confidence interval for the treatment effect is

$$-42\,\text{mm} \leqslant \tau \leqslant -16.5\,\text{mm}.$$

The limits are very similar to those we found in Section 4.2.2, where we obtained values of − 41 mm and − 16 mm for the CROS analysis of the original VAS scores.

Koch's (1972) method is undoubtedly the most important of the non-parametric procedures for the *AB*/*BA* design, and because one or two features of its application were obscured by the extremely simple nature of Example 4.1 we now illustrate these by introducing a further example.

Example 4.2 In a double-blind *multiple-dose AB–BA* cross-over in stable exertional angina pectoris 60 patients were allocated at random to one of two treatment sequences *AB* or *BA*, where *A* stands for *metoprolol oros 14/190* and *B* stands for *lopresor sr*® *200 mg*. The duration of each treatment period was four weeks. There was no wash-out between treatments. At the end of the treatment period, patients performed an exercise test consisting of 3 minutes work at 50 W with 25 W increases every 2 minutes up to a possible maximum of 200 W. The exercise test was terminated as soon as the patient suffered an anginal attack or was otherwise unable to continue due to exhaustion or dyspnoea. The trial was carried out in four centres. Table 4.8 reports the results from centre 3 for heart rate in beats/minute (bpm) at the end of the exercise test.

Remark Note the precision to which the values have been recorded in Table 4.9. Fourteen of the original heart-rate measurements have 0 as the last digit, seven have 5 and three have 8. Again, we have an illustration of an extremely common feature of real data (Preece, 1981).

If the rank sum is obtained for the period differences for the second sequence group it will be found to equal 37. Subtracting 21 leaves 16 which is well above the critical value of 5 and is clearly not significant. There is no evidence of a difference between treatments. We shall use the results, however to explain the relationship between the means of the basic estimators and the Mann–Whitney *U*.

First we arrange the period differences in order and underline those belonging to the 'lop/met' sequence. The list then looks like this:

$$-\underline{40}\quad -15\quad -\underline{10}\quad -7\quad -\underline{2}\ \ 0\ \ \underline{5}\ \ 5\ \ \underline{5}\ \ 15\ \ 22\ \ \underline{25}\ \text{bpm} \tag{4.11}$$

Table 4.9 (Example 4.2) Results of a cross-over trial in angina pectoris.

		Heart rate in beats/minute (rank)				
Sequence	Patient	Metoprolol oros	Lopresor sr®	Basic estimator	Period difference	Total
met/lop	301	88 (2)	95	– 7(3)	– 7(4)	183 (2)
	304	130(11)	125	5(8)	5(8)	255(11)
	305	115(7.5)	130	– 15(2)	– 15(2)	245(9.5)
	307	125(9.5)	110	15(10)	15(10)	235(7)
	310	140(12)	118	22(11)	22(11)	258(12)
	311	110(6)	110	0(6)	0(6)	220(5)
lop/met	302	110	115(7.5)	– 5(4)	5(8)	225(6)
	303	80	105(5)	– 25(1)	25(12)	185(3)
	306	120	125(9.5)	– 5(5)	5(8)	245(9.5)
	308	140	100(4)	40(12)	– 40(1)	240(8)
	309	100	98(3)	2(7)	– 2(5)	198(4)
	312	90	80(1)	10(9)	– 10(3)	170(1)

met/lop = metoprolol oros followed by lopresor sr®
lop/met = lopresor sr® followed by metoprolol oros

Where observations have been tied with others in another sequence the mean rank has been allocated.

Now, if for every underlined observation we count the number of preceding non-underlined observations and form the total we shall get

$$0 + 1 + 2 + 3 + 4 + 6 = 16.$$

This is the alternative formulation of the Mann–Whitney *U*. Its equivalence to the other form may easily be seen as follows. Consider the highest value 25, its rank is 12 but it exceeds 6 values in the other group in size. The difference is made up by its rank, 6, within the same group. Similarly the next value in the same group has a rank of 9 and exceeds 4 values in the other group (if we accept the way in which the ties have been broken). The difference is made up by its rank in the same group. In fact for every observation the difference between its overall rank and its rank within the same group equals the number of observations in the other group which it exceeds. Hence we have the identity *for a given group*:

$$\text{total ranks overall} = \text{total ranks within group} + \text{number of observations in other group exceeded.}$$

We may rewrite this as:

$$\text{number of observations in other group exceeded} = \text{total rank overall} - \text{total ranks within group.}$$

Here, however, the left-hand side and right-hand side are simply the two formulations of the Mann–Whitney *U* statistic, which is what we had to show.

Table 4.10 presents the 36 pairwise means obtained by taking one basic estimator from each sequence group. (These means may equivalently be calculated as half the difference between the corresponding period differences.) It will be seen that 15 of these means are negative and two are equal to zero. The two that are zero, however, correspond to differences between identical period differences of value 5. In (4.11) above, when calculating the Mann–Whitney *U*, we noticed these ties and broke them in what seemed a fair manner. If we now similarly declare one of the zeros to be negative and one positive we shall then have 16 negative values. The number of negative values is thus equal to the Mann–Whitney *U* statistic.

The reason that this is so we can see by studying one of the rows within the table, say the 4th row: −6.5, −2.5, 1, 3.5, 8.5, 12 bpm. This row presents the result of subtracting the period difference of − 2 bpm in the second group from each of the period differences in the first group and dividing the result by 2. Wherever the result of this calculation is negative it shows that the value in the second group was higher than that in the first. Thus the entries in the table show us that −2 bpm is greater than −15 bpm and −7 bpm but not greater than the other values in the first sequence. Thus two is the relevant contribution to the Mann–Whitney *U* statistic from this row.

The way in which we may use the table to calculate confidence limits may now be described (Hauschke *et al.*, 1990). Any hypothesis about the value of τ becomes a new origin from which we assess the pairwise means for their unusualness. Bearing in mind that the relevant critical value of *U* corresponding to a two-sided test at the 5% level is 5 and that the 6th pairwise mean is −10 bpm, we may subtract any value greater than or equal to −10 bpm from

Table 4.10 (Example 4.2) Means of two basic estimators (one from each sequence) in beats per minute (bpm) and associated (ranks).

		metoprolol oros/lopresor sr® sequence					
		−15	−7	0	5	15	22
lopresor sr®/ metoprolol oros sequence	−25	−20 (1)	−16 (2)	−12.5 (3)	−10 (4)	−5 (10)	−1.5 (15)
	−5	−20 (5)	−6 (8)	−2.5 (11)	0 (16)	5 (21)	8.5 (25)
	−5	−10 (6)	−6 (9)	−2.5 (12)	0 (17)	5 (22)	8.5 (26)
	2	−6.5 (7)	−2.5 (13)	1 (18)	3.5 (20)	8.5 (27)	12 (28)
	10	−2.5 (14)	1.5 (19)	5 (23)	7.5 (24)	12.5 (29)	16 (31)
	40	12.5 (30)	16.5 (32)	20 (33)	22.5 (34)	27.5 (35)	31 (36)

the table and this will still leave us with more than 5 negative differences (if we conservatively assign the zero). At the other extreme, however, 5 or fewer positive differences would give us a significant result, hence we may subtract any value less than or equal to 16.5 bpm. Thus the 95% confidence limit is

$$-10\,\text{bpm} \leqslant \tau \leqslant 16.5\,\text{bpm}.$$

The Hodges–Lehmann estimator is the mean of the 18th and 19th values and equals 1.25 bpm.

4.3.10 A permutation test allowing for the period effect

Again, as was the case with the Wilcoxon signed rank test, we can produce a permutation test using the original values instead. Choosing a convenient sequence group, we thus use the sum of the original period differences rather than the sum of the ranks as our statistic. Again the task is to establish the distribution of this statistic under the null hypothesis. In general, there will be $(n_1 + n_2)!/(n_1!n_2!)$ ways of doing this. We need to calculate the sum for each of these combinations and compare the sum actually obtained to see how 'unusual' it is. This is then the basis of our significance test.

For Example 4.1 this is trivial. Since we have the most extreme case, we can spare ourselves the trouble of calculating the sums for the other permutations. Just as with the ranks, there are $12!/(6!6!) = 924$ possible sums (distinguishing between ties) corresponding to the number of ways in which 12 values can be split into two groups of six. Thus, as for the rank test, our P value is 1/924.

However, for Example 4.2, things would not be nearly so easy. The sum of the period differences is $-7 + 5 - 15 + 15 + 22 + 0 = 20$ in the first sequence group and -17 in the second. However, the sum in the first is a long way short of the maximum possible, which is $25 + 22 + 15 + 5 + 5 + 5 = 77$, and clearly there are a great number of values in between 77 and 20, the finding of which would be extremely tedious. We shall not do this but will get StatXact® to do it for us in the next section.

4.3.11 Computer analysis allowing for the period effect

Most of the tests described above, or at least approximations to them, can be performed using *proc npar1way* in SAS®. We must first of all calculate period differences for each estimate. If we create the variable *PDIFF* to store the values of these differences and the variable *GROUP* to record to which sequence group the patient belonged then the following code:

```
proc npar1way;
  var PDIFF;
  class GROUP;
run;
```

will automatically carry out a number of non-parametric tests. Included in the output will be a Wilcoxon–Mann–Whitney test in its Wilcoxon's rank sum form.

Using this on Example 4.1, SAS® provides the same rank sum, 21, as we obtained by hand. The probability calculation is slightly different, however. It provides a number of solutions, all of them approximations. The first consists of calculating the expected rank sum and its variance and hence calculating the probability using a Normal approximation.

If S is the rank sum for the group with the smaller rank sum, m is the number of patients in the group for which the sum is calculated and n is the number of patients in the other group, then we have

$$\begin{aligned} E(S) &= \frac{m(m+n+1)}{2} \\ \text{var}(S) &= mn(m+n+1)/12 \end{aligned} \tag{4.12}$$

In this case with $m = 6$ and $n = 6$ this yields an expected rank sum, $E(S)$, of 39 with a variance of 39 and a standard error of 6.245. SAS® then calculates the standardized difference of the rank sum from the expected value, applying a continuity correction by adding a half, to obtain

$$Z = \frac{S - E(S) + 0.5}{\text{sd}(S)} = \frac{21 - 39 + 0.5}{6.2445} = -2.80.$$

The two sided tail probability corresponding to this value of Z is 0.005.

It also provides a median test and again an approximation is used to obtain the P value. There is not space to cover these issues and readers desirous of obtaining further details should consult the relevant manuals. Sprent and Smeeton (2001) and Lehmann (1975) also give computational advice concerning large sample approximations to a large variety of non-parametric tests.

In this section we assume, unless specific mention is made to the contrary, that we are dealing with Example 4.1, although, in fact, most of the remarks are valid for any example.

Again StatXact® provides simple-to-use and accurate analysis. The commands for the median test, Wilcoxon–Mann–Whitney test and Hodges–Lehmann confidence intervals applied to the CaseData in Table 4.5 are

```
KI ME/EX/RO=SEQUENCE /CO=SEMIDIFF /ME=CO
TI WI/EX /RO=SEQUENCE /CO=SEMIDIFF /ST=
TI HL/EX/RO=SEQUENCE /CO=SEMIDIFF.
```

respectively. (Again analysis is extremely easily performed using menus instead.) Here *EX* requests an exact analysis as before, *RO* defines the 'row variable' to be *SEQUENCE* and *CO* defines the column variable to be *SEMIDIFF*, the semi-period difference. *TI* stands for two independent samples and *WI* and *HL* for Wilcoxon and Hodges–Lehmann, respectively. *KI* stands for '*K* independent samples' (because the median test is more generally available than for the two-sample case) and *ME* for median. *ME=CO* states that the median of the column variable is being used for the test. StatXact® offers the ability to calculate stratified Wilcoxon tests, and although we consider such tests in Chapter 6, they are not needed here. *ST=* shows that the test is unstratified. To carry-out the period adjusted sign test we work with TableData rather than case data and assume that a table such as Table 4.5 has been constructed. We then do a Fisher's exact test on this, and this may be done via a menu or with the command

```
TB FI/EX.
```

StatPlus® also provides an attractive and simple approach to analysis (Berk and Carey, 2000). The module described as the two-sample Mann–Whitney rank test may be used. Data can be input in two forms: either in two columns of period differences, one for each sequence, or as one column of differences with a second column of labels for sequences. In addition to the *P* value being calculated, confidence intervals are produced. Again the results are dynamic. The confidence levels can be changed, as indeed can the value for the null hypothesis. This makes is easy to check the confidence intervals in the way described in connection with SAS® above.

Details of analysis using GenStat® and S-Plus® are given in Appendix 4.1 and 4.2, respectively. Like SAS®, these packages do not provide non-parametric confidence intervals.

This is perhaps an appropriate point to add a general warning about computer routines for performing rank tests. The results produced on a given data set can differ considerably from package to package. Bergmann *et al.* (2000) discuss the performance of several packages in detail. They consider three computational aspects. First, how large the sample is before large-sample approximations rather than exact permutation distributions are used. Second, whether the large-sample approximation is applied with or without correction for continuity. Third, how ties are dealt with when large-sample approximations are used. They compared 11 packages, including S-Plus and SAS, on a genuine data set which had a large number of ties and found considerable differences in results between many of the packages.

Table 4.11 gives the performance of the five packages we have considered here. It will be seen that on this data set, StatXact® scores 100% for coverage and accuracy. The simple spreadsheet package StatPlus® also scores 100% for accuracy and appears to be the only package providing confidence intervals

Table 4.11 Comparison of various packages applied to Example 4.1

		Computer package				
	By hand	StatXact®	SAS®	StatPlus®	GenStat®	S-Plus®
Sign test	0.0063	0.0063	0.0063	0.0063	0.006	0.0063
Wilcoxon signed rank test	0.0010	0.0010	0.0010	0.0010	0.003	0.0033
Period adjusted sign test	0.0152	0.0152	0.015		0.015	0.0152
Median test	0.0022	0.0022	0.0009		0.002	0.0022
Wilcoxon rank sum	0.0022	0.0022	0.0051	0.0022	0.004	0.0022

for the sign test. However, neither it nor Excel® provides the means of carrying out median tests, nor, since Fisher's exact test is not provided, period adjusted sign tests. It should be noted that neither StatPlus® nor StatXact® is limited to the production of 95% confidence intervals but, for example, can handle 90% limits, commonly used in bioequivalence.

Of the three major programmable packages, SAS® appears to require the least code to produce all five statistics. (But, of course, this is at least partly dependent on the programmer's familiarity and skill with each package!) On the other hand, GenStat® and S-Plus® have rather greater flexibility if one wishes to use the results further. As far as I am aware, GenStat® is the only package to provide mid *P* values and it does this in connection with the Fisher's exact test. (This is used to perform the period adjusted and median tests.)

It remains to discuss the application, using StatXact®, of the permutation test to Example 4.2. This can be done with code

```
TI PE/EX /RO=SEQUENCE /CO=PERDIFF /ST=
```

or very easily using menus. The only difference with respect to the code for the Wilcoxon–Mann–Whitney test is the specification of *PE* for permutation test rather than *WI*. Amongst the output will be found the following:

```
Summary of Exact distribution of PERMUTATION statistic:

   Min      Max     Mean    Std-dev   Observed   Standardized
 -74.00    77.00    1.500    30.39     20.00      0.6088
```

This confirms what we had already established, that the maximum possible value is 77 and that the observed sum is 20. Of more interest to us, however, is the significance of the observed result. The two-sided *P* value is 0.6126. This can be compared to the corresponding value for the Wilcoxon–Mann–Whitney test, which is 0.8160.

4.3.12 Further non-parametric tests*

Koch (1972) describes a number of non-parametric tests apart from the within-patient test for treatment which may be carried out in connection with Grizzle's (1965) model.

Grizzle (1965) had pointed out that in the parametric case the within-patient treatment test requires an assumption of no carry-over to be valid and that a valid test for carry-over may be provided by using the totals of responses for both periods. We made a use of such totals to test for carry-over in Section 3.8 above and the only difference which a non-parametric approach makes is to use a rank test rather than a *t* test on the totals. For Example 4.2 the totals have been ranked in Table 4.8 and a test for carry-over is provided by calculating the Mann–Whitney *U* for the totals. The rank sum for the second group is 31.5 and hence the Mann–Whitney *U* is 10.5 and not significant. As an alternative non-parametric procedure to Koch's proposal the Brown–Mood median test may also be used on such totals. A sign test is not an option.

Koch (1972) also shows that a valid test for period on the assumption of no carry-over may be provided by making the same use of the basic estimators as was made for the period differences. (This is a general feature of the *AB/BA* cross-over trial which appears over and over again in tests. Differences between period differences correspond to averages of basic estimators and hence to treatment effects and vice versa.) These have also been ranked in Table 4.8. The *U* statistic is 17 and not significant. Again an even more robust alternative to Koch's (but less powerful under many plausible circumstances) is a Brown–Mood median test on the basic estimators. A treatment-corrected sign test is also a possibility by using the signs of the basic estimators rather than the period differences.

Koch also presents a valid test of the treatment effect in the presence of carry-over. This uses a Wilcoxon–Mann–Whitney test to compare the first period values between sequence groups. These are to be found in different columns of Table 4.8. The lower rank sum is provided by the period 1 values from the second sequence group and equals 30. Subtracting 21 we find that the *U* statistic is 9, which again is not significant. Again a median test would be an option.

Finally, Koch describes a *multivariate non-parametric test* which may be used to test simultaneously the null hypothesis that there is no treatment effect and that there is no carry-over effect. This is a test of the hypothesis that the two observations made on each patient come from the same distribution irrespective of the treatment sequence to which the patient was assigned. Its parametric equivalent would be a *Hotelling* T^2 test. The description of both of these tests is beyond the scope of this book. Their role in practical analysis of cross-over trials is limited.

4.3.13 Discussion

It should be noted that Freeman's result concerning the two-stage analysis for *AB*/*BA* trials, although derived in terms of a parametric analysis, will apply in very similar form to any two-stage analysis of *AB*/*BA* trials in which the choice of a within-patient test or a between-patient test using first period values is made dependent on the results of a between-patient test for carry-over. Thus the same warning applies to non-parametric tests which applies to parametric ones: *on no account must the choice of which test for treatment to use be made by carrying out a test for carry-over first*. In my opinion the logical conclusion to which this leads is that the within-patient test for treatment must always be performed.

This does not imply, however, that one should be complacent regarding carry-over. It is clearly potentially more of a worry in Example 4.2 than in Example 3.1 since the former consists of a multiple-dose trial over several weeks and the latter of a single-dose trial in products renowned for the reversibility of their effects. In Chapter 9 an examination will be made regarding practical steps which may be undertaken in designing trials to minimize the possibility of carry-over.

It would also be a mistake, however, to come away with the impression that the most serious problem in interpreting the heart rate data from Example 4.2 has to do with carry-over. As for Example 4.1 the trial is a titration. The more efficacious the drug the greater the amount of exercise that the patient will tolerate and receive. The more efficacious the drug the less will be the effect of a given constant amount of exercise on heart rate. The greater the amount of exercise, other things being equal, the greater the effect on heart rate. Putting these three things together it is not at all clear what we should expect heart rate measured at the end of an exercise test to tell us about the efficacy of the drug. Another problem concerns the objective of the trial. Metoprolol oros and lopresor sr® (sr for slow release) are two different formulations of the same drug. The object of the trial was to demonstrate their equivalence. Such demonstrations are fraught with difficulty (Makuch and Johnson, 1989; Senn, 1993a; Temple, 1982). For example, the trial was carried out double-blind but such blinding could not possibly afford the same degree of protection as in a placebo-controlled trial. If patient or investigator prejudice leads patients to believe the drugs are equivalent they do not need to know which drug was which to produce similar results. It is also fair to note that carry-over is more likely to lead to problems in an equivalence trial than in placebo-controlled trials since the most plausible consequence of carry-over is to reduce disparity between treatments.

As regards the position of non-parametric methods in cross-over trials in general I am of the opinion that in most cases it will be more fruitful to search for a suitable transformation of the data and proceed to a parametric analysis. Nevertheless it is always useful to have at one's disposition many different ways of looking at data. Non-parametric methods provide an alternative which is helpful on occasion.

4.4 BINARY OUTCOMES

A very common question which is asked of the patient in clinical trials is whether he would be prepared to take the medication again. If we exclude the possibility of a non-response or an incorrect answer then the only possible answers are NO or YES. Such outcomes are referred to as *binary outcomes* and are conveniently coded 0 and 1. Again it is quite common (although not always wise) to classify patients according to whether they improved or deteriorated during the course of a treatment. If we may exclude the possibility of the patients' being exactly the same at the end of the trial as at the beginning this is also a binary outcome. In this section we consider the analysis of such outcomes.

Curiously enough, although binary outcomes are extremely simple, their analysis is a delicate and controversial matter. A thorough treatment of the subject is quite beyond the scope of this book. All that will be attempted is a heuristic presentation of some simple techniques.

Before going on to describe the methods, however, we shall now introduce another example for the purpose of illustrating the analysis of binary outcomes.

Example 4.3 In an *AB*/*BA* double-blind cross-over trial, children aged 7 to 13 suffering from exercise-induced asthma were randomized to one of two treatment sequences. A single dose of 12 μg formoterol solution aerosol followed by a single dose of 200 μg salbutamol solution aerosol or vice versa. On an initial examination day a suitable exercise challenge was established for each child. That is to say, an exercise work load was established which induced a given response in terms of its effect in reducing forced expiratory volume in one second (FEV_1), a measure of lung function. For a given child the same challenge was used on each treatment day and was carried out twice: once 2 hours after treatment and again 8 hours after treatment. Various lung function measurements were taken throughout the trial. Amongst many variables recorded was the investigator's overall subjective assessment of efficacy. This was recorded on a four-point scale, $1 =$ poor, $2 =$ fair, $3 =$ moderate, $4 =$ good. The results are given in Table 4.12. The four-point scale will be used in an analysis below in due course. For the moment we imagine the outcome reduced to a binary scale where $- = 1$, 2 or 3 and $+ = 4$. We now proceed to consider some possible analyses of these data.

4.4.1 Ignoring the effect of period

We suppose that the four-point assessments are not available to us and that all we have are the pluses or minuses which we take to correspond to 'good' and 'not good' respectively. We also suppose for the moment that the patients have been allocated completely at random to the two sequences. In fact, this is far

Table 4.12 (Example 4.3) Investigator's assessment of efficacy on a four-point and two-point scale for a cross-over trial.

		Efficacy	
Sequence	Patient number	Formoterol	Salbutamol
for/sal	3	4 +	4 +
	4	3 −	1 −
	7	4 +	1 −
	8	4 +	3 −
	9	4 +	4 +
	11	4 +	3 −
	15	4 +	3 −
	16	4 +	1 −
	19	4 +	3 −
	20	4 +	1 −
	22	4 +	3 −
	23	4 +	2 −
sal/for	1	4 +	2 −
	2	4 +	3 −
	5	4 +	4 +
	6	4 +	4 +
	10	4 +	4 +
	12	4 +	4 +
	13	4 +	4 +
	14	3 −	4 +
	17	4 +	3 −
	18	4 +	2 −
	21	4 +	2 −
	24	4 +	3 −

from being the case. Clearly a block size of 4 has been used, since, starting with patient 1 every 4 consecutive patient numbers show an even 2/2 split to the two sequences. (This is not a good idea in my opinion and my practice is to use the largest block size possible in randomizing patients. Thus I should have preferred a block size of 24 for this trial.)

We now reduce the plus and minus data to a 'preference' by classifying patients into three groups. Those who had a plus under formoterol but a minus under salbutamol will be said to 'prefer' formoterol. Those who had a plus under salbutamol but a minus under formoterol will be said to prefer salbutamol. Those with either 2 pluses or 2 minuses have no preference. We note the following results:

Prefer formoterol	No preference	Prefer salbutamol
15	8	1
a	*b*	*c*.

We may now use *McNemar's test* (McNemar, 1947; Campbell and Machin, 1990 pp. 139–41). This consists of conditioning on the total of patients who observed a preference and calculating the probability of observing this many or more preferences for formoterol under the hypothesis of equality of the two treatments. We have 16 patients for which the investigator's rating provides a preference. Under the null hypothesis of equality of the treatments those who showed a preference would have a 0.5 probability of preferring a given drug. Thus the *P* value is calculated from the binomial distribution with $n = 16$, $p = 0.5$ using (4.7) and we have the probability of 15 or more preferences for formoterol is

$$0.5^{16} + 16 \times 0.5^{16} = 0.0003.$$

Alternatively, the value may be obtained from tables of the binomial distribution. For a two-tailed test we double the probability.

A large-sample approximation to this exact probability may be obtained using the chi-square distribution. The general formula for a chi-square goodness of fit statistic for a set of observations classified in a number of cells is

$$\sum (O_i - E_i)^2/E_i,$$

where O_i stands for the observed number of observations in the *i*th cell and E_i for the corresponding expected number. Substituting a for O_1, and c for O_2 and $(a + c)/2$ for E_1 and E_2 and making a continuity correction, we obtain

$$\frac{(|a - c| - 1)^2}{(a + c)}.$$

(The correction is 1 and not 0.5 since the difference between successive possible values of $a - c$, given a total value of $a + c$, is 2.) In this case, with $a = 15$ and $c = 1$ this yields a value of 10.56, which may be referred to a table of the chi-square with one degree of freedom and yields a two-tailed probability of 0.001.

The similarity of McNemar's test to the sign test covered in Section 4.3.2 above should be noted. The difference is that when continuous data are reduced to a preference, then in theory there are no ties. When binary outcomes are reduced to a preference a large number of ties may result and the calculation of the *P* value is conditional on the total number of patients showing a preference.

4.4.2 Allowing for period: the Mainland–Gart test

An alternative classification of the results to that used above is as follows: we note for each sequence group the number of patients who 'preferred' the treatment in the first period and the number who 'preferred' the second. Applying this representation to Example 4.3 we get the representation in Table 4.13.

Under the null hypothesis that there is no treatment effect, the pattern of preferences over time will simply reflect period effects. A treatment effect, however, would be reflected in differences from one sequence group to another for a given period. If, as in the case of McNemar's test, we ignore the no-preference patients as being uninformative we get the position in Table 4.14 below.

The Mainland–Gart test (Mainland, 1963; Gart, 1969) consists of examining this table for an association between period and sequence. A difference between treatments will tend to manifest itself as such an association. For example, the fact that in the for/sal sequence the first period values are preferred whereas in the sal/for sequence the second period values are preferred is suggestive of a preference for formoterol over salbutamol.

The association in the table may be examined using Fisher's exact test or a large sample approximation such as the chi-square test. The similarity to the *period adjusted sign test* considered in Section 4.3.6 should be noted. The difference, as for the difference between the sign test and McNemar's test, is that

Table 4.13 (Example 4.3) Classification of binary outcomes.

Sequence	Prefer first period	No preference	Prefer second period	Total all preferences
for/sal	9 d	3 e	0 f	12 r_1
sal/for	1 g	5 h	6 k	12 r_2
total both sequences	10 s_1	8 s_2	6 s_3	24 N

Table 4.14 (Example 4.3) Classification of binary outcomes by period preference.

	Preference		
Sequence	First period	Second period	Total
for/sal	9 d	0 f	9 $d+f$
sal/for	1 g	6 k	7 $g+k$
total both sequences	10 $d+g$	6 $f+k$	16 $d+f+g+k$

the period adjusted sign test, being based on a reduction of continuous data to preference data, will not (in theory) have to discard any ties. (In practice, of course, because of finite precision of measurement, there may be ties.) The Mainland–Gart test, requiring a reduction of binary outcomes to preferences, is likely to require the discarding of a number of tied values.

If we apply the chi-square approximation in this example, making a correction for continuity, then making the necessary changes of symbol to (4.9) we obtain

$$\frac{M\{|dk - fg| - M/2\}^2}{(d + g)(f + k)(d + f)(g + k)},$$

where $M = d + f + g + k$.

This yields a value of 8.96 which must be referred to a chi-square table with one degree of freedom. Alternatively we may refer the square root to tables of the standard Normal and double the corresponding tail probability. In this case the P value is 0.003.

4.4.3 Allowing for period: Prescott's test

Prescott (1981) proposed an alternative to the Mainland—Gart test which does not discard the information from patients with tied preferences.

To illustrate the test we return to Table 4.11. We condition on all the margins since we regard the number of patients allocated to each sequence, $r_1 = d + e + f$ and $r_2 = g + h + k$ as uninformative (it may, indeed, have been fixed) and we wish to look for treatment effects in the presence of a possible general trend which, for the experiment as a whole, is reflected in the three column margin values: $s_1 = d + g$, $s_2 = e + h$ and $s_3 = f + k$. Rather in the same way, however, as for continuous measures, the treatment effect can be defined in terms of a difference in period effects between sequences. Now, for the first sequence (for/sal) the difference $d - f$ (9 in this case) is a measure of the extent to which period 1 is preferred to period 2, whereas for the second sequence the corresponding difference is $g - k$ ($= -5$ here). The difference between these two figures $d - f - g + k$ ($= 14$ here) is simply the excess of preferences for one treatment (formoterol) over another (salbutamol).

Given, however, that the margins are known, then once we know what $d - f$ is we know what $g - k$ is since, $g - k = s_1 - s_3 - (d - f)$. Hence we only need to consider the difference $d - f$ when looking for evidence of treatment effects. Prescott's test then consists of examining all possible contingency tables with values of $T = d - f$ equal to or exceeding that observed. In this particular case $T = 9$ and the three possible tables are:

$$\begin{array}{ccc} 9 & 3 & 0 \\ 1 & 5 & 6 \end{array} \qquad \begin{array}{ccc} 10 & 1 & 1 \\ 0 & 7 & 5 \end{array} \qquad \begin{array}{ccc} 10 & 2 & 0 \\ 0 & 6 & 6 \end{array}$$

Thus

$$\Pr(T \geqslant 9) = \left\{ \binom{10}{9}\binom{8}{3}\binom{6}{0} + \binom{10}{10}\binom{8}{1}\binom{6}{1} + \binom{10}{10}\binom{8}{2}\binom{6}{0} \right\} \Big/ \binom{24}{12}$$
$$= 0.00024.$$

This is the exact probability calculation. For a two-tailed test we double the value, so we have $P = 0.0005$.

This sort of analysis can be applied very easily via the Jonckheere–Terpstra test in StatXact®. We use the fact that the outcomes ‘Prefer first period’, ‘No preference’ and ‘Prefer second period’ are ways of classifying a temporal trend. They are thus just the sort of data to which a trend test may be applied. If a data set with two rows and three columns corresponding to the frequencies of Table 4.13 is entered, application of the command *KI JT/EX*, where *KI* stands for *K independent samples JT* stands for *Jonckheere–Terpstra* and */EX* requests an exact test, yields the result calculated above. Alternatively, the pull-down menus will also perform the analysis. (The global parameters of the program may also have to be set via the pull-down menu under ‘options’ to deliver the number of decimal places we have used in calculation above.) Alternatively, we can use the Wilcoxon–Mann–Whitney test, and applying this with StatXact® gives the same result.

Prescott (1981) also provides the expected value and variance of his statistic. These are respectively

$$E(T) = r_1(s_1 - s_3)/N$$

and

$$\mathrm{var}(T) = r_1 r_2 \left\{ (s_1 + s_3) - (s_1 - s_3)^2/N \right\} / \{(N(N-1)\}.$$

In this particular case we have $E(T) = 2$ and $(T) = 4$, from which the standard error $SE(T) = 2$. Using a Normal approximation and calculating $z = \{|T - E(T)| - 1/2\}/SE(T)$ we obtain $z = 3.25$, which corresponds to a two-tailed probability of 0.001.

Farewell (1985) provides a useful discussion of binary data from cross-over trials and is particularly illuminating on the relationship between the Mainland–Gart (Gart, 1969; Mainland, 1963) approach and that of Prescott (1981). A mixed model approach defined in terms of marginal probabilities is used by Fidler (1984) and the relationship to Prescott’s test is discussed by Fidler (1986).

4.4.4 Ezzet and Whitehead's random effect approach*

Ezzet and Whitehead (1992) have proposed a random effects version of logistic regression. Suppose that

$$
\begin{aligned}
(\eta_{ij}) &= \log\left(\frac{p_{ij}}{1-p_{ij}}\right),\\
p_{ij} &= \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})},
\end{aligned}
$$

where p_{ij} is the probability of a 'success' for patient i in period j. Although p ranges from 0 to 1, η ranges on the real line and may be modelled using regression techniques. This is, of course, one of the central ideas behind logistic regression. We then model

$$\eta_{ij} = \mu_{ij} + u_i,$$

where μ_{ij} is a fixed effect predictor reflecting the fixed effects (for example, period and treatment effects) applying to patient i in period j and u_i is a random effect due to patient i. In turn we model u as a random normal variate with expectation zero and variance to be determined.

This conceptually simple and attractive model is extremely awkward to fit, involving as it does numerical optimization of a likelihood which is itself a function of integrals that have to be numerically evaluated. Although this model (Ezzet and Whitehead, 1992) was referred to in the first edition of this book, it was not illustrated because it did not at that stage form a practical option for analysis for any but the most skilful and dedicated. However, `proc nlmixed` of SAS® is a powerful procedure which has been developed in the meantime and which permits this model (and many others) to be fitted relatively easily. Although Example 4.3 is not ideal to illustrate this approach since the data set is rather small, we shall nevertheless illustrate how it may be analysed with SAS®.

We suppose that the data have been read into the program in an SAS® datastep and consist of the patient number, *PATIENT*, a dummy variable for treatment, *TREAT*, coded 0 or 1, a similar dummy variable for *PERIOD*, and a variable *BINEFF* coded 0 if the response was 3 or less and 1 if the response was 4. (In other words, 0 if the response was coded − in Table 4.12 and 1 if it was coded +.) The data are to be arranged so that there is one line for each measured outcome (two for each patient).

We may then carry out the analysis using the following code:

```
proc sort;
  by PATIENT;
run;
```

```
proc nlmixed data=one;
  parms mu=0 tau=0.0 pi=0 s2u=2;
  bounds s2u>0;
  pred=mu + tau* TREAT + pi* PERIOD +u;
  p=exp(pred)/(1+exp(pred));
  model BINEFF~binomial(1,p);
  random u~normal(0,s2u) subject=PATIENT;
run;
```

We start by looking at the second to last line of this code. This informs *proc nlmixed* what term provides the random effect (using the keyword *subject*) and how this effect is to be modelled. Here we declare *PATIENT* to be the random effect. Note that it is necessary to make sure that the data are sorted by patient order since `proc nlmixed` assumes that every time a new value of the *subject* variable is encountered a new subject is being analysed.

We now look at the code line by line from the second line of `proc nlmixed`. The `parms` statement sets initial values for the parameters `mu`, the general intercept, `tau`, the treatment effect `pi`, the period effect and `s2u` the variance of the random effect. (An alternative parameterization of this variance will be considered in Chapter 6.) The *bounds* statement constrains this variance to be positive. The linear predictor `pred` is then expressed as a function of fixed period, treatment and random subject effects. The probability, `p`, is then expressed as a function of the linear predictor using the inverse link function. Finally, `BINEFF` is declared as a binomial variate with probability `p` and $n = 1$.

The output from the program includes the following:

```
                              Parameter Estimates

                       Standard
Parameter   Estimate     Error    DF   t Value   Pr > |t|   Alpha     Lower      Upper

beta0         3.1233    0.9504    23      3.29    0.0032     0.05    1.1573     5.0893
tau          −3.2884    0.9114    23     −3.61    0.0015     0.05   −5.1738    −1.4029
pi           −1.1729    0.8054    23     −1.46    0.1588     0.05   −2.8390     0.4932
s2u             1E-8         .    23         .         .     0.05         .          .
```

Bearing in mind that salbutamol has been coded 1 and formoterol has been coded zero, the value of −3.29 for *tau* suggests an odds ratio of exp(3.29) = 26.8 in favour of formoterol. The columns headed 'Lower' and 'Upper' give 95% confidence intervals for the parameters.

If we wish to carry out an analysis of deviance, then we need to fit a model with the period effect only and look at the change in deviance when then fitting the treatment effect as well. In the first case (fitting period only) SAS® gives a deviance (labelled '−2 Log-Likelihood') of 62.1. In the second case (fitting treatment also) the deviance is 42.0. We thus have a difference of

20.1 associated with the one degree of freedom for treatment, which, when compared to tables of the chi-square distribution, is clearly highly significant.

Remark The estimate of the random-effect variance is at the lower bound of zero in this example. As we stated before, this example is not ideal for illustrating this method of analysis. The reason is that the binary values are actually slightly negatively correlated, a phenomenon which is not accommodated by the form of conditional model used by Ezzet and Whitehead (1992). In my opinion, this negative correlation is a purely chance phenomenon here.

4.4.5 Other approaches*

In an extremely important paper, Kenward and Jones (1987b) develop a log-linear model for binary outcomes from cross-over trials (see also Jones and Kenward, 1989, Chapter 3). (We illustrate this analysis, but without further explanation, using GenStat® in Appendix 4.1.) The modelling is in terms of the original responses (0, 1) and allows for period as well as treatment effects. The carry-over effect may also be included but if this is done the same problem surfaces as with continuous data. Jones and Kenward (1989) discuss a number of ways in which the between/within-subject distinction (which is the essential feature of a cross-over trial) may be reflected in the model. The Mainland–Gart test, for example, may be obtained by allowing a fixed effect for each subject and conditioning on the subject totals (i.e. the sum of the binary responses for the two periods for the given subject) (Gart, 1969). Alternatively, dependency may be introduced, using a bivariate logistic model, and Jones and Kenward (1989) develop this approach in great detail. There is no space here to discuss this work but the references given are essential reading for anyone who is interested in this subject.

I must issue one particular warning, however, and that is that the bivariate logistic model in its most general form incorporates more parameters (a carry-over parameter and a group by dependence parameter) than it will in practice be desirable to incorporate for the purpose of examining treatment effects. There are a number of tests for these additional parameters which may be used to determine whether these parameters may be dropped from the final model but to use an approach which relies on such tests courts the same dangers as the two-stage analysis of continuous data and I do not recommend it.

One further approach to analysing binary data will be considered. This consists of reducing the information from the two periods for each patient to a single score as was done for Prescott's test and then using logistic regression. This approach may also be used for ordered categorical data and will be considered in the section below that deals with such outcomes.

4.5 ORDERED CATEGORICAL DATA

The original responses for Example 4.3 given in Table 4.15 were on a four-point scale with 1 = poor, 2 = fair, 3 = moderate, and 4 = good. They were reduced to a binary outcome for the purpose of illustrating the various techniques covered in Section 4.4. Both the original outcomes and the binary outcomes derived from it are examples of *ordered categorical data*: the outcomes may be classified into categories but unlike many categorical classifications (e.g. marital status: never married, married, widowed, divorced, or religion: Christian, Jewish, Moslem, Sikh, Hindu, Buddhist etc.) there is an ordering which is important. We now consider a possible analysis of such data. The technique proposed, which is similar in spirit to a proposal of Hills and Armitage (1979), is rough and ready but easy to implement. Further discussion will be found in (Senn 1993b).

The first step in the technique consists of taking the two responses from the two periods and reducing them to another ordered categorical response for each patient. The way this may be done is illustrated in the table below. Thus, the original four-point scale is turned into a five-point score illustrating change over time. The categories 'better' and 'worse' might be further subdivided but this is not simple. We know that an improvement from poor to moderate exceeds that from fair to moderate or from poor to fair but we do not know that it exceeds one from moderate to good. Thus it is probably safest to accept the loss of information brought about by this grouping.

We now recode the data from Table 4.12 in terms of change scores and present it in the form of a contingency table (Table 4.16). The difference between the two sequence groups is an indication of the treatment effect and this contingency table may now be examined either using the *Armitage trend test* (Armitage, 1955) or *logistic regression for ordered categorical variables* or indeed any of the many other methods available for analysing ordered categorical data. We shall illustrate a possible approach using the *empirical logit transform* (McCullagh, 1980). Such an approach is perhaps not entirely suitable for a trial as small as this one but its use is illustrated here for didactic purposes.

Table 4.15. (Example 4.3) Possible categorization of patients by change in status from first to second period based on four original ordered categories.

		Period 2			
		Poor	Fair	Moderate	Good
	Poor	Same	Better	Better	Much better
Period 1	Fair	Worse	Same	Better	Better
	Moderate	Worse	Worse	Same	Better
	Good	Much Worse	Worse	Worse	Same

Table 4.16. (Example 4.3) Frequency distribution of patients cross-classified by sequence and change score.

	Change score				
Sequence	Much worse	Worse	Same	Better	Much better
for/sal	3	7	2	0	0
sal/for	0	1	5	6	0
both	3	8	7	6	0

First, we consider the four possible dichotomizations of the five-point change score: much worse versus the rest, much worse or worse versus the rest etc. In the first sequence these produce divisions of patients as follows: 3:9, 10:2, 12:0, 12:0. The presence of zeros can make the following steps awkward and it is best to add a half to each category of the dichotomy. Of the original categories, however, the much better category, giving no values for either sequence group, is probably better abandoned altogether, thus leaving us with a four-point scale. Carrying out these steps for both sequence groups we obtain

for/sal	3.5 : 9.5,	10.5 : 2.5,	12.5 : 0.5,
sal/for	0.5 : 12.5,	1.5 : 11.5,	6.5 : 6.5.

We now take logs of the ratios corresponding to the divisions to obtain the following *logits*:

for/sal	– 0.999	1.435	3.219
sal/for	– 3.219	– 2.037	0.000
differences	2.220	3.472	3.219

Thus, $-0.999 = \log(3.5/9.5)$ etc.

We also calculate the differences and these provide treatment estimates on the logit scale corresponding to each of the three possible dichotomizations of the (reduced) four-point change score. We already encountered a logit type transformation when considering VAS scores in Section 4.2.2, where we noted that its advantage is that it transforms values which are measured on a bounded scale into ones which are measured on the ordinary linear scale. Since the original change score is defined in terms of period differences the differences between the logits associated with the change score correspond to a treatment effect. Thus we have obtained three possible estimates in terms of logits. The next step is to combine them to form one estimate.

Obviously a very simple solution would be to obtain the arithmetic mean of all three. The variances, however, of the three estimates are not the same and a superior solution is to use the so-called *general empirical logit transform weights* (McCullagh, 1980). These are obtained as proportional to the product of the totals associated with each of the dichotomized categories and the sum of the totals of the categories on either side of the 'cut'. Thus, we have

$$\begin{aligned} w_1 &\propto 3 \times 21 \times (3 + 8) = 693 \\ w_2 &\propto 11 \times 13 \times (8 + 7) = 2145 \\ w_3 &\propto 18 \times 6 \times (7 + 6) = 1404 \end{aligned}$$

which when scaled to add to 1 gives $w_1 = 0.163$, $w_2 = 0.506$, $w_3 = 0.331$. Applying these to the three treatment estimates obtained we have

$$\hat{\tau} = 0.163 \times 2.220 + 0.506 \times 3.472 + 0.331 \times 3.219 = 3.184.$$

The variance of this estimate may be calculated according to the formula (McCullagh, 1980):

$$1/\mathrm{var}(\hat{\tau}) = n_1 n_2 (1/3 - \textstyle\sum \pi_j^3/3)/n, \tag{4.13}$$

where n_1 and n_2 are the numbers in each sequence group and $n = n_1 + n_2$ and π_j is the proportion of patients in the jth change score category. Thus for this example we have

$$n_1 = 12, \quad n_2 = 12, \quad n = 24 \quad \text{and} \quad \begin{aligned} &\pi_1 = 3/24, \quad \pi_2 = 8/24, \\ &\pi_3 = 7/24, \quad \pi_4 = 6/24 \end{aligned}$$

yielding $\mathrm{var}(\hat{\tau}) = 0.5431$ and $SE(\hat{\tau}) = 0.74$. Although according to McCullagh (1981) the estimate of the variance of $\hat{\tau}$ can have serious bias if $|\hat{\tau}| > 1$ it is clear that the treatment effect is highly significant.

The labour involved in this analysis is not excessive and it is probably always worth undertaking this work as a rough check. In practice, however, a more satisfactory answer is obtained by estimating the treatment parameter via maximum likelihood: by carrying out, in fact, a logistic regression for ordered categorical variables. This may be done quite simply using *proc logistic* in SAS®. It is necessary to create first of all, a data set consisting of two variables *CHANGE* (say) recording the change score for each patient and *GROUP* (say) recording as a dichotomy the sequence group to which he belongs with 0 as group 1 (say) and 1 as group 2. Then this simple code:

```
proc logistic;
  model CHANGE = GROUP;
run;
```

will carry out the desired analysis.

Applying this to this example we obtain the following estimates: $\hat{\tau} = -4.2517$ with est $SE(\hat{\tau}) = 1.2976$. The negative sign is due to our having used 1 for group 1 and 2 for group 2 and the particular formulation of the logistic model within SAS®. We need to reverse it to make it comparable to the empirical logistic transform estimate. The agreement between the two estimates is not close and the same is true of the estimate of the standard error. Nevertheless, both analyses indicate a strong treatment effect. The example, having small numbers and categories with zero counts, is not particularly suitable for producing close agreement in methods of estimation.

4.5.1 Discussion of the logistic regression procedure

It should be noted that there is a connection between this procedure and Prescott's (1981) test since the logistic regression procedure outlined above may also be applied to Prescott's score. There a binary outcome is reduced to a three-point change score. As soon as three or more categories are noted a five point change score may be obtained. If there are exactly three original categories then no partial ordering within the five change score categories can be established. If there are four or more original categories then a partial ordering with the second and the fourth change score categories can be observed and is ignored in the procedure outlined above. I suspect that in such cases the loss of information is slight.

If there are a large number of original categories then the number of patients in the most extreme change score categories may be small (or even zero). It may then be better from the practical point of view to collapse the original scores into a small number of categories (three will still provide five change scores). The disadvantage of this is that the collapsing will always involve some degree of judgement and an assumption which goes beyond that which is strictly justifiable for the ordered categorical problem. If, for example, we had collapsed the 'poor' and 'fair' categories in Example 4.3 then a change from 'fair' to 'good' is classified as 'much better' and therefore more important than a change from 'poor' to 'moderate'. Previously, these two changes were included together.

4.5.2 An alternative approach using the original categories*

Ezzet and Whitehead (1991) describe an approach to analysis that uses a modification of logistic regression applied to the original scores. This allows dependence between the first period and second period values by using a random effect for patients. Obviously, since binary outcomes are a special case of ordered categorical outcomes, the method can also be applied to dichotomies, and this is specifically considered by the authors in another paper (Ezzet and Whitehead, 1992). This is a theoretically attractive approach which has the added advantage that it generalizes to design with more than two periods but suffers from the practical difficulty, common in random effect modelling of non-Normal outcomes, that the log-likelihood is then the sum of a number of integrals. Finding the maximum likelihood solution then involves numerical methods. As far as I am aware, this approach is not readily available in any standard software packages and, until it is, it will remain awkward to implement. Since the analogous approach for binary data can be implemented using *proc nlmixed* of SAS®, it is possible that it may be capable of implementation using that procedure.

An alternative would be to adapt the continuation-ratio approach of Lindsey *et al.* (1997). In that paper, the authors commented that dependence amongst observations on the same patient would be difficult to handle, but this should be feasible using *proc nlmixed* of SAS®.

4.6 FREQUENCY DATA

Although cross-over trials are most often applied to study the effect of treatments on continuous data that arise as a result of measurement, occasionally they are used to study discrete data that arise as a result of counting. For example, such data arise where the numbers of seizures are counted in a trial in epilepsy or the number of asthma attacks or exacerbations are counted in a trial in asthma. A natural candidate distribution for modelling such data is the Poisson distribution, and indeed Poisson regression is sometimes used in parallel group trials where such data arise. It might be thought that the correlation between frequencies measured repeatedly on the same patient in a cross-over trial would make Poisson regression difficult to apply. In fact its application is simple and there is the added bonus that the conditions likely to ensure its validity are more likely to apply in a cross-over trial than in a parallel group trial. Before discussing this, however, we divert to consider an example.

Example 4.4 A double-blind cross-over trial in asthma compared the effect of the long-acting beta-agonist salmeterol (50 μg twice daily) to placebo (Wilding *et al.*, 1997). One hundred and one patients were treated for two six-month periods, having been randomized to one of two sequences: either placebo followed by salmeterol or salmeterol followed by placebo. There was a one-month wash-out between treatment periods. A number of different measures were compared. We concentrate here on the number of exacerbations of asthma. Fifteen patients did not complete the trial. A summary of the results for the 86 patients who did complete the trial is given in Table 4.17.

Remark Hitherto our examples have involved fewer subjects. They have been so-called single-dose pharmacodynamic studies and common of the sort carried out in phase II in drug development. This, however, is a phase IV trial and involves studying patients over an extended period in which they are given regular therapy. The risk of drop-out is greater in such trials, and this is reflected in the number of patients discontinuing in this study. This problem will be ignored in the analysis that follows.

Table 4.17 shows in compressed form all the data that are necessary to construct a listing, patient, by patient of the original data (provided the individual identities are irrelevant), which is usually what is needed to carry out an analysis of a cross-over trial. As it turns out, in the case of Poisson regression this is not necessary and an even more compressed summary is all that is required. It is, in fact, sufficient to list the totals by each sequence under each treatment (or equivalently under each period, since for a given sequence the period determines the treatment and vice versa). This has been done in Table 4.18, which obviously can be constructed from Table 4.17.

The reason why this is all that is necessary is as follows. Consider patient *i* in the sequence placebo/salmeterol, which we will call sequence I. Suppose that this patient's exacerbations of asthma in period 1 follow a Poisson distribution with mean μ_i. Now suppose that the effect of treating the patient in period 2 is to

Table 4.17 (Example 4.4) Exacerbations of asthma in a cross-over comparing salmeterol 50 mg b.i.d. with placebo. Number of patients classified by sequence and number of exacerbations under both treatments.

		Salmeterol				
	Placebo	0	1	2	3	Total
Placebo/Salmeterol	0	27	3	0	0	30
	1	9	1	0	0	9
	2	0	0	1	0	1
	3	1	0	0	0	1
	Total	37	4	1	0	41
Salmeterol/ Placebo	0	24	7	0	0	31
	1	10	0	0	1	11
	2	0	0	1	0	1
	4	0	1	0	0	1
	6	0	1	0	0	1
	Total	34	9	1	1	45

Table 4.18. (Example 4.4) Total number of exacerbations cross-classified by sequence and treatment.

Sequence	Placebo	Salmeterol	Total
Placebo/salmeterol	14	6	20
Salmeterol/placebo	23	14	37
Total	37	20	57

multiply this mean by a factor π for the change of period and a further factor τ for the effect of salmeterol compared to placebo. The expected number of exacerbations for this patient is now $\pi\tau\mu_i$. If period and treatment effects do not vary from patient to patient, then analogous expressions apply for every patient in this sequence. However, the sum of independent Poisson variables is itself Poisson. If we now let the expectation of this Poisson sum of all exacerbations in sequence I in period 1 be μ_{I1} we have $\mu_{I1} = \sum \mu_i$, where the summation is over all patients in this sequence. If we now consider the corresponding expectation in period 2 for this sequence, we have $\mu_{I2} = \sum \pi\tau\mu_i = \pi\tau\sum \mu_i = \pi\tau\mu_{I1}$. We shall not reproduce the argument here, but it turns out that the maximum likelihood estimator from this sequence, for the product $\pi\tau$, is simply the ratio of the observed second sequence total exacerbations to the first sequence total.

If we now turn to sequence II, by a similar argument we can show that the expectation in the first period (under salmeterol), μ_{II1}, expressed as a function of the second period expectation, μ_{II2}, is $(\tau/\pi)\mu_{II2}$. (The reason why the period effect now appears as a divisor is that we are comparing the first period to the second.) The maximum likelihood estimator of the ratio is τ/π simply the ratio of the observed first period total (under salmeterol) to the second (under placebo).

To obtain an estimate of τ we form the product of these two ratios (whereby the effect of π cancels out) and take the square root. In our example we thus have

$$\hat{\tau} = \sqrt{(6/14) \times (14/23)} = 0.51.$$

This is our point estimate of the treatment effect of salmeterol, namely that it reduces exacerbations to approximately one-half.

For the purpose of calculating standard errors of treatment estimates, it is better to work on the log scale. This is the scale for the usual 'link function' for Poisson regression which is, of course, a special case of generalized linear models (McCullagh and Nelder, 1989). Our estimate of $\log(\hat{\tau})$ is simply the average of the differences from each sequence between the logarithms of the totals under salmeterol and those under placebo. In this case we have

$$\begin{aligned}\log(\hat{\tau}) &= \log\left(\sqrt{(6/14)(14/23)}\right)\\ &= [\{\log(6) - \log(14)\} + \{\log(14) - \log(23)\}]/2 = -0.672.\end{aligned}$$

We may now use the fact that the variance of the logarithm of a Poisson variable is approximately the reciprocal of its expectation. For reasons that are not immediately obvious but which will be explained below, the four totals may be treated as if independent, so that we simply add these variances and divide by 4 (to reflect the fact that we have divided the sum of the within-sequence differences by 2). We thus obtain

$$\begin{aligned}\mathrm{var}\{\log(\hat{\tau})\} &\approx \frac{1}{4}\left(\frac{1}{6}+\frac{1}{14}+\frac{1}{14}+\frac{1}{23}\right) = 0.088,\\ SE\{\log(\hat{\tau})\} &\approx \sqrt{0.088} \approx 0.297.\end{aligned}$$

There are several features of the argument above that require comment. First, one red herring needs to be avoided. It is well known that the mixture of a number of Poisson variables is not itself a Poisson. For example, if the proneness or 'frailty' of the patients to have exacerbations follows a gamma distribution over all patients, then the number of exacerbations per patient follows a negative binomial distribution. This distribution has a higher variance than the Poisson. However, in a cross-over trial each patient acts as his or her own control. The between-patient element is thus fixed over all randomizations and thus contributes nothing to the overall variability of the treatment estimate.

Remark Note that in a parallel group trial, such between-patient variability could not be ignored. Since, in practice, differences between patients will almost always obtain, it is almost inevitable that such trials will show 'extra-Poisson' variation. With a cross-over trial, such variation between patients provides no problem, although if the patients themselves exhibit extra-Poisson variability over time, this feature *is* a problem.

It remains to be explained why the four Poisson sums may be treated as if they were independent. To show this, consider the simpler case where we assume that there is no period effect. We may then use a simplified form of the estimator above where we just use the totals for each treatment, ignoring the sequence. We thus obtain

$$\begin{aligned}\hat{\tau} &= 20/37 = 0.54,\\ \log(\hat{\tau}) &= \log(20/37) = -0.615,\\ \mathrm{var}(\hat{\tau}) &= \frac{1}{20}+\frac{1}{37} = 0.077,\\ SE(\hat{\tau}) &= 0.278.\end{aligned}$$

The reason why the variance of the difference is simply the sum of the variances may be explained quite simply. Consider the case where the active treatment is identical to placebo. We have, by virtue of having fixed the patients, fixed the process. This is now no different, in principle, from any stable process. Provided that the patients themselves do not have varying proneness over time, we will then have a Poisson process. If we now consider the effect of treatment, then, provided there is no patient by treatment interaction, this will simply be to move the patients from one Poisson to another.

Remark It is, of course, a strong assumption to assume that there is no patient by treatment interaction. However, since under the null hypothesis the active treatment is a placebo, this assumption belongs to the hypothesis being tested and does not cause a problem for significance tests. It is, however, more important in the construction of confidence intervals. On the other hand, the assumption of stability over time for the patients is important for significance tests. My advice when using a Poisson model is to back this up with some other way of looking at the data, perhaps a rank test.

A further approach for modelling frequency data would be to use a non-linear mixed model and analyse this using proc nlmixed of SAS®. Other possibilities are discussed in Longford (1998) and in the important paper by Patefield (2000), which also discusses the analysis of binary data.

4.7 'SURVIVAL' DATA*

The field of survival analysis is an extremely important one in clinical trials and one which has attracted considerable attention in recent years. Where survival has to do with mortality, the topic clearly has no relevance to cross-over trials. It is not uncommon, however, in certain fields to test patients' state of health by subjecting them to a *provocation* or *challenge*. Sometimes such provocations may be fixed and the response is then recorded. An alternative approach may be to give the patient an increasing 'dose' of the provocation until a particular response is observed. This is a form of *titration* and in such cases the outcome variable becomes the 'dose' giving a particular response. For example, in asthma trials it is not uncommon to give the patient increasing *concentrations* of methacholine to breathe until a 20% reduction in forced expiratory volume in 1 second (FEV_1) is observed (Senn, 1989a). Such a concentration of methacholine is then known as the PC_{20}. Similarly, patients suffering from angina pectoris may be asked to undertake exercise until the patients' electrocardiogram (ECG) is affected in some predefined way. The time to this event or the total exercise workload then becomes the outcome. Such measures may be censored in that the test may have to be stopped for other reasons before the defined response is seen. In such cases all that can be said about the 'true' response is that it is at least as high as that recorded.

The results of any trial run in this way may consist of a mixture of censored and uncensored results. Such data may be suitably analysed using survival analysis. Trials of this nature may frequently be run as cross-overs. In fact all of the three examples introduced so far in this chapter come from provocation trials. Furthermore, Example 4.1 (Leuenberger and Gebre–Michel, 1989) was a trial in which the provocation was titrated, although we did not, when describing it in Section 4.2.2, consider the analysis of a survival-type measure.

Example 4.2 also involved a titration. Thus there is a role for survival analysis of cross-over studies.

4.7.1 A method based on 'preferences'

A paper by France *et al.* (1991) describes how such analyses may be undertaken. Starting with a model commonly employed in survival analysis, the proportional hazards model (Cox, 1972), they show how survival data may be analysed in terms of *patient 'preferences'*. A patient preference for treatment *A* is recorded if he 'survived' longer under *A* than under *B*. A preference is recorded for *B* if he 'survived' longer with *B*. A tie will occur where both measurements are censored. In other cases it should usually be possible to record a preference.

Survival analysis is a topic which has itself been the subject of several books and the details of the arguments of France *et al.* (1991) are beyond the scope of this book. The reader who is interested should consult their paper. We shall limit ourselves to illustrating some simple features of their method using an example.

Example 4.5 For the-trial reported in Example 4.1 the main outcome variable was PD_{20}, defined as the cumulative *dose* of methacholine required to produce a 20% fall in FEV_1. The data are given in Table 4.19 and are partially reported in Leuenberger and Gebre–Michel (1989) who also discuss the trial in 'survival' terms. The methacholine was delivered in discrete steps. Since in such a trial the highest dose given to the patient will not produce a fall of exactly 20%, the dose

Table 4.19 (Example 4.5) PD_{20} readings for a methacholine challenge in a trial comparing salbutamol 200 μg (sal) and formoterol 12 μg (for).

		$PD_{20}(\mu\text{mol})$					
Sequence	Patient number	Formoterol	Salbutamol	Basic est.	Period diff.	Treat pref.	Period pref.
for/sal	1	7.29	6.68	0.61	0.61	for	1
	3	7.84	2.42	5.42	5.42	for	1
	5	4.65	1.93	2.72	2.72	for	1
	6	12.40c	9.47	> 2.93	> 2.93	for	1
	10	10.57	8.96	1.61	1.61	for	1
	12	8.17	7.60	0.57	0.57	for	1
sal/for	2	5.06	2.33	2.73	−2.73	for	2
	4						
	7	12.40c	10.33	> 2.07	< -2.07	for	2
	8						
	9	7.33	6.97	0.36	−0.36	for	2
	11	11.11	10.80	−0.69	0.69	sal	1

which would produce such a fall has been estimated by interpolation using the last two doses. For two patients (patients 4 and 8) problems in recording the data mean that the values are missing. (This point will not be discussed further and the data will be treated as if they came from a trial in 10 patients.) For two other patients (patients 6 and 7) on one of the two treatment days the PD_{20} recorded is the actual highest possible dose delivered during the manoeuvre. These values may be regarded as being censored and are marked 'c'.

France *et al.* (1991) show that a proportional hazards model applied to survival data of this sort leads to an analysis in terms of 'preferences'. Note that we are able to establish a 'preference' for patients 6 and 7 because we are able to establish minimum values for the basic estimators. Using the period preference data we can now reduce Table 4.19 to a summary of the preferences as given in Table 4.20.

It will be seen that this is a simple contingency table of the sort we considered for the period adjusted sign test of Section 4.3.6 and indeed, since the period preference is simply determined by the sign of the period difference, we may analyse this table using the methods of that section. For example, the reader may check for himself using (4.8), that Fisher's exact test (Fisher, 1990a; p. 96, Sprent, 1989, p. 172) yields a two-sided P value of 0.067.

This is, however, not the only possible analysis. France *et al.* (1991) show that the maximum likelihood estimator of the *hazard ratio* is simply the geometric mean of the ratios of the *treatment* preferences within each sequence, namely $\{(a/b)(d/c)\}^{\frac{1}{2}}$. Note that if we take the log of the estimated hazard ratio the result may be expressed as

$$\text{estimated log hazard ratio} = \tfrac{1}{2}\{\log(a/b) - \log(c/d)\}. \tag{4.14}$$

This is extremely interesting because Table 4.20 could equally well be expressed in terms of ordered categorical variables. Suppose we rate the period differences (2 − 1) using a change score of 'better' or 'worse', then, although the differences in PD_{20} are not categorical, we have reduced them to a categorical change score, so we have a table of the form overleaf.

Table 4.20 (Example 4.5) Patients cross-classified by treatment sequence and period preference.

	Period preferred		
Sequence	1	2	Total
for/sal	6(a)	0(b)	6
sal/for	1(c)	3(d)	4
Total	7	3	10

	Worse	Better
sal/for	6 a	0 b
for/sal	1 c	3 d

But this is just a simple example of the sort of table we considered in Section 4.5.1. We have an ordered categorical variable with a single cut point. If we were to calculate the empirical logistic transform then (ignoring for the moment the addition of $\frac{1}{2}$ to each of the frequencies) we should obtain $\log(a/b) - \log(c/d)$. But for the factor $\frac{1}{2}$, which in any case reflects a particular parameterization, this is the same as that given by (4.14).

In this example, (4.14), is undefined, as is the estimated hazard ratio, because b is 0. If we use the same general approach of adding a half to each of the frequencies we obtain a value of 1.706 for (4.14). Because this statistic is $\frac{1}{2}$ of the empirical logistic transform its variance will be $\frac{1}{4}$ of that given by applying (4.13). Using the notation of that formula we have $n_1 = 6$, $n_2 = 4$, $n = 10$, $\pi_1 = \frac{7}{10}$, $\pi_2 = \frac{3}{10}$. Substituting in (4.13) we obtain a value of 1.984 which, on division by 4, yields a variance of 0.496. Hence the standard error of (4.14) is 0.70. Thus we have an estimated log hazard ratio of 1.706 with a standard error of 0.70. (But note that this example is simply included for illustration and that the numbers involved are not really large enough to justify this sort of analysis.)

France *et al.* (1991) give a rather different development in terms of the hazard ratio itself and provide a formula for its variance as well as many other results of interest, including some rather controversial advice regarding the estimation of treatment effects in terms of the original survival variable.

4.7.2 An adaptation of the Wilcoxon test

A disadvantage of the approach of France *et al.* (1991) is that it makes no use of the magnitude of the differences between one period and another, merely noting a preference. Feingold and Gillespie (1996) consider how this extra information may be recovered. Their approach is very similar to that of Koch (1972) discussed above. In fact, they calculate semi-period differences for each subject. (As regards testing for the effect, however, it makes no difference if period effects are used.) If no observations were censored, the results could then be analysed using the Wilcoxon–Mann–Whitney procedure. However, the presence of censored observations means that in such cases we will not be able to establish what these differences are. This requires an adjustment to the procedure as follows.

Consider Example 4.5 again. The column of Table 4.19 headed ‘Period diff.’ has the relevant information for differences (period 1 − period 2). For patient 6, however, we know only that the difference is at least 2.93, since the first period value was censored at 12.40. For patient 7, on the other hand, we know only that the difference is no more than −2.07, since the second period value was censored at 12.40. Feingold and Gillespie (1996) refer to such period differences as 'right censored' (R) and 'left censored' (L), respectively. Of course, if both values are censored we cannot say anything about the difference. From a given difference, d_i, we compute a score U_i, which is the number of differences clearly smaller than d_i minus the number clearly greater than d_i. For this example, the differences in order are:

−2.73, − 2.07L, − 0.36, 0.57, 0.61, 0.69, 1.61, 2.72, 2.93R, 5.42.

Hence, the scores are −8, −8, −5, −3, −1, 1, 3, 5, 8, 8.

Table 4.21 lists differences and scores by patient and sequence group. Taking the smaller of the two sequence groups, we have a total score of −15. We now use the familiar argument of assuming that there is no treatment effect. That being so, the distribution of the ranks of the scores amongst the two groups should be random and arbitrary. We use the standard significance test approach of calculating the probability of observing a total score as extreme as or more extreme than that observed. Clearly there are two lower total scores possible: that of −21, corresponding to individual scores −8, −8, −5 in the second sequence group, and −19, corresponding to −8, −8, −3. Thus three of the possible tables we would construct, given the four to six split, would yield a test statistic as extreme as or more extreme than that observed. However, there are 10!/(6!4!) = 210 tables in total. Hence the P value is 3/210 = 0.0143.

Table 4.21 (Example 4.5) Period differences and scores by patient and sequence.

Sequence	Patient number	Period difference	Score
for/sal	1	0.61	−1
	3	5.42	8
	5	2.72	5
	6	2.93R	8
	10	1.61	3
	12	0.57	−3
sal/for	2	−2.73	
	4	NA	
	7	−2.07L	−8
	8	NA	
	9	−0.26	−8
	11	0.69	1

4.7.3 Other approaches

Feingold and Gillespie (1996) also discuss various other approaches, including estimates and confidence intervals. Survival analysis is also considered by Lindsey *et al.* (1996). The reader is referred to those papers for details. The important book by Hougaard (2000) also covers correlated survival data, and some of the techniques covered in that book could be adapted for cross-over designs. However, we shall not discuss this issue further but instead now proceed to consider designs with more than two treatments.

4.8 FINAL REMARKS

There has been a considerable expansion of methods for dealing with non-Normal outcomes for the *AB/BA* design since the first edition of the book. Currently rapid development is taking place in the field of generalized linear models with random effect terms, with several different approaches being developed by different authors. For example, one very promising approach we have not covered, because software to carry it out is not yet readily available, is the hierarchical generalized linear model approach of Lee and Nelder (1996). It looks as if this will be an extremely useful approach for modelling frequency and binary outcomes from cross-over trials. The reader is warned to be prepared for important developments in this area.

APPENDIX 4.1 ANALYSIS WITH GENSTAT®

As in Chapter 3, code for carrying out a number of analyses presented in this chapter using GenStat® is given below.

The first example covers the various non-parametric analyses. The data that have been input are the semi-period differences. Since confidence intervals are not calculated, the period differences could be used just as well, but the semi-period differences are consistent with the way the data were input to StatPlus®. Basic estimators are calculated from period differences by multiplying by 2 and reversing the sign in one sequence.

The *FEXACT2X2* procedure is used to carry out Fisher's exact test, which is the way that the period adjusted sign test and Brown–Mood median test are performed. It is necessary to create tables to which the exact test can be applied first of all. It will sometimes happen, although this is not the case here, that one of the period differences will be zero. Similarly, if there are an odd number of patients altogether, then at least one of the period differences is equal to the median. Analysis in these two cases must be restricted to use the non-zero or non-median values only. The code has been written to deal with this case and employs the *RESTRICT* procedure of GenStat® to do this.

```
"Analysis of Example 4.1"
"Cross-over Trials in Clinical Research"
"Input data values"
"n1 is number of values in first sequence,"
"n2 is number in second"
"sequence, n is total"
"Sequence is factor of sequence labels"
"PerDiff is semi-period differences"
SCALAR n1,n2,n
READ n1,n2
6 6:
CALCULATE n=n1+n2
FACTOR[VALUES=#n1(1),#n2(2);LEVELS=2;LABELS=\
!t(forsal,salfor)] Sequence
VARIATE[nvalues=n] PerDiff, BasicEst, MedDiff
READ PerDiff
9.5 30 17 9 25.5 4 −14 0.5 −18 −9 −12 −24.5 :

"Begin calculations"
"Calculate basic estimators"
"and compare period differences to their median"
CALCULATE BasicEst= −2* PerDiff
& BasicEst$[1...n1] =2* PerDiff$[1...n1]
& MedDiff=PerDiff-median(PerDiff)

"Create table for period corrected sign test"
"NB RESTRICT used to keep non-zero differences only"
RESTRICT PerDiff;\
CONDITION=(PerDiff.LT.0).OR.(PerDiff.GT.0)
FACTOR[LABELS=!t(neg,pos)] signpd
"RHS of next line codes 1 if PerDiff negative, 2 if positive"
CALCULATE signpd=(PerDiff.LT.0)+2* (PerDiff.GT.0)
TABULATE[PRINT=*; CLASSIFICATION=Sequence,signpd;\
COUNTS=SignTab; MARGIN=no]
PRINT SignTab

"Create table for median test"
"NB RESTRICT used to keep non-zero differences only"
RESTRICT MedDiff;\
CONDITION=(MedDiff.LT.0).OR.(MedDiff.GT.0)
FACTOR[LABELS=!t(neg,pos)] medpd
"RHS of next line codes 1 if MedDiff negative, 2 if positive"
CALCULATE medpd=(MedDiff.LT.0)+2* (MedDiff.GT.0)
TABULATE[PRINT=*; CLASSIFICATION=Sequence,medpd;\
COUNTS=MedTab; MARGIN=no]
PRINT MedTab

"Begin tests"
PRINT 'Sign test for basic estimators'
SIGNTEST[NULL=0] BasicEst
PRINT 'Wilcoxon signed ranks test on basic estimators'
WILCOXON BasicEst
PRINT 'Period corrected sign test'
FEXACT2X2 SignTab
PRINT 'Brown-Mood median test'
FEXACT2X2 MedTab
PRINT 'Wilcoxon-Mann-Whitney test of period differences'
MANNWHITNEY[GROUPS=Sequence] PerDiff
VARIATE[nvalues=12] test
```

The second example covered is Example 4.3. For the sake of brevity details of data processing are omitted. First, we consider logistic regression applied to categorical change scores and fitting a period effect.

We assume that the pointer *CatFreq* contains *CF2*, *CF2*, *CF3*, *CF4*, which in turn hold frequencies for the four types of change score based on categorical outcomes. (Five were theoretically possible, but only four were represented in the data set.) *Seq* is a factor with two values.

The data look like this:

Seq	CF2	CF3	CF4	CF5
forsal	0	2	7	3
salfor	6	5	1	0

and the code is:

```
MODEL[DISTRIBUTION=multinomial; YRELATION=cumulative]\
CatFreq[]
TERMS Seq
FIT Seq
```

The output includes the following:

```
*** Estimates of parameters ***

                                                    antilog of
                  estimate      s.e.       t(*)     estimate

Cut-point 0/1     −4.28         1.27       −3.38    0.01385
Cut-point 1/2     −1.660        0.779      −2.13    0.1901
Cut-point 2/3      1.117        0.666       1.68    3.057
Seq salfor        −4.25         1.30       −3.28    0.01424
* MESSAGE: s.e.s are based on dispersion parameter with value 1
```

Parameters for factors are differences compared with the reference level:

Factor	Reference level
Seq	forsal

The relevant line of the output is that beginning `Seq salfor`, which gives the treatment estimate on the log-odds scale. If this is compared to the result we obtained with SAS® it will be seen to be the same.

GenStat® also has a special procedure *XOCATEGORIES* that can be used for analysing binary data from cross-over trials. This uses a technique developed by Jones and Kenward (1989, pp. 124–9). Code to perform this analysis applied to Example 4.3 is as follows:

```
FACTOR[nvalues=8;levels=2;labels=!t(‘AB’,‘BA’)]\
Sequence
READ Sequence; frepresentation=ordinal
1 1 1 1 2 2 2 2 :
FACTOR[nvalues=8;levels=2] BinEff1, BinEff2
READ BinEff1; frepresentation=ordinal
1 2 1 2 1 2 1 2 :
READ BinEff2; frepresentation=ordinal
1 1 2 2 1 1 2 2 :
VARIATE[nvalues=8] Freq
READ Freq
1 0 9 2 0 6 1 5 :
POINTER[VALUES=BinEff1,BinEff2] BinEff
PRINT Sequence, BinEff1, BinEff2, Freq
XOCATEGORIES[METHOD=loglinear] SEQUENCE=Sequence;\
RESULTS=BinEff; NUMBER=Freq; SAVE=Fsave
```

Here the first part of the code simply reads in the necessary data. Note that all vectors and factors used are of length 8 (corresponding to two sequences times two possible first period outcomes times two possible second period outcomes). The binary outcomes are stored in two variates *BinEff1* and *BinEff2*. A ‘pointer’ *BinEff* points to these variates and *Freq* records with what frequency the particular combination occurs. The procedure has three essential parameters. *SEQUENCE* must be set equal to the factor (in this case *Seq2*) storing the sequences, *RESULTS* must be set equal to the pointer (in this case *BinEff*) referring to the variates storing the results, and *NUMBER* must be set equal to the variate (in this case *Freq*) giving the frequency with which each of the eight possible combinations of sequence, first period and second period results occur.

The analysis of Example 4.5 can be carried out using the following code.

```
"Analysis of Example 4.5"
"Cross-over Trials in Clinical Research"
"Input data values"
FACTOR[nvalues=4;levels=2;labels=!t(‘PS’,‘SP’)]\
sequence
READ sequence; frepresentation=ordinal
1 1 2 2 :
FACTOR[nvalues=4;levels=2;labels=!t(‘P’,‘S’)] treat
READ treat; frepresentation=ordinal
1 2 2 1 :
FACTOR[nvalues=4;levels=2] period
READ period; frepresentation=ordinal
1 2 1 2 :
VARIATE[nvalues=4] exacerb
READ exacerb
14 6 14 23 :

"Analysis using Poisson regression"

MODEL[DISTRIBUTION=poisson] exacerb
TERMS[FACT=1] sequence+period+treat
FIT[PRINT=*;CONSTANT=estimate; FACT=1] sequence+period
ADD[PRINT=model,summary,estimates,deviance;\
CONSTANT=estimate; TPROB=yes; FACT=1] treat
```

Here the `MODEL` statement is used to specify that a Poisson regression is being carried out. A log link is assumed by GenStat® unless otherwise specified. The terms to be fitted are `sequence`, and `period` in the first instance. The term `treat` is deliberately added subsequently using `ADD` in order that an ‘analysis of deviance’ can be performed, that is to say, to check by how much the fit is improved by adding the `treat` term. The output includes the following:

*** Summary of analysis ***

	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	3	10.65	3.550	3.55	0.014
Residual	0	0.00	*		
Total	3	10.65	3.550		
Change	−1	−5.49	5.485	5.49	0.019

*** Estimates of parameters ***

	estimate	s.e.	t(*)	t pr.	antilog of estimate
Constant	2.639	0.267	9.87	<.001	14.00
sequence SP	0.672	0.297	2.26	0.024	1.958
period 2	−0.175	0.297	−0.59	0.555	0.8391
treat S	−0.672	0.297	−2.26	0.024	0.5108

The estimate (on the log scale) and standard error for the treatment effect can be seen to be exactly as we obtained by hand when adjusting for the period effect above. The significance or otherwise of the effect of treatment is better judged by looking at the *P* value of 0.019 associated with the analysis of deviance rather than the value of 0.024 associated with the *t* statistic.

APPENDIX 4.2 ANALYSIS WITH S-PLUS®

The code that follows is for the analysis of Example 4.1. As was the case with GenStat®, the data that have been input are the semi-period differences. Since confidence intervals are not calculated, the period differences could be used just as well, but the semi-period differences were used when inputting the data to StatPlus®. Basic estimators are calculated from period differences by multiplying by 2 and reversing the sign in one sequence.

The *fisher.test* function is used to carry out Fisher's exact test, which is the way that the period adjusted sign test and Brown–Mood median test are performed. It is first necessary to create tables to which the exact test can be applied. It will sometimes happen, although this is not the case here, that one of the period differences will be zero. Similarly, if there are an odd number of patients altogether, then at least one of the period differences is equal to the median. Analysis in these two cases must be restricted to use the non-zero or non-median values only. The code has been written to deal with this case. For example, the statement

```
medpd<-sign(perdiff-median(perdiff))
```

compares the period differences to their median and returns the value −1, 0 or 1 to *medpd* depending on whether these are lower than, equal to or higher than the median, and *medpd<-medpd[medpd!=0]* checks to see that *medpd* is not equal to zero, discarding those values that are zero.

```
#Analysis of Example 4.1 Cross-over Trials in Clinical
#Research
#VAS data
```

```
#Input data
#n1 = numbers in first sequence, n2 in second
#seq = sequence, perdiff = semi period differences
n1<-6
n2<-6
seq<-factor(c(rep("forsal",n1),rep("salfor",n2)))
perdiff<-c(9.5,30,17,9,25.5,4,−14,0.5,−18,−9,−12,
−24.5)

#Various calculations follow
#Calculate basic estimators from semi-period differences
basicest<-2*(ifelse(seq=="forsal",perdiff,-perdiff))
#Count number of non-zero basic estimators
nnzb<-sum(abs(sign(basicest)))
#Count number of positive basic estimators
npb<-sum(sign(abs(basicest)+(basicest)))
#Calculate signs of semi period differences
#Only retain non-zero cases
signpd<-sign(perdiff[perdiff!=0] )
seqsign<-seq[perdiff!=0]
#Calculate if semi-period difference is above or below median
#Only retain non-zero cases
medpd<-sign(perdiff-median(perdiff))
seqmed<-seq[medpd!=0]
medpd<-medpd[medpd!=0]
#Split perdiff into two groups by sequence
perdiff.1<-perdiff[seq=="forsal"]
perdiff.2<-perdiff[seq=="salfor"]

#begin tests
#sign test
binom.test(npb, nnzb, p=0.5, alternative="two.sided")
#perform Wilcoxon signed rank test
wilcox.test(basicest, alternative="two.sided", mu=0,
paired=F, exact=T, correct=T)
#Perform period adjusted sign test
fisher.test(signpd, seqsign)
#perform Brown−Mood median test
fisher.test(medpd, seqmed)
#perform Wilcoxon Mann-Whitney test
wilcox.test(perdiff.1, perdiff.2, alternative=
"two.sided", mu=0, paired=F, exact=T, correct=T)
```

S-Plus does not have a routine for analysis of ordered categorical data, although the *polr* function is available in the MASS library created by

Venables and Ripley and described in their book (Venables and Ripley, 1999). Its use is not covered here and the analysis of Example 4.3 will not be described.

The analysis of Example 4.5 may be achieved using the following code:

```
#Analysis of Example 4.5
#Cross-over Trials in Clinical Research
#Input data
sequence<-factor(c("PS","PS","SP","SP"))
period<-factor(c(1,2,1,2))
treat<-factor(c("P","S","S","P"))
exacerb<-c(14,6,14,23)
#Define contrasts
options(contrasts=c(factor="contr.treatment",
ordered="contr.poly"))
#Carry out Poisson regression
fit<-glm(exacerb~sequence+period+treat,family=poisson)
summary(fit, corr=F)
anova(fit,test="Chi")
```

Note the use of the *glm* command and that the error is declared to be Poisson using *family=Poisson*. The log link is assumed. The *summary* command requests estimates and standard errors and the *anova* command calls for an analysis of deviance.

The output includes the following:

```
Coefficients:
                    Value    Std. Error      t value
 (Intercept)    2.6390573    0.2672612     9.8744484
    sequence    0.6718674    0.2970699     2.2616475
      period   -0.1754305    0.2970699    -0.5905361
       treat   -0.6718674    0.2970699    -2.2616475
```

and

```
Analysis of Deviance Table

Poisson model

Response: exacerb

Terms added sequentially (first to last)
          Df   Deviance Resid.  Df  Resid. Dev      Pr(Chi)
    NULL                         3   10.65077
sequence   1   5.148151          2    5.50262   0.0232711
```

```
period  1   0.017545          1     5.48508  0.8946231
 treat  1   5.485075          0     0.00000  0.0191795
```

The parameter estimate for *treat* and standard error are as we obtained by hand. The analysis of deviance gives a P value of 0.019 for treat. These results agree with those obtained with GenStat®.

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

5

Normal Data from Designs with Three or More Treatments

5.1 WHY DO WE HAVE DESIGNS WITH THREE OR MORE TREATMENTS?

The *AB/BA* design is extremely popular. It is the natural way to study, through the medium of the cross-over, the effect of an active treatment using a placebo control, and it is the simplest cross-over design permitting a comparison of a new treatment with an old one. Nevertheless, not all clinical trials, whether parallel or cross-over, consist of a simple comparison between treatment and control. In this chapter ways of analysing Normally (or approximately Normally) distributed data from cross-over designs with three or more treatments will be considered. In fact the scope of the chapter will be more restrictive than implied by the title since all of the designs considered will not only have exactly the same number of periods as treatments but will have the further restriction that every patient will receive every treatment being studied once and once only. Thus excluded are *incomplete block designs*, in which no patient receives every treatment, *extra period designs*, in which every patient not only receives each treatment once but at least one treatment at least twice, *n of 1 designs*, in which a single patient receives every treatment many times (although these trials are in any case usually restricted to two treatments) as well as other more complex mixed designs. These will be dealt with in differing degrees of detail in Chapters 7 and 10.

Before considering the design and analysis of cross-over trials with three or more treatments we shall first consider briefly some general circumstances which may lead an investigator to wish to study more than two treatments. The remarks which follow have no necessary connection to cross-over trials and may also be relevant for parallel group trials.

5.1.1 Dose-finding trials

In investigations of tolerability for a new drug in Phase I of a clinical programme or initial investigations of efficacy in Phase II, it is usual to run trials in which a number of different doses are used within one trial with a view to choosing a particular dose for examination for the major claims of efficacy in Phase III (Pocock, 1983).

Since a dose recommendation consists not only of a *unit dose* (the amount which is given in a single administration) but also a *treatment regimen* (the frequency with which a treatment is to be given: twice daily, three times daily etc.), since establishing a *therapeutic window* requires both *minimum effective* and *maximum tolerated* doses to be determined, since in any case efficacy may have many aspects to it (for example *onset of action* and *duration of action*) and finally since different patient populations (e.g. children and adults) commonly require different doses, many trials are necessary in a drug development programme as part of the process of establishing final doses.

Because we both expect a non-linear dose response and wish to investigate departure from linearity, three doses at least are usually studied. On occasion a placebo control may take the place of the lowest dose. Designs with four or five treatments in total are not uncommon.

5.1.2 Combination therapy trials

In certain indications it is common to treat patients with combinations of therapies. Suppose we are interested in studying the combination *AB* of two treatments *A* and *B*. It is desirable under such circumstances to show that *AB* is superior to *A* alone and also superior to *B* alone. In a double-blind trial we would then have at least three 'treatments' under study:

$$A \text{ and placebo to } B$$

$$B \text{ and placebo to } A$$

$$A \text{ and } B.$$

(Two placebos are necessary in practice since it is most unlikely that two active treatments will be perfectly matched. This method of maintaining blindness is referred to as *double dummy loading*.)

5.1.3 Factorial designs

There are good reasons (Pledger, 1989) for adding a fourth 'treatment' to the basic combination therapy trial in which neither active treatment is given. Thus in terms of the double-blind design discussed above this would require a further treatment:

Placebo to *A* and Placebo to *B*.

One possible way of looking at the resulting trial is as a factorial experiment with two factors, 'treatment *A*' and 'treatment *B*', each at two levels, 'present' and 'absent'.

There are other circumstances under which factorial designs may be useful. In asthma the protective effects of bronchodilating treatments against 'challenges' by stimuli, such as exercise, cold air, allergens etc. are frequently studied. In a placebo-controlled trial of a bronchodilator it would then be sensible to have four 'treatments' defined by the four possible combinations of the factors 'therapy' (verum or placebo) and 'challenge' (present or absent) (Senn, 1989a). See also Chapters 7 and 9.

5.1.4 Response surface designs

Sometimes it is desirable to study various doses of treatments in combination. Consider, for example, the combination therapy trial discussed in Section 5.1.2. Simply proving that the combination of *A* and *B* is superior to *A* and superior to *B* does not in itself suggest that the combination therapy is a good treatment. A double dose of *A* or a double dose of *B* might be superior to the combination. Of course these higher doses may bring problems in terms of tolerability but then so may the combination itself. The problem of deciding upon an optimal therapy is then one of studying the response surfaces for tolerability and efficacy as functions of changes both in doses of *A* and doses of *B*. Such investigations are extremely complex and are rarely concluded successfully through the medium of a single trial. Nevertheless, the individual trials which are used as part of the programme of investigation will include a number of treatments.

5.1.5 Equivalence studies

It is an advantage for the physician to have at his disposal a number of possible alternative therapies even if in terms of their average effect these are roughly equivalent. There are various reasons why: some patients may respond better to one drug, some to another; other patients may be allergic to a particular drug; patients may build up a tolerance to a particular therapy and need to be switched to something else. Sometimes a new therapy is simply an alternative formulation (for example, dispersible, slow release, suppository etc.) of an existing treatment. The new formulation may bring practical benefits in terms of route of administration to a particular group of patients. For example powder aerosols require less coordination for inhalation than do pressurized solution aerosols and so are more suitable for young children.

Under such circumstances it is sometimes sufficient for a new treatment, whether a new chemical entity or a new formulation, to show equivalent average effects (to some acceptable degree of equivalence) to an existing therapy. In the case of a new formulation, special *pharmacokinetic* studies, known as bio-equivalence studies, comparing new with standard formulation in terms of levels over time of active ingredient in blood and urine, may be sufficient to establish equivalence. However, for these studies it may be necessary (and for studies involving new chemical entities it *will* be necessary) to compare the treatments in terms of efficacy.

There is a problem, however, if this comparison is the only purpose of the trial (Temple, 1982). If at the end of the trial new and standard treatments are shown to be equivalent there is no proof within the trial that either therapy was effective. If at all ethically possible, it is desirable, therefore, to include placebos in such trials in order to demonstrate 'assay sensitivity'. Demonstrated sensitivity makes any apparent equivalence of the treatments more convincing (Hasselblad and Kong, 2001; International Conference on Harmonisation, 2000).

5.1.6 'Gold standard' trials

Designs with a new therapy, a standard therapy and a placebo are sometimes referred to as gold standard trials. They are not limited, however, to cases in which it is desired to assert the equivalence of two active treatments. It may be that the main interest of the trial is the comparison of the new treatment to placebo but that a standard therapy is included in order to provide, via its comparison with placebo, some assessment of the sensitivity of the trial. Such three armed trials are extremely common and we shall encounter an example in this chapter.

5.2 SEQUENCES FOR TRIALS WITH THREE OR MORE TREATMENTS

There is far more to designing a cross-over trial than simply specifying the sorts of sequences of treatments patients will receive and describing how patients will be allocated to them. Some of the wider issues affecting design will be discussed in more detail in Chapter 9. In this section we consider issues affecting sequences and allocation.

Suppose we have a trial in which we wish to study k treatments, A, B, C etc. and we decide that each patient will receive each treatment at least once. We then have k choices for the first treatment he will receive, $k-1$ choices for the second and so on. There will thus be $k! = k(k-1)(k-2)\ldots(2)(1)$ possible sequences. For example for a three treatment trial we have $k! = 3! = 6$ possible sequences and they are: *ABC, ACB, BAC, BCA, CAB, CBA*. For $k = 4$ there are 24 possible sequences and for $k = 5$ and 6, 120 and 720 sequences respectively. Cross-over trials in which patients each receive more than 6 treatments are rare.

If the investigator is totally unconcerned about the effect of time on his experiment, then (always assuming that carry-over is not an issue) he may simply allocate patients at random to a sequence of treatments (since essentially it is totally irrelevant to him which sequences are represented in his experiment and which are not). There is no reason, then, why his trial should present any degree of balance with respect to sequences or with respect to the association between periods and treatments.

If, on the other hand, he is concerned about possible trends over the study as a whole and consequently about potential differences between periods he will wish to introduce a degree of balance into the experiment. The way in which he can do this is best illustrated by example. Suppose he wishes to study three treatments, A, B, C, then if he allocates patients at random and in equal numbers to the sequences

$$\begin{array}{ll|ccc|} \multicolumn{2}{c}{} & \multicolumn{3}{c}{\text{Period}} \\ \multicolumn{2}{c}{} & \multicolumn{1}{c}{1} & 2 & \multicolumn{1}{c}{3} \\ \cline{3-5} & \text{I} & A & B & C \\ \text{Sequence} & \text{II} & B & C & A \\ & \text{III} & C & A & B \\ \cline{3-5} \end{array} \tag{5.1}$$

every treatment will be represented in every period with the same frequency. The sequences represented in (5.1) form, when taken together, a *Latin square*, every treatment being represented once and once only in each column and in each row. Cross-over trials using such sequences are often referred to as *Latin square designs* but several important differences to Latin square designs in agricultural field experiments, where they were first introduced, must be noted. The first is that in a cross-over design the sequences are likely to be replicated many times, a number of patients being allocated to each sequence. The second is that in a cross-over trial we may conceive of a patient's second treatment affecting his third treatment but not vice versa, whereas if treatments in a field are arranged in a Latin square the treatment given to the third plot in a row may just as conceivably affect growing conditions in the second as vice versa. The third is that there is an important qualitative difference between patients and periods compared to the difference between rows and columns of a field. Nevertheless, we shall refer to such arrangements of sequences in cross-over trials as Latin square designs, but at the same time warn the reader against assuming that everything which applies to Latin square designs in general is also valid for Latin square designs in cross-over trials.

The Latin square given by (5.1) is not the only one available. We could also have used the sequences

$$ACB/BAC/CBA. \tag{5.2}$$

Not all sets of three sequences constitute a Latin square, however. In total there are $6!/(3!3!) = 20$ ways of choosing three sequences from 6 and only the two sets given in (5.1) and (5.2) are Latin squares. If we choose to balance a three-treatment three-period cross-over design for period we shall have to choose, therefore, to treat a total number of patients which is a multiple of 3. If instead of simply using one of the two sets of sequences given by (5.1) and (5.2) we use them both we shall then have to study a multiple of 6 patients.

Perfect balance with respect to sequences, is not, however, an absolute requirement for an analysis which allows for period effects as well as treatments: it is simply that other things being equal such designs are more efficient. The search for balance should not become a holy grail which has precedence as an objective over all others (Mead, 1990). If, for example, we have a three-period three-treatment cross-over with 14 patients it cannot be balanced for period. On the other hand things could be arranged so that a subset of 12 of the patients could constitute a balanced set. The design with 14 patients would then be more efficient than a design with 12.

For designs with four treatments, as noted above, there are 24 possible sequences. It also so happens that if we take the sequences four at a time then there are 24 sets which will form a Latin square. The sequences are given in Table 5.1. In the table all 24 sequences will be found to be represented in a given group of six Latin squares defined by a Roman numeral. The first sequence in the Latin squares in a given column defined by an Arabic numeral is always the same. For group II the second sequence from group I is also retained, for group III the third sequence and for group IV the fourth. The six Latin squares marked * are *Williams squares* (Williams, 1949). They have the property that every treatment follows every other once. Particular advantages are sometimes claimed for these squares. These claims will be critically examined in Chapter 10. The squares marked + (all those in group I) have a property which may be useful on occasions. If any pair of treatments in a given sequence is chosen it will be found that there is another sequence in which the same pair appears with periods reversed. For example if we look at square 13 and take treatments *A* and *D* they appear in periods 1 and 4 in the first sequence but also in periods 4 and 1 in the fourth sequence. Similarly they appear in periods 3 and 2 in the second sequence and in periods 2 and 3 in the third. A use which may be made of this property will be examined in Chapter 6.

Frequently there is no good reason to choose any one Latin square in preference to another and the choice may then be made at random. There is often no reason either why a single Latin square should be used in a given trial (although

Table 5.1 The 24 possible treatment sequences for a four-period, four-treatment cross-over design arranged in the 24 possible Latin squares.

	1	2	3	4	5	6
I	ABCD +	ABDC +	ACBD +	ADBC +	ACDB +	ADCB +
	BADC	BACD	BDAC	BCAD	BDCA	BCDA
	CDAB	CDBA	CADB	CBDA	CABD	CBAD
	DCBA	DCAB	DBCA	DACB	DBAC	DABC
II	ABCD	ABDC	ACBD	ADBC	ACDB*	ADCB*
	BADC	BACD	BDAC	BCAD	BDCA	BCDA
	CDBA	CDAB	CBDA	CADB	CBAD	CABD
	DCAB	DCBA	DACB	DBCA	DABC	DBAC
III	ABCD	ABDC*	ACBD	ADBC*	ACDB	ADCB
	BCDA	BCAD	BDCA	BACD	BDAC	BADC
	CDAB	CDBA	CADB	CBDA	CABD	CBAD
	DABC	DACB	DBAC	DCAB	DBCA	DCBA
IV	ABCD*	ABDC	ACBD*	ADBC	ACDB	ADCB
	BDAC	BDCA	BADC	BCDA	BACD	BCAD
	CADB	CABD	CDAB	CBAD	CDBA	CBDA
	DCBA	DCAB	DBCA	DACB	DBAC	DABC

some analyses are simpler if this is done) rather than a number of squares, apart from convenience in preparing trial material. It has been my habit in designing cross-over trials to use all six sequences for a three-period three-treatment design and to use a single Latin square for a four-treatment design.

For the purpose of randomizing patients, a design may be chosen at random from a number of suitable candidates and the patients are then allocated at random to the appropriate sequences. To ensure balance they may be allocated in equal numbers to the sequences chosen. I do not recommend using sub-centre blocks. For example, in a four-treatment trial in 20 patients in a single centre I would simply ensure that five patients were allocated at random to each of the four sequences chosen (assuming I had used a single Latin square) but place no further restriction on allocation.

For Latin squares for higher-order designs with five or even six treatments the reader should consult Fisher and Yates (1974). More general approaches and the use of SAS® or Genstat® to find designs will be considered in Chapter 9.

5.3 ANALYSES IGNORING THE EFFECT OF PERIOD

We now introduce an example which will be used to illustrate various possible analyses. Following the plan of previous chapters we shall consider first of all analyses which ignore any possible effect of period.

Table 5.2 (Example 5.1) Three-treatment cross-over. Forced expiratory volume (FEV_1) after an exercise test.

		FEV_1 (ml)				
Patient	Sequence*	Period 1	Period 2	Period 3	Contrast†	Mean
1	*FSP*	3500	3200	2900	300	3200
10	*FSP*	3400	2800	2200	600	2800
17	*FSP*	2300	2200	1700	100	2066.7
21	*FSP*	2300	1300	1400	1000	1666.7
23	*FSP*	3000	2400	1800	600	2400
4	*SPF*	2200	1100	2600	400	1966.7
8	*SPF*	2800	2000	2800	0	2533.3
16	*SPF*	2400	1700	3400	1000	2500
6	*PFS*	2200	2500	2400	100	2366.7
9	*PFS*	2200	3200	3300	−100	2900
13	*PFS*	800	1400	1000	400	1066.7
20	*PFS*	950	1320	1480	−160	1250
26	*PFS*	1700	2600	2400	200	2233.3
31	*PFS*	1400	2500	2200	300	2033.3
2	*FPS*	3100	1800	2400	700	2433.3
11	*FPS*	2800	1600	2200	600	2200
14	*FPS*	3100	1600	1400	1700	2033.3
19	*FPS*	2300	1500	2200	100	2000
25	*FPS*	3000	1700	2600	400	2433.3
28	*FPS*	3100	2100	2800	300	2666.7
3	*SFP*	2100	3200	1000	1100	2100
12	*SFP*	1600	2300	1600	700	1833.3
18	*SFP*	1600	1400	800	−200	1266.7
24	*SFP*	3100	3200	1000	100	2433.3
27	*SFP*	2800	3100	2000	300	2633.3
5	*PSF*	900	1900	2900	1000	1900
7	*PSF*	1500	2600	2000	−600	2033.3
15	*PSF*	1200	2200	2700	500	2033.3
22	*PSF*	2400	2600	3800	1200	2933.3
30	*PSF*	1900	2700	2800	100	2466.7

**F*: Formoterol 12 μg
P: Placebo
S: Salbutamol 100 μg
†Contrast $= F - S$
Means: formoterol = 2720.67 salbutamol = 2296 placebo = 1621.67

Example 5.1 Table 5.2 is based on data considered in Senn and Hildebrand (1991) which come from a single centre in the multi-centre trial reported by Tsoy *et al.* (1990). The data present values of forced expiratory volume in one second (FEV_1) obtained after an exercise challenge in a three-period three-treatment double-blind cross-over trial comparing the protective effect of a single dose of an experimental treatment, formoterol solution aerosol

(12 μg), to a single dose of a standard therapy, salbutamol suspension aerosol (100 μg) and placebo for patients suffering from exercise-induced asthma. (The trial might thus be regarded as an example of the so-called gold standard trial described in Section 5.1.6.) In a titration study carried out at the beginning of the trial an appropriate exercise test was established for each patient. The patient was then asked to perform this exercise test two hours after treatment on each of the three treatment days. The values reported are the lowest of a number of determinations of FEV_1 made in the period after the exercise test. All six possible sequences of formoterol, salbutamol and placebo were used. We shall assume that no attempt was made to balance for period and patients were allocated at random to the sequences. The data are illustrated in Figure 5.1.

Three points regarding the data should be noted. First, that patient number 29 is missing and there are only 30 patients in total. Second, that patient 20 has been measured to a different standard of precision. Third, that the representation in Figure 5.1 alone is sufficient to show that there are genuine differences between the treatments. We now consider what a formal analysis might show.

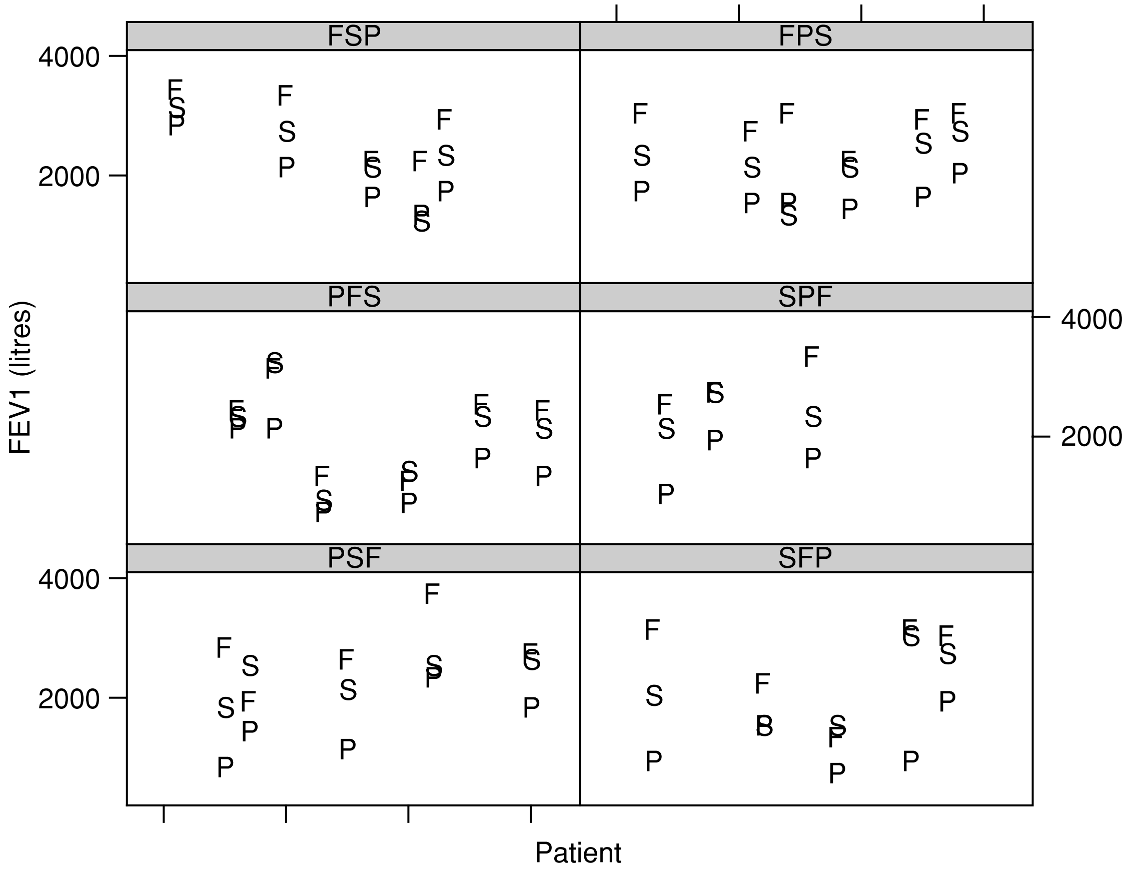


Figure 5.1 (Example 5.1) Forced expiratory volume in one second (FEV_1) recorded after exercise provocation in a cross-over comparing formoterol, salbutamol and placebo.

5.3.1 Basic estimator approach (matched-pairs *t*)

This is essentially the same approach which Student (1908) used to analyse the data of Cushny and Peebles (1905). Suppose that we are interested in analysing the contrast formoterol–salbutamol. We calculate a basic estimator corresponding to this contrast for each patient as given in the second to last column of Table 5.2. Now, if we are happy to ignore the effect of period then each of these estimates may be regarded as measuring the same thing. We may analyse the data just as we did in Section 3.2. First, we calculate the arithmetic mean of the 30 contrasts, $\bar{X}$, and their standard deviation, s, obtaining the results:

$$\bar{X} = 424.7\,\text{ml}, \quad s = 484.42\,\text{ml}.$$

We may regard $\bar{X}$ as a point estimate, $\hat{\tau}$, of the treatment effect τ. Its estimated standard error is then simply

$$\text{est } SE(\hat{\tau}) = s/\sqrt{n} = 88.44\,\text{ml}.$$

The t statistic is thus 4.80 on 29 degrees of freedom and highly significant. The critical value of t corresponding to a two-tail probability of 0.05 and 29 degrees of freedom is 2.045, and multiplying this by the standard error we obtain 180.9 ml. This may be added and subtracted from the point estimate 424.6 ml to obtain the 95% confidence limits for τ. We thus have the confidence interval for τ

$$244\,\text{ml} \leqslant \tau \leqslant 606\,\text{ml}.$$

There are various ways this analysis could be carried out using SAS®. One of the simplest is to use *proc univariate* on the 30 basic estimators. This will provide a number of summary statistics, including those necessary for calculating the confidence limits as well as calculating the t statistic and the associated P value.

5.3.2 Estimating the variance from the whole experiment

By using the matched-pairs t approach as we did in Section 5.3.1, we only used salbutamol and formoterol readings for the purpose of estimating the difference between formoterol and salbutamol and did not use the placebo values at all. This is eminently reasonable. There is no direct information concerning the difference between formoterol and salbutamol to be had in studying the placebo readings. In a design in which some patients had received formoterol and placebo but not salbutamol and others salbutamol and placebo but not formoterol there might be some indirect information to be had which was not otherwise available but this is not the case here. If, however, we believe that the variability that is to be observed within patients is the same whatever treatment they are given, then although there is no information on the formoterol–salbutamol difference to be had from studying the placebo values they may help us make a better estimate of the variability in the experiment as a whole.

The way in which the calculation may be performed with a pocket calculator will now be demonstrated. This is done purely for illustrative purposes. In practice one would use a computer package nowadays.

The relevant variance can be estimated using a two-way analysis of variance for the experiment as a whole. (Although it has been assumed in this book that the reader has a background in statistics which covers analysis of variance we shall remind him of various points regarding the calculation as we go through it.) The relevant effects are treatments and patients and the identity for the sums of squares (*SS*) is

$$SSTotal = SSPatients + SSTreatments + SSError.$$

If in general we have n patients and k treatments, the overall mean is $\bar{Y}_{..}$, the mean response over all patients for the ith treatment is $\bar{Y}_{i.}$ and the mean over all treatments for the jth patient is $\bar{Y}_{.j}$, then formulae for the sums of squares are

$$SSTotal = \sum\sum(Y_{ij} - \bar{Y}_{..})^2 \; SSPatients = k\sum(\bar{Y}_{.j} - \bar{Y}_{..})^2$$
$$SSTreatments = n\sum(\bar{Y}_{i.} - \bar{Y}_{..})^2 \; SSError = \sum\sum(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2.$$

For this example we have $n = 30$ and $k = 3$.

Traditionally one would have used computing formulae for these sums of squares but with a scientific calculator it is actually easier to obtain the first three directly. Such calculators usually provide a mean and standard deviation facility. Each of the first three sums of squares is simply *nk* times a sample variance, where this sample variance may be obtained by squaring the result given by pressing the σ_n key, having entered the relevant values. For *SSTotal* these are all the original *nk values*, for *SSPatients* these are the n means for each patient, and for *SSTreatment* these are the k means for each treatment. These means are all presented in Table 5.1. The *SSError* term is obtained by subtraction.

Performing the calculations we may construct the analysis of variance table below. (In order to obtain results to this precision it is necessary to use the patient and treatment means quoted to four decimal places. Thus for patient 17, use 2066.6667, not 2066.7. The precision illustrated, however, is not necessary to come to a reasonable conclusion.)

Source	Sum of squares (ml^2)	Degrees of freedom	Mean square (ml^2)	*F* ratio
Patients	21 279 472	$n - 1 = 29$	733 775	6.42
Treatments	18 428 682	$k - 1 = 2$	9 214 341	80.58
Error	6 632 452	$(n - 1)(k - 1) = 58$	114 353	
Total	46 340 606	$nk - 1 = 89$		

Various checks are possible. Thus for example, the mean of the patient means equals the mean of the treatment means equals the mean of the overall values and these means are produced automatically by the pocket calculator as a by-product of calculating the variances. Similarly, if when calculating the treatment means, a note is kept of the k sample variances (calculated using the divisor n) then the sum of these when multiplied by n will equal the sum of *SSPatients* and *SSError*. In this example the three variances are 365 719.6 ml^2 (formoterol), 311 397.3 ml^2 (salbutamol) and 253 280.6 ml^2, yielding a total when multiplied by 30 of 27 911 925 ml^2 which agrees closely with the total of *SSPatients* and *SSError* from the table above.

The mean squares are obtained by dividing the sums of squares by the corresponding degrees of freedom. The F-statistics or ratios are obtained by dividing patient and treatment mean squares respectively by the error mean square. Although our main purpose in constructing the table is to obtain the mean square error the F ratios are of minor interest in themselves. Under the null hypothesis that there is no systematic difference between patients in the values of FEV_1, the F ratio is distributed according to the F distribution with 29 and 58 degrees of freedom. The observed value is highly significant and the null hypothesis may be rejected, but it is a hypothesis which is scarcely of any interest. We take it for granted that patients vary. Similarly the F ratio for treatments provides a test of the equality of all the treatments and may be compared to an F distribution with 2 and 58 degrees of freedom. This F ratio is also highly significant. Here the hypothesis being tested is of slightly more interest, although in practice the pairwise comparison of treatments is more interesting than the simultaneous comparison of them all.

The mean square error in the table, 114 353 ml^2, provides an estimate of the variability in the experiment once patient and treatment effects have been accounted for. If we have the general model:

$$\text{observed value} = \text{overall mean} + \text{patient effect} + \text{treatment effect} + \text{error}, \tag{5.3}$$

with the patient effects summing to zero and the treatment effects summing to zero, then the mean square error is an estimate of the variance of the error term.

A more formal rendition of expression (5.3) is

$$Y_{ij} = \mu + \pi_j + \tau_i + \epsilon_{ij}, \quad i = 1 \text{ to } k, \ j = 1 \text{ to } n, \tag{5.4}$$

where Y_{ij} is the observation made on the jth patient under the ith treatment, π_j is the effect due to the jth patient, τ_i is the effect due to the ith treatment and $E(\epsilon_{ij}) = 0$, with $\sum_j \pi_j = 0$ and $\sum_i \tau_i = 0$.

The basic estimators we formed for each patient eliminated all effects but two: the difference between the two treatments in question and the errors. In terms of model (5.3), if we calculate the basic estimator for the jth patient treatment r compared to treatment s, then we have

$$C_{i,\ r-s} = Y_{ir} - Y_{is} = \tau_r - \tau_s + \epsilon_{ir} - \epsilon_{is},$$

where $C_{i,\ r-s}$ is the basic estimator for the ith patient.

If we now make the further assumption that the variances of the error terms in our models are identical and that there are no covariances between them (or put another way, that once we know what the true treatment effect is and once we know what a given patient's true mean level is, everything which is explainable in the model has been explained) then, since it involves the difference between two independent errors, the variance of the basic estimator is $2\sigma^2$, where σ^2 is the variance of a single error term. It thus follows, by (2.3), that the mean of n such estimators has variance $2\sigma^2/n$.

The constant variance and zero covariance assumption can be expressed symbolically as

$$\begin{aligned} E\left(\epsilon_{ij}\epsilon_{gh}\right) &= \sigma^2 \quad \text{if } i = g \text{ and } j = h, \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{5.5}$$

Hence to use the mean square error which we obtained from the analysis of variance to help us estimate the variance of our estimated contrast we need to multiply it by 2 and divide by n. In this case $n = 30$ so that dividing 114 353 ml^2 by 15 we obtain 7623.53 ml^2. The square root of this gives the standard error of 87.3 ml, which we may compare to the figure of 88.44 ml which we obtained in Section 5.3.1.

5.3.3 Discussion

The estimate of the formoterol–salbutamol contrast is no different in Section 5.3.2 than it was in Section 5.3.1. Indeed the estimators are identical. Since the estimators are identical the variances are also identical. Put symbolically the variance of 5.3.1 is γ^2/n, whereas that of 5.3.2 is $2\sigma^2/n$, but $\gamma^2 = 2\sigma^2$ and so the two are identical. What is different is that we don't know what the unknown γ^2 and σ^2 are and we have estimated them differently.

For a given patient, the basic estimator has a variance γ^2 (or $2\sigma^2$) and this is what we estimated directly in Section 5.3.1 by comparing the variability of the basic estimators amongst themselves. In so doing we only used readings obtained under salbutamol and formoterol and completely ignored the values obtained under placebo. The degrees of freedom available for estimating γ^2 were $n - 1$ or, in this case, 29. In Section 5.3.2 we also used the placebo readings to estimate variability and this led us to an indirect calculation. The degrees of freedom available are the degrees of freedom for the mean square error in the analysis of variance table. In general these are $(n - 1)(k - 1)$ or, in this case, 58.

The value of these extra degrees of freedom to us can be judged (partly) by looking at the critical values of the t distribution for a 5% (two-tailed) test. For 29 degrees of freedom these are ±2.045 and for 58 degrees of freedom ±2.002. Thus, other things being equal, we are able to make more precise statements using the variance with the higher degrees of freedom, since we will be able to use a lower multiplier, for example, when calculating 95% confidence intervals. In this case we can use 2.002 rather than 2.045.

This does not mean that the variance with more degrees of freedom is automatically to be preferred. The degrees of freedom exceed the number of patients and if the model for the error terms given in (5.3) were incorrect this could lead to problems. For example, it could be the case that variances increased with increasing FEV_1. This would make the variability under placebo, for which the values are in general lower, unreliable as an estimate of the variability under the bronchodilators. A further point is that the t statistic is extremely robust providing that the estimate of variance is internal: that is to say exactly the same observations contribute to the variance in the denominator as contribute to the estimate in the numerator. For these and other reasons (Senn and Hildebrand, 1991), on the whole I prefer an estimate of variance which is internal such as that in Section 5.2.1 to one which is partly external such as in Section 5.3.1. (See also discussion in sections 3.16.2 and 3.18.1.) In practice in many cases, as is the case here, it makes little difference which is chosen.

5.4 ALLOWING FOR PERIOD EFFECTS

In Example 5.1 we assumed that patients were allocated at random to the various treatment sequences. Had it been the intention to adjust for period effects a more efficient procedure would have been to allocate patients in equal numbers to the six sequences: five per sequence. It is interesting that the first 18 patients *are* perfectly balanced, as if a block of 6 was used but later abandoned.

It is generally possible in such an unbalanced design, however, to allow for the period effect. Exceptions may occasionally arise where the degree of imbalance is so severe that no adjustment is possible. We now illustrate two analyses which are possible providing some minimal balance is kept. As was discussed in Section 5.2 such analyses will be more efficient, other things being equal, if balance is maintained. If, however, the investigator has forced balance in his method of allocating patients then I consider that he is obliged to adjust for the period effect because his actions show that he considers that it is important.

5.4.1 Basic estimators approach allowing for period

A very simple analysis of Example 5.1 allowing for period effect may be based on basic estimators (Senn and Hildebrand, 1991).

The first step is to average the basic estimators within each group. This has been done in Table 5.3 which gives, for each sequence group, the number of patients (m), the mean of the basic estimators, the variance (calculated by dividing by $m - 1$), the corrected sum of squares (the variance multiplied by $m - 1$), as well as the mean of the means over all groups and the sum of the corrected sums of squares.

We assume that for a given patient the basic estimator will reflect three things: first, the treatment contrast being estimated (formoterol–salbutamol in this case), second, a difference between periods which depends on the sequence group to which the patient was allocated and third, random variation. By averaging the estimators within a sequence group we reduce the relative importance of the random element whilst preserving the treatment and period effects. If we go one step further and average the six sequence means then, on the assumption that the period effect is additive and that there is no period by treatment interaction, providing the sequences are balanced, we eliminate the effect of difference between period. Note that it is only necessary for the sequences when taken together to form a Latin square or a series of squares. It is not necessary for there to be an equal number of patients per sequence.

Table 5.3 (Example 5.1) Illustration of basic estimator approach adjusting for period.

Sequence group (ml^2)	Number of patients m	Degrees of freedom $m-1$	Statistics for basic estmators		
			Mean (ml)	Variance (ml^2)	CSS
FSP	5	4	520	117 000	468 000
SPF	3	2	466.7	253 333.3	506 667
PFS	6	5	123.3	48 866.7	244 333
FPS	6	5	633.3	318 666.7	1593 333
SFP	5	4	400	260 000	1040 000
PSF	5	4	440	523 000	2092 000
Overall	30	24	430.6		5944 333

$\hat{\sigma}^2 = 5944\,333/(4+2+5+5+4+4) = 5944\,333/24 = 247\,680.5\,\text{ml}^2$
est var $= 247\,680.5(1/5+1/3+1/6+1/6+1/5+1/5)/36 = 8715\,\text{ml}^2$
est standard error $= 93.4\,\text{ml}$

Table 5.4 Illustration of how period effects are eliminated by averaging.

Sequence	Period difference		
FSP	$+\beta_1$	$-\beta_2$	
SPF	$-\beta_1$		$+\beta_3$
PFS		$+\beta_2$	$-\beta_3$
FPS	$+\beta_1$		$-\beta_3$
SFP	$-\beta_1$	$+\beta_2$	
PSF		$-\beta_2$	$+\beta_3$
Total	0	0	0

Table 5.4 shows why the period effects are eliminated. In the table β_1, β_2 and β_3 represent the (unknown) effects of period 1, 2 and 3 respectively. On the assumption that these effects are additive, then, for example, in using the difference between the formoterol reading and the salbutamol reading as a basic estimate of the difference between these treatments, we are also estimating for a patient in sequence *FSP*, the difference between the first and second period, $\beta_1 - \beta_2$. If all of these period differences are added together they will be seen to sum to zero.

Note that the same would not happen if we simply averaged the 30 basic estimators for this example in one go (as we did in Section 5.3) since we have an unequal number of patients per sequence. The reader may check for himself that if that were to be done we should be left with a bias equal to $(3\beta_1 + \beta_2 - 4\beta_3)/30$.

The sum of the corrected sum of squares is 5944 333 ml^2 and this figure has been achieved by adding together the six corrected sums of squares corresponding to the six sequence groups. These six sums of squares being formed within each sequence group and hence comparing like with like as regards any effect of period, have had the period effects eliminated from them. The sum is therefore also uninfluenced by the period effects and in order to provide an unbiased estimate of the variability of the basic estimators it remains simply to divide this figure by the total degrees of freedom. This total is obtained by summing the degrees of freedom for each sequence group and thus equals $4 + 2 + 5 + 5 + 4 + 4 = 24$. Therefore our estimate of the variance of a basic estimator is $5\,944\,333\,\text{ml}^2/24 = 247\,681\,\text{ml}^2$.

To obtain the variance of our overall estimated treatment effect we now argue as follows. Each sequence mean has a variance of the form σ^2/m_s, where m_s is the number of patients in the sequence groups. If we add r such sequence group means together and divide by r we have:

$$\text{var}\,(\text{est treatment effect}) = \sigma^2 \sum_s (1/m_s)/r^2. \tag{5.6}$$

In Example 5.1, m_1, m_5 and $m_6 = 5$, $m_2 = 3$, and m_3 and $m_4 = 6$, whereas $r = 6$ and so the variance of the estimated treatment effect is $0.0352\,\sigma^2$. These values have been used in the calculation in Table 5.3. Finally we take the square root to obtain the estimated standard error which is 93.4 ml.

These calculations may be carried out very simply using SAS®. The technique we have just described for estimating the variance within sequence groups is essentially that of a one-way analysis of variance with the basic estimators as the observations and the sequence groups as 'treatments'. An analysis using *proc glm* can then be obtained as follows. First we need to calculate a basic estimator for each patient which we call *BASICEST*. We then regress this on the sequence group *SEQ*, having first defined this as a classification variable. This may be achieved using the following code:

```
proc glm;
  class SEQ;
  model BASICEST = SEQ;
  estimate 'Treatment' intercept 1;
run;
```

The *estimate* statement is used to obtain the estimated treatment effect and its standard error. 'Treatment' is the label which has been assigned to the estimate in this code. (The data analyst is free to choose what label he wishes.) Intercept is the name by which SAS® recognizes the intercept parameter in *proc glm*. In the code it has been assigned a weight of 1.

The intercept provides the required estimate in this model because the treatment is obtained as an average of the means. *Proc glm* automatically provides estimates for the *SEQ* terms which sum to 0. For packages which do not do this a linear constraint would have to be introduced, forcing them to add to 0.

For the purpose of calculating confidence limits using the results of this analysis it should be noted that the degrees of freedom to be used with the t distribution are the degrees of freedom used in calculating the variance (i.e. 24 in this case). Thus in this example, for calculating 95% confidence limits we should use a t value of 2.064, which corresponds to a tail area of $2\frac{1}{2}\%$ on 24 degrees of freedom. Multiplying this by the standard error we obtain a figure of 192.8 ml which may be added and subtracted from the point estimate of 430.6 to obtain confidence limits for the treatment effect τ of

$$238\,\text{ml} \leqslant \tau \leqslant 623\,\text{ml}.$$

5.4.2 Degrees of freedom in the basic estimator approach*

The basic estimator approach reduces a number of measurements made on a given patient to a single estimate. It thus follows as a consequence that the degrees of freedom for error for the estimate are fewer than the number of patients. As discussed in Section 5.3.3 and in Senn and Hildebrand (1991), t statistics formed using this approach are fairly robust. It is a simple technique to apply and understand but it is unnecessarily wasteful of degrees of freedom if the sequences which have been used come from more than one Latin square.

In Example 5.1, once the 60 readings under formoterol and salbutamol have been reduced to 30 basic estimators, there are 29 degrees of freedom available. Since there are three period effects but these sum to 0, there are two degrees of freedom which have to be assigned to periods if we are to adjust for any period effect. This would still leave us with 27 degrees of freedom whereas in fact we were left with only 24. The difference arises because we adjusted for differences between six sequences, thus losing five degrees of freedom rather than two. Had we used only a single Latin square we would have lost only two degrees of freedom to sequences.

In fact, in the example above, there are two uses which we can make of the missing degrees of freedom. One is that we could produce a slightly more efficient estimate of the treatment effect. The second is that we could produce a slightly more efficient estimate of the error variance. We now illustrate the first of these points with the help of Table 5.5.

In Table 5.5 the period effects have been expressed as in Table 5.4. Also included is a general scheme of weights by which the means of the basic estimators from the given sequence groups may be combined. If these weights are studied the following will be noted. First, they sum to 1. Second, the sum of the product of the period differences and the weights is zero. It thus follows that this is a scheme of weights for combining the means of the basic estimators for the sequences given which will produce unbiased estimates of the treatment effect. The scheme which was used in Section 5.4.1, where each sequence was given the equal weight of $\frac{1}{6}$, is a particular example of the general scheme given in Table 5.5 (as may be verified by substituting $\frac{1}{6}$ for w_1, w_2 and w_3) but it is by no means the only one. Now, if an equal number of patients had been studied in each sequence there would in fact be no advantage in using any other approach than weighting each sequence mean equally. However, since the number of patients varies from sequence to sequence so does the variance of each sequence mean. Other things being equal we might assume that the variance of these was inversely proportional to the number of patients studied.

If that were the case and we knew that there were no period effect, then the optimal scheme would in fact be to weight each sequence mean according to the number of patients, i.e. using the scheme of weights 5/30, 3/30, 6/30, 6/30, 5/30, 5/30. (All that happens if we do this is that we get the estimate we obtained in Section 5.3.) As the reader may check for himself, however, this is not a scheme which satisfies the conditions in Table 5.5. However, if we adapt expression (5.6) for a weighted estimator we obtain

Table 5.5 A general scheme of weights for combining sequence means of basic estimators for a three-treatment cross-over in six sequences.

Sequence	Period effect	Weight
FSP	$\beta_1 - \beta_2$	w_1
SPF	$-\beta_1 \quad + \beta_3$	w_2
PFS	$+\beta_2 - \beta_3$	w_3
FPS	$\beta_1 \quad + \beta_3$	$w_4 = \frac{1}{3} - 2w_1/3 + w_2/3 - 2w_3/3$
SFP	$-\beta_1 + \beta_2$	$w_5 = \frac{1}{3} + w_1/3 - 2w_2/3 - 2w_3/3$
PSF	$-\beta_2 - \beta_3$	$w_6 = \frac{1}{3} - 2w_1/3 - 2w_2/3 + w_3/3$

$$\text{var (est treatment effect)} = \sigma^2 \sum_s \left({w_s}^2 / m_s \right). \tag{5.7}$$

If this is minimized subject to the constraints in Table 5.5 we obtain $w_1 = 0.1576$, $w_2 = 0.1284$, $w_3 = 0.1622$ and hence $w_4 = 0.1629$, $w_5 = 0.1921$ and $w_6 = 0.1967$. Multiplying these weights by the means of the basic estimators for each of the six sequences we obtain as an estimate for the treatment effect, τ:

$$\hat{\tau} = 428.4\,\text{ml},$$

which is scarcely different from that obtained before. Substituting for w_1 to w_6 in (5.7) and using the value for $\hat{\sigma}^2$ obtained before of 247 680.5 ml^2, we obtain an estimate of the variance of the treatment effect of 8518 ml^2 and a standard error of 92.3 ml, which is also almost identical to that obtained before.

The second consequence of this waste of degrees of freedom is that the error variance is less precisely based than it need be. For example, since each of the Latin square means provides an estimate of the treatment effect then, if we can conceive of no systematic difference between them, that difference also reflects random variation. This fact could be used to recover degrees of freedom.

We shall not pursue this argument further here. The approach outlined in Section 5.4.1 is simple and robust. It is not optimal. It is less wasteful of degrees of freedom if based on a single Latin square. We now consider as an alternative a very common form of analysis.

5.4.3 The analysis via ordinary least squares (linear regression)

A very common simple general approach to analysing clinical trials, whether parallel or cross-over, is to use *ordinary least squares*. The response is expressed as a function of various systematic effects, including treatment, via a model. The parameters of the model are estimated using a general algorithm which minimizes the sum of the squares of the differences between the observed responses and those which would be predicted under the model. This general topic is known in statistics as linear regression. Linear regression is an extremely general and powerful way of analysing unbalanced experiments. It can also, of course, be used to analyse balanced experiments, and indeed nearly all of the analyses we have considered so far in Chapters 3 and 5 can be regarded as special cases of applying ordinary least squares. For example the model in Section 5.3.2 defined by expressions (5.4) and (5.5) is an example of the sort of model which is commonly analysed using least squares.

Another example occurred in Section 5.4.1 where, in using *proc glm* to analyse the basic estimators, we were using least squares on the basic estimators (not on the original observations). As we have already pointed out, however, the model was one in which we fitted enough parameters to account for differences between all of the groups (i.e. five parameters) whereas we only needed two to account for differences in period. Various interpretations could be given to the three extra parameters fitted, but if we truly believe that they are not necessary, then the procedure has some redundancy in it and it may be possible by eliminating it to improve estimates and also leave more degrees of freedom for error.

Nobody, except possibly the student of statistics working on artificial examples, does linear regression by hand. We shall not illustrate it here. All we shall do is describe how *proc glm* in SAS® may be used to provide an analysis of Example 5.1 which makes the adjustment necessary for period effects whilst using up no more degrees of freedom than is strictly necessary to achieve this. This method may be regarded as the natural extension of that of Section 5.3.2 to allow for a period effect. Formally, it corresponds to replacing (5.4) by

$$Y_{ij} = \mu + \pi_j + \beta_i + \tau_{(h)ij} + \epsilon_{ij}, \quad i = 1 \text{ to } n, \quad j = 1 \text{ to } k, \quad h = 1 \text{ to } k. \quad (5.8)$$

In this formulation, Y_{ij} is the measurement made on the ith patient in the jth period, β_i is the ith patient effect π_j, and $\tau_{(h)ij}$ is the effect due to the hth treatment (what the particular value of h is, is determined by i and j since the treatment chosen for a given period varies from patient to patient). This, together with (5.5), defines the model.

The method of analysis we use is essentially the same approach as we used for the *AB*/*BA* design in Chapter 3. We shall need to have the FEV_1 data stored as a single variable of $3 \times 30 = 90$ observations rather than as three variables of 30 observations each. (A feature of the ordinary least squares approach is that the repeated measures nature of the data is ignored.) For each observation we shall need a variable which records for which *PATIENT* it was obtained, in which *PERIOD* it was observed and what the *TREAT*ment was. *PATIENT* in this example is a classification variable with 30 levels, one for each patient and *PERIOD* is a classification variable with 3 levels, as is *TREAT*.

For example the first four data lines and the last two corresponding to Table 5.1 might look like this:

FEV1	PATIENT	PERIOD	TREAT
3500	1	1	FORM
3200	1	2	SALB
2900	1	3	PLAC
3400	10	1	FORM
⋮	⋮	⋮	⋮
2700	30	2	SALB
2800	30	3	FORM

Having organized the data in this form we now express the response as a function of the variables in our model and analyse the data using *proc glm* thus:

```
proc glm;
  class PATIENT PERIOD TREAT;
  model FEV1 = PATIENT PERIOD TREAT;
  estimate ‘For - Sal’ TREAT 10 − 1;
run;
```

As was explained in Chapter 3, the *class* statement is necessary to define the type of variable we are dealing with. Were it not included, for example, SAS® would regard *PATIENT* as being a single variable with values between 1 and 31. Using the class statement causes SAS® to replace the *PATIENT* variable with a series of dummy variables taking the values 0 or 1, thus permitting a separate effect for each patient to be fitted. The *model* statement defines the particular regression model being used: FEV_1 is to be regarded as determined by the patient's own particular level, an effect which depends on the period in which he is being treated as well as, of course, the treatment being received. Any other factors which affect *FEV1* are regarded as being unidentifiable and treated as random. Thus, the *model* statement expresses in computer code the relationship defined by (5.8). The *estimate* statement defines the particular contrast being examined. Taking the levels of *TREAT* in alphabetical order (formoterol, placebo, salbutamol) we define the particular contrast we are interested in, the difference between formoterol and salbutamol, using the weights 1, 0, − 1. (There is no limit to the number of estimate statements we can include and if, for example, we were interested in the contrast of the average of the active treatments to placebo we could simply add another line:

estimate ‘Act – Plac’ TREAT 0.5 –1 0.5;.)

Included in the SAS® output are the following lines:

```
Source            DF              Sum of squares    F value     Pr > F

Model             33              39 891 002.1686   10.50       0.0001
Error             56               6 449 603.3870
Corrected total   89              46 340 605.5556
                                  T for H0:         Pr > |T|    Std error of
Parameter         Estimate        Parameter = 0                 estimate
for-sal           422.621 977 89           4.79     0.0001      88.264 715 4.
```

The first part of this output shows the breakdown in analysis of variance terms of the total sum of squares into a component due to the model (i.e. all the terms fitted: patient, period and treatment effects) and an error term. Because there are 30 patients each measured three times, there are $89 = 90 - 1$ degrees of freedom altogether. Of these $29 = 30 - 1$ are used up in fitting patient effects, $2 = 3 - 1$ for periods and $2 = 3 - 1$ for treatments. This makes the total of $33 = 29 + 2 + 2$ given for the model and leaves $56 = 89 - 33$ associated with the error sum of squares. Thus there are 56 degrees of freedom to be used in connection with the t distribution for the purpose of establishing critical values for using in calculating confidence intervals or performing hypothesis tests.

The second part of the output gives the estimated treatment effect as well as its estimated standard error. If these are compared to the results obtained in Section 5.4.1 they will be seen to be very similar. Here the estimates and standard errors are 422.6 ml and 88.3 ml, whereas previously they were 430.6 ml and 93.4 ml.

5.4.4 Treating the patient effect as random

We have illustrated the analysis with *proc glm*, treating the patient effect as fixed. Of course, *proc mixed* of SAS® could also be used for the analysis, and this would be particularly useful if we wished to treat the patient effect as random. The code is virtually as before:

```
proc mixed;
  class PATIENT PERIOD TREAT;
  model FEV1 = PERIOD TREAT;
  random PATIENT PATIENT* TREAT;
  estimate 'For-Sal' TREAT 1 0 −1;
run;
```

This makes no difference to any results of interest. This is because all patients have data from all three periods and because there is no adjustment for carry-over. This means that there is no inter-patient information to recover.

If, on the other hand, we have missing data, or a so-called 'incomplete blocks' design, then there is some inter-patient information to recover and *proc glm* and *proc mixed* will produce different results. Incomplete block designs are considered in Chapter 7, where the issue of random effects will be picked up in more detail. Also if a carry-over effect is fitted, there will be a difference. That subject is covered in Chapter 10. For general advice on the use of *proc mixed*, including some examples applied to cross-over trials, the reader should consult Brown and Prescott (1999). Random effects analyses in S-Plus® and GenStat® are included in the appendices to this chapter.

5.5 OTHER MISCELLANEOUS ISSUES

We now discuss various other issues with relevance to the design and analysis of designs with three or more treatments.

5.5.1 Designs with four or more treatments

There is no particular difficulty in applying any of the methods discussed so far to higher-order designs involving four or more treatments. If the design is balanced so that the treatment sequences when taken together form a Latin square (or a number of Latin squares), the basic estimator approach of averaging within sequences and then over sequence can be applied quite generally. Similarly no change is necessary in the way in which an ordinary least squares analysis is performed. The computer code used, whether that of *proc glm* in SAS® or in any other package, will not require modifying (except that the *estimate* statement must include a weight for every treatment), it is simply necessary that the structure of the data reflects the design.

5.5.2 Use of baselines

Similar considerations apply here to those in Chapter 3. If the wash-out period is long compared to the treatment period, baselines taken at the beginning of each treatment period may be used as concomitant variables. If the patients are subject to individual trends over time these measurements may contain useful information and may be included quite easily as covariates in an ordinary least squares analysis. (An alternative interpretation in which the baselines are modelled in a similar way to outcomes and the two taken together are analysed via *generalized least squares* is discussed by Jones and Kenward, 1987a.)

Table 5.6 (Example 5.1) Three treatment cross-over. Forced expiratory volume (*FEV1*) before treatment (baselines).

			FEV1 (ml)		
Patient	Sequence	Period 1	Period 2	Period 3	Contrast
1	*FSP*	3500	3600	3400	−100
10	*FSP*	2800	3200	2600	−400
17	*FSP*	2100	2100	2100	0
21	*FSP*	2100	2100	1900	0
23	*FSP*	2500	2700	2800	−200
4	*SPF*	2100	2400	2200	100
8	*SPF*	2500	2300	2100	−400
16	*SPF*	2300	2300	1800	−500
6	*PFS*	2100	2200	2400	−200
9	*PFS*	3300	2800	2900	−100
13	*PFS*	1400	1500	1500	0
20	*PFS*	1890	1900	1600	300
26	*PFS*	2400	2000	2000	0
31	*PFS*	1800	1900	1900	0
2	*FPS*	2500	2800	2000	500
11	*FPS*	2200	1500	1800	400
14	*FPS*	1900	1800	1800	100
19	*FPS*	1900	1900	2100	−200
25	*FPS*	2300	2400	2300	0
28	*FPS*	2800	2900	2800	0
3	*SFP*	2600	2600	2200	0
12	*SFP*	1700	1700	2100	0
18	*SFP*	1200	1200	1200	0
24	*SFP*	2600	2700	2800	100
27	*SFP*	2600	2900	2900	300
5	*PSF*	1400	1600	2500	900
7	*PSF*	2300	2200	1900	−300
15	*PSF*	1200	1100	1900	800
22	*PSF*	2900	2500	2900	400
30	*PSF*	2600	2900	2900	0

For the ordinary least squares analysis with SAS® *proc glm* it is simply necessary to adapt the model statement to include a *BASE*line term. Thus for Example 5.1 we would write:

model FEV1 = PATIENT PERIOD BASE TREAT;.

For the basic estimator approach it is necessary first of all to calculate the basic estimators in baselines corresponding to those in outcomes. If we call the baseline difference *BASEDIF* then the *model* statement will look like this:

model BASICEST = SEQ BASEDIF;.

Table 5.6 gives the baseline readings for Example 5.1. Adjusting for these readings using the ordinary least squares approach produces an estimate of the treatment effect of 409.3 ml with an estimated standard error of 88.0 ml on 55 degrees of freedom. The corresponding analysis for the basic estimator approach adjusting for baseline differences gives a treatment estimate of 413.7 ml with standard error 90.0 ml on 23 degrees of freedom. These results, although very similar to each other, are slightly lower than those obtained when baselines were not fitted, reflecting the fact that the mean baselines are slightly higher under formoterol than under salbutamol and also reflecting the positive correlation between baseline and outcome. As a result these analysis of covariance techniques adjust the estimate of the treatment effect downward slightly.

5.5.3 Estimating other effects apart from treatment

We carry out cross-over trials to study the effects of treatments and these are always of primary interest in the analysis of any trial. For the purpose of planning future trials there may be some interest in studying other effects which we model. For example, the importance of the patient effects gives some clue as to the relative merits of parallel group and cross-over trials in the given field of study. If the patient effect is very important, other things being equal, it will be all the more advantageous to perform cross-over trials. Similarly there may be some interest in looking at the effect of periods to see whether they are worth allowing for in future analyses.

If we use computer packages such as SAS® to analyse such models there is no particular difficulty in obtaining the necessary information, although interpretation requires some care. For example, the output of the ordinary least squares analysis in Section 5.4.3 included the following lines (which were not previously presented):

Source	DF	Type III SS	*F* value	Pr > *F*
PATIENT	29	21 279 472.2222	6.37	0.0001
PERIOD	2	182 847.7241	0.79	0.4571
TREAT	2	18 531 247.7241	80.45	0.0001

Here ‘Type III SS’ stands for ‘Type III sum of squares’ and represents the contribution to the total corrected sum of squares made by a particular effect once all other effects have been taken into account. In a sense it is the variation observed which cannot be explained in terms of other effects.

The *F* value provides a test of the null hypothesis that the particular effect does not exist and is calculated by referring the mean square for that particular effect to the mean square error. In Section 5.4.3 we saw that the error sum of squares was 6 449 603.3870 ml^2 with 56 degrees of freedom. Dividing the sum of squares by the degrees of freedom we obtain mean square error $= 115\ 171.5\,\text{ml}^2$. The mean square error for *PERIOD*s is 91 423.9 ml^2 $= 182\ 847.7241\,\text{ml}^2/2$. The *F* ratio for *PERIOD*s is then $0.79 = 91\ 423.9\,\text{ml}^2/115\ 171.5\,\text{ml}^2$. The *P* value associated with this effect is 0.46 and so the effect is not significant.

The significance, or otherwise, of an effect, however, is not, in general, a good criterion for retaining or dropping it from current models. The cross-over design used, for example, may not be sufficiently sensitive to detect via a hypothesis test a difference between periods. Such differences might nevertheless be postulated to exist and they might still be important enough to make a difference to the overall interpretation of the experiment as regards treatment. (This is not, however, the case here.) In judging the importance of such effects it is probably also wise to compare their mean square to that for the treatment effect as well as to the mean square error.

5.5.4 Adjusting for carry-over

All of the methods outlined in this chapter rely on the assumption that carry-over is unimportant. By relaxing this assumption and replacing it with one which asserts that if important carry-over takes place, its effects will be negligible in all but the first following period (i.e. it will be first-order carry-over) and that the size of the effect will depend almost entirely on the treatment having carry-over and scarcely at all on the treatment into which it carries over, models may be developed in which carry-over effects can be not only estimated but eliminated from the treatment estimates.

I have already given reasons why I do not consider that such models are serious practical contenders for analysis. Nevertheless, they are commonly employed. A discussion of their use is deferred to Chapter 10 where these matters will be dealt with more fully.

5.6 RECOMMENDATIONS

The following recommendation regarding the design and analysis of cross-over trials with three or more treatments are made.

If the investigator is confident that period effects are negligible he may choose to allocate patients completely at random to possible treatment sequences. An analysis which then makes no allowance for period is a valid option.

In general, however, a more cautious approach will be to allow for a possible period effect and allocate patients at random to the chosen sequences in a way which ensures a reasonable balance. Designs based on Latin squares should then be employed. A method of analysis should then be chosen which allows for the period effect. Such methods may, however, still be considered as an option even if the patients have been allocated completely at random.

I do not advise forcing a greater degree of balance into the trial than that suggested above. It should be sufficient to use a block size for randomization equal to the number of patients in the trial.

If the wash-out period is such that the investigator may confidently expect that the effect of the previous treatment has been eliminated by the beginning of the following treatment period, and if it is considered that patients may be subject to individual and specific trends over time (independently of the treatments to which they are allocated), then there may be some value in fitting baselines in an analysis of covariance.

It is my opinion that, as with the *AB*/*BA* design, carry-over can only be dealt with by adequate wash-out. If necessary active wash-out periods may be employed but measurements should then be limited to the latter part of treatment periods. General planning issues affecting cross-over trials are considered in Chapter 9 and some conventional advice concerning modelling for carry-over is considered in Chapter 10.

Methods based on conventional ordinary least squares analyses are often used and well accepted. For designs in more than two periods they produce estimates for error variances based on more degrees of freedom than there are patients. In many cases this will present no difficulties but in some it may be problematic. The basic estimator approach may prove useful as a robust alternative.

APPENDIX 5.1 ANALYSIS WITH GENSTAT®

Since we have 30 patients each with three observations, we assume that the data consist of vectors of length $90 = 30 \times 3$. (Unlike Chapter 3, we do not illustrate the inputting of data in order to save space.) Here we have six such vectors. `FEV1` is a variate containing the results of the trial and `Base` is a similar variate for baseline values. `Patient` is a factor with 30 levels, `period` and `treat` are both factors with three levels and `Seqn` is a factor with six levels corresponding to the six sequences. In fact, this last factor is redundant in the analyses that follow and does not need to be included unless there is specific desire to examine differences between sequences.

Here we illustrate the approach to modelling in which terms are sequentially added to the current model using the *ADD* directive:

```
"Fixed effects model"
MODEL FEV1
TERMS[FACT=1] Patient+Period+Treat+Base
FIT[PRINT=model,summary,estimates,accumulated;\
CONSTANT=estimate; FPROB=yes; TPROB=yes; FACT=1]\
Patient+Treat
ADD[TPROB=yes; FPROB=yes] Period
ADD[TPROB=yes; FPROB=yes] Base
```

```
"Random effects model"
VCOMPONENTS[FIXED=Seqn+Period+Treat; FACTORIAL=1] \
RANDOM=Patient+Patient.Treat;CONSTRAIN=positive
REML[PRINT=model,components,effects,waldTests;\
PSE=alldifferences; MVINCLUDE=*; METHOD=fisher] FEV1
```

APPENDIX 5.2 ANALYSIS WITH S-PLUS

As was the case with GenStat®, we omit presentation of data input and manipulation in order to save space. For most of the analyses we use data vectors of length determined by number of patients times number of periods. We suppose that `fev1` and `base` are variates, that `patient` is a factor with 30 levels, that `period` and `treat` are factors with three levels and that `seqn` is a factor with six levels. All the above are vectors of length 30 × 3 = 90.

The analysis of basic estimators, however, reduces the data to a single contrast per patient. We thus have a vector `basicest` of length 30. The vector `seq` also has length 30 and is a factor with six levels.

The code is very similar to that used in Chapter 3. A difference here is that we have chosen to illustrate the approach of adding terms to an existing fit. This is done using the `update` function.

```
#Define contrasts
options(contrasts=c("contr.treatment", "contr.poly"))
#Begin fixed effects analysis
#Model with period and treatment effect
fit1<-lm(fev1~patient+treat)
summary(fit1,corr=F)
fit2<-update(fit1,.~.+period) #Add period effect
summary(fit2,corr=F)
fit3<-update(fit2,.~.+base) #Add baseline effect
summary(fit3,corr=F)
```

```
#Fit random effects model
fit4<-lme(fev1~seqn+period+treat, random=~1|patient)
summary(fit4)

#Reset definition of contrasts
options(contrasts=c(factor="contr.sum",
ordered="contr.poly"))

#Fit basic estimators approach
fit5<-lm(basicest~seq)
summary(fit5, corr=F)
```

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

6

Other Outcomes from Designs with Three or More Treatments

6.1 INTRODUCTION

The first edition of this book drew attention to the lack of methods for analysing non-Normal data from designs with three or more treatments. (It might have been even more appropriate to talk of difficulties with designs with three or more periods.) This situation has been partially remedied. For example, computational developments in SAS®, which now has a procedure, *proc nlmixed*, which permits analysis of random (or mixed) effect models with non-Normal outcomes, has made practicable more sophisticated approaches to designs with three or more periods. We shall cover the analysis of an example of such a design later in the chapter. Nevertheless, the various analyses introduced in the first edition are still useful as a way of getting a simple feel for data, and we shall cover these first.

Most of these simpler methods rely on tricks to reduce the particular aspect under investigation to a representation that resembles that of the *AB*/*BA* design.

6.2 ANALYSES WHICH TAKE NO ACCOUNT OF PERIOD EFFECTS

If the period effect is ignored, a very simple approach may be used to reduce an analysis from a cross-over design with more than two treatments to a form that makes it suitable for any of the methods for the *AB*/*BA* cross-over ignoring period outlined in Chapter 4. All that it is necessary to do is to treat the data for any two treatments being compared as if they were the only treatments which had been studied for that patient. Logically, such an approach ought to be accompanied by an unrestricted randomization. Thus for a four treatment (*A*, *B*, *C* and *D*) cross-over we would allocate patients at random to receive any one of the 24 possible treatment sequences without caring whether or not any balance between the periods was maintained. In comparing treatments A and B for a given patient we should take no note of the period in which they were given.

Using this approach, for example, we could use the *sign test* (Section 4.3.2) or the *Wilcoxon signed ranks test* (Section 4.3.3) for continuous outcomes. For binary outcomes we could use *McNemar's test* (Section 4.4.1).

In practice this approach is used quite often even for cases where it is allowed that the period effect may be important and the treatments have been balanced by period. I cannot declare any great enthusiasm for such a strategy and should always try to avoid it myself, but I must confess that my objections are largely aesthetic. I do not honestly think that in practice in most cases there is likely to be a serious danger of being badly deceived in doing this.

We now consider below what may be done to deal with period effects.

6.3 NON-PARAMETRIC ANALYSES ADJUSTING FOR PERIOD EFFECTS

We introduce this topic by considering an example.

Example 6.1 A placebo (P) controlled, multicentre trial in Sweden and Finland (Dahlof and Bjorkman, 1993) was run of two doses ($D1 = 50$ mg and $D2 = 100$ mg) of the potassium salt of diclofenac, a non-steroidal anti-inflammatory drug (NSAID), to establish the efficacy of this product in patients suffering from migraine attacks. The main outcome variable was pain intensity two hours after treatment measured on a 100 mm visual analogue scale (VAS) running from 0 = no pain to 100 = unbearable pain. The trial was double-blind: the treatments consisting of two active tablets ($D2$), two matching placebo tablets (P), or one of each ($D1$). All six possible treatment sequences were used. Each patient was given a treatment pack with the treatments marked 'attack 1', 'attack 2' and 'attack 3' (in the local language). Patients were not to use the first treatment until they had observed a treatment-free and attack-free period of at least one week. Thereafter they were to treat the first three attacks with the treatments provided in the order indicated.

Remark An interesting feature of this trial is that the interval between treatments is determined by the frequency of attacks. This raises several fascinating possibilities. For example, the onset of the next attack might be delayed by a successful treatment (this is a point which could be studied). If it were the case that a successful treatment affected the severity of the next attack, then this would be a form of carry-over. The analysis must proceed on the assumption that this is not the case.

We shall consider the analysis of data from country 2 and these are given in Table 6.1. It should be noted, however, that the trial was not analysed separately for each country: this is done here simply for convenience in order to reduce the number of data to be presented. Furthermore, we shall assume for simplicity that randomization took place over the country as a whole and we shall make no distinction between results from different centres.

The following should be noted about the data in Table 6.1. First that there are a number of missing values. Not all patients completed the trial and not all patients who completed the trial recorded all the measurements asked for.

Table 6.1 (Example 6.1) Three-treatment cross-over in three periods and six sequences. VAS for pain 2 hours after treatment for a migraine attack.

		VAS (mm)		Difference	
Sequence	Period 1	Period 2	Period 3	(*P* − *D*1) or (*D* − *P*1)*	Rank
D1 D2 P	0	2	42	42	10
	25	3	27	2	4.5
	55	.	.	.	
	0	68	9	9	7
	9	0	11	2	4.5
P D2 D1	0	8	31	31*	9
	81	47	6	−75*	1
	44	2	33	−11*	3
	18	0	22	4*	6
	33	39	17	−16*	2
	24	10	36	12*	8
D2 D1 P	27	39	56	17	11
	0	29	38	9	10
	0	0	0	0	6
	41	15	18	3	8
	45	66	41	−25	3
	59	16	24+	8	9
D2 P D1	0	33	1	−32*	2
	1	66	58	−8*	5
	0	100	0	−100*	1
	11	48	25	−23*	4
	19	10	79	69*	12
	33	37	38	1*	7
D1 P D2	15	0	0	−15	5
	49	11	12	−38	3
	38	77	28	39	8
	1	0	36	−1	6
	0	53	.	53	9
	9	29	71	20	7
P D1 D2	87	2	38	−85*	1
	51	1	10	−50*	2
	1	.	.	.	
	85	53	.	−32*	4

Where a value was missing 2 hours after treatment on a given day but at least one post-treatment VAS was recorded, then the last available value has been used for the 2 hour reading. (See Section 9.6 for a discussion of problems with missing data.) Where this has happened the value substituted has been marked '+'. Where the patient has discontinued the trial or where no post-treatment value at all has been recorded a '.' has been entered. The second point to note is that the data have been arranged in a particular way to facilitate the analysis of the comparison of *D1* and *P*.

The six sequences have been arranged in pairs. Within each pair it will be found that *D1* and *P* occur in the same periods. For the first sequence in a pair *D1* occurs in the earlier position and *P* in the later position and for the second sequence the position is reversed. If the results for *D2* are ignored it then follows that each of the pairs of sequences forms a standard *AB*/*BA* cross-over, with $D1 = A$ and $P = B$. The analysis now consists of two stages: first performing a *Wilcoxon–Mann–Whitney rank sum* test as we did in Section 4.3.8 (Koch, 1972) on each of the three *AB*/*BA* cross-overs formed by pairing the sequences and secondly combining the results in a way we shall discuss in due course.

The preliminary calculations have already been performed in Table 6.1. A difference in the VAS score has been calculated for each patient. For the first sequence in a pair this has been calculated as $P - D1$, for the second as $D1 - P$. For a given pair of sequences the difference thus incorporates an identical trend effect should there be one. (For example for the first pair the difference between period 3 and period 1 is calculated.) Since, however, the treatment effects are reversed for the second sequence pair compared to the first, a difference between treatments will be reflected by a difference between sequences in a pair. To perform the Wilcoxon–Mann–Whitney test, therefore, we calculate ranks within a pair of sequences. These are given in the final column of the table.

We shall illustrate the further working of this example using the Wilcoxon form of the rank sum statistic. Further calculations are given in Table 6.2, which gives for each of the first sequences in a pair of sequences, the rank sum as well as the expected value and variance under the null hypothesis calculated according to the formula (4.12). Also given is the *Z* statistic calculated as $\{W - E(W)\}/SE(W)$ as well as the form with continuity correction for which the difference between *W* and $E(W)$ has been reduced by $\frac{1}{2}$.

These individual *Z* statistics, calculated on the same basis as we did for Example 4.1 in Section 4.3.9, are of interest in looking at the individual contribution of the pairs of sequence to the overall probability calculation we shall perform. It is noticeable immediately that each is positive, reflecting the higher pain scores on average under placebo and that the value for the third sequence pair alone, using the continuity corrected form, is nearly 'significant' as the 5% level two sided, since the critical values for such a test using the Normal distribution are −1.96 and +1.96. In fact an exact calculation of the probability shows that the result is 'significant'. The ranks in the sequence with three patients are 1, 2, 4. No other set of three ranks gives the same score and

Table 6.2 Various calculations for the analysis of Example 6.1.

	Sequence pair		
	D1 D2 P/P D2 D1	D2 D1 P/D2 P D1	D1 P D2/P D1 D2
Wilcoxon rank sum, W, for first sequence	26	47	38
Patients in first sequence (m)	4	6	6
Patients in second sequence (m)	6	6	3
Expected sum, $E(W)$	22	39	30
Variance of sum, var (W)	22	39	15
Standard error of sum, $SE(W)$	4.690	6.245	3.873
Z statistic	0.853	1.281	2.066
continuity corrected Z statistic	0.746	1.201	1.936

the only lower score is given by 1, 2, 3. Altogether, however, there are $9!/(6!3!) = 84$ ways of allocating 9 patients into a group of 6 and one of 3, hence the P value, one-sided, is $2/84 = 0.0238$ and two-sided it is 0.048.

We are not much interested, however, in the results of the individual Wilcoxon tests but in a combination of the results. For this reason it is important to make sure, in calculating the Wilcoxon rank sums, that these are always formed in the same way. For each pair of sequences in Table 6.2 we have always calculated the sum of the sequence for which the VAS under P was subtracted from the value under $D1$. This was a purely arbitrary decison and we could have done it the other way round. What is essential, however, is to use the same convention for each sequence.

A simple way of combining the individual rank sums is to add them up. If we do this we obtain a total rank sum of 111. It is theoretically possible, but it would involve much tedious labour, to count all the possible ways in which this rank sum could have been reached, and to proceed to an exact probability calculation. We shall not discuss how this might be done but instead illustrate the large-sample approximation. By our rules for linear combinations of Section 2.2.1 the expected value of this sum is simply the sum of the individual expected values and thus equals $22 + 39 + 30 = 91$. Similarly the variance of this sum is the sum of the variances $22 + 39 + 15 = 76$. Taking the square root we obtain a standard error of 8.7178. From this we may calculate the Z statistic as $(111 - 91)/8.7178 = 2.294$ or, applying a continuity correction, as $(|111 - 91| - \frac{1}{2})/8.7178 = 2.237$. Using the former figure we may calculate from statistical tables of the standard Normal distribution that the P value, two-sided, is 0.022; the latter figure yields a P value of 0.025. (Thus the continuity correction makes hardly any difference to the final result.)

Table 6.3 (Example 6.1) Calculations for a weighted Wilcoxon test.

Sequence pair	Wilcoxon statistic	Expected value	Variance	Weight	Weighted statistic	Weighted exp. value	Weighted variance
a	*b*	*c*	*d*	*e*	$e \times b$	$e \times c$	$e^2 \times d$
D1 D2 P *P D2 D1*	26	22	22	1/11	2.3636	2	0.1818
D2 D1 P *D2 P D1*	47	39	39	1/13	3.6154	3	0.2308
D1 P D2 *P D1 D2*	38	30	15	1/10	3.8000	3	0.1500
Total					9.7790	8	0.5626

Standard error = $\sqrt{0}.5626 = 0.7501$
$Z = (9.779 - 8)/0.7501 = 2.372$
P value = 0.018

A refinement of the above calculation, originally proposed by van Elteren (1960), is to use a weighted sum which leads to a more powerful test (Lehmann, 1975, pp. 132–41). Here we multiply each rank sum by the factor $1/(m + n + 1)$ before combining in one total. We use our formulae from 2.2.1 to obtain the expected value of the total and its variance. The calculations are set out in Table 6.3. It will be seen that the value of the Z statistic, of 2.372, is little different from that of 2.294 obtained previously. (The application of a continuity correction is not so simple a matter for the weighted sum and we shall not attempt to apply it here.)

6.3.1 Analysis using SAS®

To perform an analysis using SAS® we must first of all obtain the data in an appropriate form. For the comparison of *P* and *D1* for Example 6.1 we should need a data set with three variables and 31 observations: one observation for each of the 31 patients. The three variables might be as follows. For each patient we record the *BLOCK* defining the pair of sequences to which the patient belongs. *BLOCK* would then be a categorical variable with three values: say I, II and III. We should also record for each patient whether he belongs to the first *SEQ*uence within a *BLOCK*. *SEQ* is thus a categorical variable with two values: say *X* and *Y*. Finally we record the actual *DIFF*erence in VAS scores as defined in Table 6.1.

It is useful to have the data sorted by *BLOCK* and *DIFF* and if this is not already done this may be achieved using the following code:

```
proc sort;
  by BLOCK DIFF;
run;
```

Ranks may be obtained using the following code:

```
proc ranks;
  by BLOCK;
  var DIFF;
  ranks RDIFF;
run;
```

This creates a new variable *RDIFF* which consists of the ranks obtained by ranking *DIFF* within *BLOCK*. (This step is not necessary to the analysis which follows but may be useful if one wishes to inspect the data.)

The code

```
proc nparlway wilcoxon;
  by BLOCK;
  var DIFF;
  class SEQ;
run;
```

produces (with one slight difference) the three Wilcoxon analyses given in Table 3.2. The difference is that SAS® uses a more sophisticated formula to calculate the variance of the rank sum than the one we have used (4.12). The SAS® formula makes an adjustment for tied ranks. For the first pair of sequences two ranks are tied and the one difference to Table 6.2 is that SAS® gives a value for the variance of the rank sum for BLOCK = I of 21.87 as opposed to 22.

The rest of the calculation can then be completed by hand along the lines given in Table 6.3. As might be expected the final result is little different to that obtained in Table 6.3. For the first method of pooling, namely simple addition, the variance is 75.87 as opposed to 76; for the van Elteren procedure the variance is 0.5615 instead of 0.5626. These two procedures yield Z statistics (without continuity correction) of 2.296 and 2.374 respectively.

An ingenious approach which permits the calculation of the whole procedure through SAS® has been proposed by Koch and Edwards (1988). They point to a connection between the van Elteren test (van Elteren, 1960) and the extended Mantel–Haenszel procedure (Mantel, 1963). A discussion of this connection is beyond the scope of this book but the reader is referred to the important article by Koch and Edwards (1988) for further information. We limit ourselves to producing the necessary SAS® code and making a few observations.

The code is as follows:

```
proc freq;
  table BLOCK* SEQ* DIFF/noprint cmh score = modridit;
run;
```

The following points may be noted.

- *proc freq* is a powerful procedure within SAS® which has a great number of different uses.
- The order in which the three variables are quoted in the *table* statement is important and should not be altered.
- *noprint* is an option used to suppress unwanted output.
- *cmh* is an option requesting the calculation of what are referred to in SAS® as Cochran–Mantel–Haenszel statistics which are used for the extended Mantel–Haenszel test.
- *score = modridit* causes SAS® to calculate within-stratum (i.e. within-BLOCK) standardized midrank scores. Such scores are referred to within SAS® as *modified ridits*. This transformation is necessary in order for the calculation to be performed correctly and is discussed in Koch and Edwards (1988).

If this procedure is applied to the data of Example 6.1 the output will be as in Table 6.4. The relevant statistic is that given in row 2. Its value is 5.637 and it has approximately a chi-square distribution with one degree of freedom. Since the square root of a chi-square with one degree of freedom is a standard Normal (see Section 2.2.5) and since we have said that the extended Mantel–Haenszel is

Table 6.4 (Example 6.1) SAS® output using *proc freq*.

```
                 SUMMARY STATISTICS FOR SEQ BY DIFF
                        CONTROLLING FOR BLOCK
     Cochran–Mantel–Haenszel Statistics (Modified Ridit Scores)
Statistic      Alternative Hypothesis      DF      Value      Prob

    1           Nonzero Correlation         1      5.604     0.018
    2          Row Mean Scores Differ       1      5.637     0.018
    3           General Association        28

At least 1 statistic not computed—singular covariance matrix.
Total Sample Size = 31
```

equivalent to the van Elteren procedure, it is gratifying to find that the square root of 5.637, namely 2.374, is the same as the value of the Z statistic which we obtained above.

6.3.2 Discussion

The method we have outlined relies on a technique which may be applied generally to a variety of different problems.

The first step is to match sequences so that for a given contrast we may reduce the design to a series of *AB*/*BA* cross-overs which we analyse separately. For example, when we wish to compare *D2* and *P* we shall have to match sequence *D1 D2 P* with *D1 P D2*, sequence *D2 D1 P* with *P D1 D2* and *P D2 D1* with *D2 P D1*. The same technique can be used for designs with four treatments, but we have to have chosen an appropriate set of sequences. If we use a single Latin square, then, as we noted in Section 5.2, any of the designs from the first set of Table 5.1 is suitable. For example, if we use square *I5*, then to compare *A* and *B* we match sequence *ACDB* with *BDCA* and *CABD* with *DBAC*.

The second step is to combine the statistics so formed in some appropriate way. We illustrated two approaches for Example 6.1: simple addition and weighting using the van Elteren (1960) approach. There was very little difference between the two. Had we had the same number of patients in each sequence pair the two approaches would have been identical. In practice we shall usually choose to arrange things so that this will very nearly be the case, and that being so there seems to be little harm in using the first and simpler method. (If, however, the analysis is performed using SAS® and *proc freq* as outlined in Section 6.3.1, then this corresponds to the more sophisticated second approach.)

The feature which is particularly unpleasing about this sort of analysis is the rather arbitrary matching of sequences in pairs. For designs such as the three-period three-treatment cross-over in six sequences this imposes a degree of constraint which goes beyond that which is simply necessary to obtain a balance for periods. Nevertheless, it provides a practical solution where it is feared that the data will show important departures from Normality.

For further advice as to how the exact calculation may be made for small samples the reader should consult Lehmann (1975).

6.4 HODGES–LEHMANN TYPE ESTIMATORS*

We may define statistics for the three-treatment, three-period cross-over in six sequences which are similar to the Hodges–Lehmann estimator we defined for the AB/BA cross-over in Section 4.3.8. One possibility is to calculate all the possible pairwise means formed by taking one basic estimator from each of the two sequences in a sequence pair. (This is of course equivalent to taking the semi-pairwise difference of the period differences.) In Example 6.1 this gives $4 \times 6 = 24$ means for the first sequence pair, $6 \times 6 = 36$ means for the second sequence pair and $6 \times 3 = 18$ means for the third sequence pair, making 78 means altogether. There are then various ways that a single statistic might be produced from these. One possibility is to use the median of the 78 means. Another possibility is to form a further set of means taken by choosing one basic estimator from each sequence pair and averaging the three values. This will produce $24 \times 36 \times 18 = 15\,552$ means. We can then take the median of these values as our estimator. This latter estimator may be equivalently viewed as the median of the means formed by taking one basic estimator from each of the six treatment sequences.

An estimator formed in this way would have the advantage of a certain robustness. If there were a few freak observations then any estimator, such as an ordinary least squares estimator, which was a weighted sum of observations would incorporate these unusual values. Some of the means of the six observations formed by taking a basic estimator from each of the six sequences would, of course, also reflect such freak values, but most would not. By taking a median as a final step the influence of the freak values would be largely eliminated.

Figure 6.1 is a cumulative plot of the 78 semi-pairwise differences from Example 6.1. The median of these values is 12.5 mm VAS.

6.5 A STRATIFIED PERIOD ADJUSTED SIGN TEST

We now describe an approach which is a generalization of the procedure of Section 4.3.6.

Consider Table 6.5, which is based on Table 6.1 with the object, as before, of comparing P and $D1$. Instead of recording the difference in VAS, as was done in Table 6.1, only the sign of the difference is noted. This means that for a given pair of sequences a given patient can be cross-classified in one of four ways

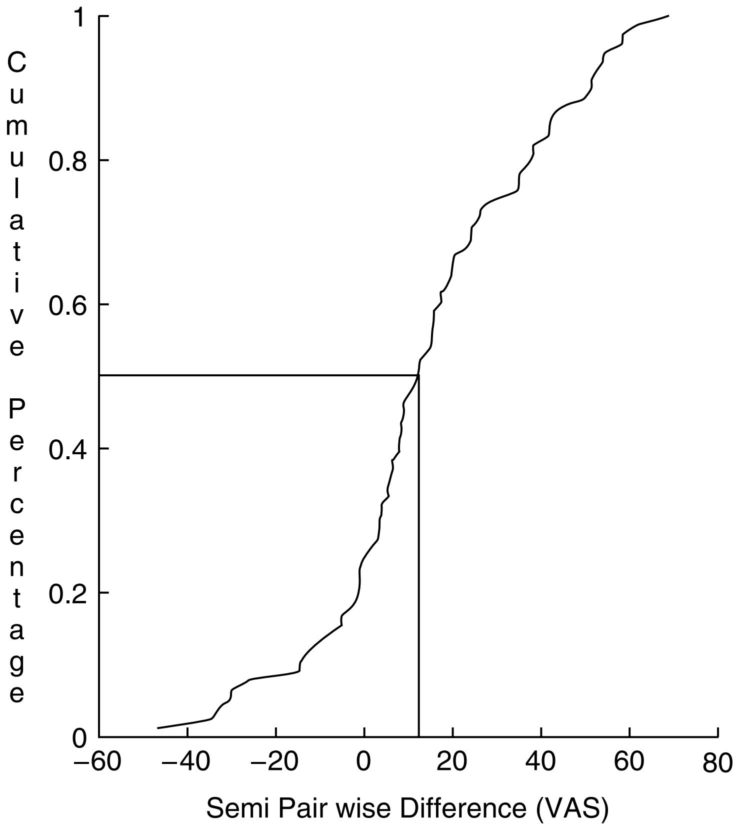


Figure 6.1 (Example 6.1) Construction of a Hodges–Lehmann type estimator: cumulative plot of semi-pairwise differences and median.

Table 6.5 (Example 6.1) Patients classified in three 2 × 2 contingency tables for the purpose of comparing *P* and *D1* according to sequence pair, earlier treatment and sign of period difference.

Sequence pair		Sign of Period Difference		
*D*1 *D*2 *P*/*P DE D*1		−	+	
earlier treatment	*D*1	0	4	4
	P	3	3	6
		3	7	10
*D*2 *D*1 *P*/*D*2 *P D*1				
earlier treatment	*D*1	1	4	5
	P	4	2	6
		5	6	11
*D*1 *P D*2/*P D*1 *D*2				
earlier treatment	*D*1	3	3	6
	P	3	0	3
		6	3	9

depending on whether he was given *D1* before *P* or vice versa and depending on whether the difference between the later treatment and the earlier was positive or negative. (One patient has a difference of zero and his result is discarded as uninformative. Table 6.5 shows the results of 30 patients, not 31.) Thus for each of the three pairs of sequences we can produce a 2 × 2 contingency table. Each of these could be analysed using the arguments and methods of Section 4.3.6.

As in Section 6.2, however, what we should like is an overall statement for the trial as a whole. A useful general approach, which may be used for combining contingency tables, is the Mantel–Haenszel procedure (Mantel and Haenszel, 1959; Somes and O'Brien, 1985) which we now illustrate.

The procedure involves our using a single corresponding cell from each of the three contingency tables. Let us nominate the top right-hand cell. This cell records the number of patients within each sequence pair for whom (a) *D1* was given before *P* and (b) the VAS for pain was higher in the second of the two periods being compared. Thus the more patients in these three cells the more effective is *D1*. From the three cells we form the total $4 + 4 + 3 = 11$. We also calculate the expected number for this total using the margins of each table. Thus for the first table the expected number for the cell is $4 \times 7/10 = 2.8$. For the other two tables the values are 2.727 and 2 respectively. Thus the total expected number is 7.527.

We can also calculate the variance of the observed sum as the sum of the variances for each table. The variance may be calculated as follows (Somes and O'Brien, 1985). First, we form the product of each of the four margins. Then we divide this by the product of the square of total for the table and the total minus one. Thus for the first table we obtain a variance of $4 \times 6 \times 7 \times 3/(10^2 \times 9) = 0.56$. The variances for the other two tables are 0.7438 and 0.5 respectively. Since the obserations in the three tables may be regarded as being independent of each other the variance of the sum is simply the sum of the variances and so we obtain a variance of our total of $0.56 + 0.7438 + 0.5 = 1.8038$.

Finally, we calculate the Mantel–Haenszel statistic as the ratio of the square of the difference between observed and expected totals (reduced by 0.5 as a continuity correction) to the variance. Thus for this example we have:

$$MH = (|11 - 7.527| - 0.5)^2/1.8038 = 4.90.$$

The Mantel–Haenszel statistic is approximately distributed as a chi-square with one degree of freedom, so that by using the tables of the chi-square, or alternatively taking the square root of the statistic and referring it to tables of the Normal distribution, we may calculate the *Z* statistic, which is 2.21, and hence the *P* value, which is 0.027. This may be compared to the results we obtained in Section 6.2 above.

If *SDIFF* is a variable taking the value −1 or 1 according to the sign of the period difference in Table 6.1, then the following simple code in SAS® produces the analysis:

```
proc freq:
  table BLOCK* SEQ* SDIFF/cmh;
run;
```

This will print contingency tables of the sort produced in Table 6.5 and also calculate the Mantel–Haenszel statistics. The value produced is 6.686. This differs from that we calculated by hand because SAS® does not make a continuity correction. Thus the calculation is equivalent to $(11 - 7.527)^2/1.8038 = 6.69$.

6.6 BINARY DATA

The method of Ezzet and Whitehead (1992) for analysing binary data, which we discussed in Chapter 4, generalizes very naturally to designs with three or more periods. Procedure *proc nlmixed* of SAS®, which we also used in Chapter 4, can be used to analyse such cases. We shall now introduce and analyse a rather complex example to illustrate the power of this approach.

Example 6.2 In a three-period design comparing three types of condom, volunteer couples were randomized to one of the six possible sequences of one type each (Bounds and Guillebaud, 2002). They were provided with packs of six of each of the three sorts of condom. Each pack was to be used sequentially until all the condoms had been used. For subsequent acts of sexual intercourse the couple was to proceed with the next variety of condom and so forth. Amongst various features recorded by the couple was whether breakage occurred. Data relating to this are recorded in Table 6.6.

Remark The data have been sorted in a way that helps to highlight the considerable degree to which data are missing in the trial. Only 15 of the 36 couples listed here carried out the trial as originally planned. (There were some who agreed to take part but recorded no values.) A further 13, although not using all condoms, did at least try each sort, so that in total there are 28 couples in this category. Four couples tried two only and four couples tried one only. We shall assume in what follows that the data are 'missing completely at random' (Little and Rubin, 1983).

We now proceed to analyse these data by adapting the method of Ezzet and Whitehead (1992). We assume that a SAS® data set, *ONE*, has been created with rows corresponding to couples × periods. There are thus $(28 \times 3) + (4 \times 2) + (4 \times 1) = 96$ rows of data. Each row records the couple's identifier (*PON*), the number of condoms used in that period, *N*, and the number of breakages, *BREAK*. In addition, two dummy variables are used for period. *PERIOD2* is coded one if the results apply to period 2 and zero otherwise. *PERIOD3* is coded one if the results apply to period 3 and zero otherwise. There are two analogous treatment dummies for the type of condom used, *TREATB* and *TREATC*. Breakages are assumed to be conditionally binomial with the probability of breakage on the logit scale being a function of the relevant period and treatment fixed effects and the random couple effect,

Table 6.6 (Example 6.2) Breakages in three periods for 36 couples trying up to six condoms of each of three types.

		Period 1		Period 2		Period 3			
Sequence	Couple	#Break	*n*	#Break	*n*	#Break	*n*	Condoms	Periods
ABC	5	0	6	3	6	1	6	18	3
ABC	17	0	6	0	6	0	6	18	3
ABC	35	0	6	0	6	0	6	18	3
ACB	6	0	6	0	6	0	6	18	3
ACB	36	0	6	0	6	0	6	18	3
ACB	39	0	6	0	6	0	6	18	3
BAC	4	0	6	0	6	0	6	18	3
BAC	8	0	6	0	6	0	6	18	3
BCA	10	0	6	0	6	0	6	18	3
BCA	40	0	6	0	6	0	6	18	3
BCA	43	0	6	0	6	0	6	18	3
CAB	26	0	6	0	6	0	6	18	3
CAB	42	0	6	1	6	0	6	18	3
CBA	3	2	6	2	6	0	6	18	3
CBA	29	0	6	0	6	0	6	18	3
BCA	13	0	6	0	5	0	6	17	3
BCA	32	0	6	0	6	1	5	17	3
BAC	14	0	6	0	3	0	6	15	3
BAC	30	0	6	0	6	0	3	15	3
BCA	20	0	3	1	6	3	6	15	3
CAB	12	0	6	0	3	0	6	15	3
CAB	31	0	6	1	6	0	3	15	3
BCA	2	0	6	0	6	0	2	14	3
CAB	1	0	6	0	2	0	6	14	3
CBA	11	0	6	0	6	0	2	14	3
CBA	19	0	6	1	2	0	6	14	3
CAB	15	0	5	0	4	0	3	12	3
ABC	7	0	1	0	1	0	1	3	3
BAC	38	0	6	0	6	0	0	12	2
CBA	41	1	6	0	6	0	0	12	2
BAC	22	0	4	0	5	0	0	9	2
CBA	33	0	4	0	3	0	0	7	2
ACB	16	0	6	0	0	0	0	6	1
CBA	18	0	6	0	0	0	0	6	1
ACB	25	0	4	0	0	0	0	4	1
BAC	34	0	3	0	0	0	0	3	1

which is assumed to be Normally distributed with mean zero and unknown variance *s2u*. We take the opportunity here to illustrate an alternative approach to handling the variance whereby modelling is done primarily in terms of *gamma = log(s2u)* because this is a more useful scale for calculating the standard error of the variance.

Analysis is achieved with the following code:

```
proc nlmixed data=one;
  parms beta0=-3 tauB=0.0 tauC=0.0 pi2=0 pi3=0 gamma=0;
  pred=beta0 + tauB* TREATB + tauC* TREATC + pi2* PERIOD2
  +pi3* PERIOD3 +u;
  p=exp(pred)/(1+exp(pred));
  s2u=exp(gamma);
  model BREAK~binomial(N,p);
  random u∼normal(0,s2u) subject=PON;
run;
```

The resulting output includes the following:

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
beta0	−5.3008	0.9881	35	−5.36	<.0001	0.05	−7.3067	−3.2950
tauB	−0.1051	0.7086	35	−0.15	0.8829	0.05	−1.5437	1.3335
tauC	−0.2898	0.6997	35	−0.41	0.6813	0.05	−1.7102	1.1306
pi2	1.2933	0.7237	35	1.79	0.0826	0.05	−0.1759	2.7626
pi3	0.6345	0.8049	35	0.79	0.4358	0.05	−0.9996	2.2686
gamma	1.1262	0.6320	35	1.78	0.0834	0.05	−0.1567	2.4092

The parameterization used expresses the treatment effects with respect to the effect of condom A. Thus *tauB* and *tauC* are estimates on the log-odds scale of the difference in probability of breakage for condoms B and C compared to A. These differences are not significant. If we wish to estimate the random effect variance we can do this by exponentiation of the estimate of gamma. Similarly, confidence limits can be produced. Thus our estimate is $s2u = \hat{\sigma}^2 = \exp(1.1262) = 3.1$ and the confidence limits are and $\exp(-0.1567) = 0.9$ and $\exp(2.4092) = 11.1$.

Remark We have repeated measures on each couple in a sense that we did not in the analogous analysis of binary data in Section 4.4.4. We now have more than two periods, but we also have repeated measures within each period on each couple. The strong assumption that we make is of independence of breakage given the period, couple and type of condom. In particular, we assume that there is no random effect of condom (in the sense of couple by condom interaction) and that it is not the case that certain couples have a particular raised risk of breakage with certain types of condom. A robust alternative would be to summarize the observations for a given patient and period and use a rank test, but this could not easily be adapted to deal with the large amount of missing data.

6.7 OTHER ANALYSES

Any analysis which can be reduced to a series of 2 × 2 contingency tables can be handled using the approach via *Mantel–Haenszel statistics* described in Section 6.5. Thus as an alternative to the stratified period adjusted sign test we could use a stratified *Brown–Mood median test* (Section 4.3.8). Where we have binary outcomes we can use a stratified *Mainland–Gart test* (Section 4.4.2). The only differences lie in the steps to produce the contingency tables in the first place.

For a stratified version of Prescott's test for binary outcomes (Section 4.4.3) the extended Mantel–Haenszel procedure of Section 6.3.1 could be used. This could also be used to analyse change scores produced from ordered categorical variables as in Section 4.5. An alternative would be to use *logistic regression* for ordered categorical variables fitting the sequence pair as a block.

APPENDIX 6.1 ANALYSIS WITH GENSTAT®

GenStat®, and for that matter S-Plus®, can be used to apply logistic regression to the analysis of binary change scores to perform the sort of analysis suggested at the end of Section 6.6, fitting sequence pair as a block. The extension to this is straightforward in principle and will not be illustrated here.

As far as I am aware, GenStat® does not provide the means of carrying out the Mantel–Haenszel test. However, S-Plus® does, and this will be illustrated below.

APPENDIX 6.2 ANALYSIS WITH S-PLUS®

Although S-Plus® does not contain a routine for the van Elteren test, it can be used to perform the Mantel–Haenszel test. To perform this analysis using S-Plus, we may use the following code:

```
#Analysis of Table 6.5 Cross-over Trials in Clinical Research
data.mh<-array(c(0,3,4,3,1,4,4,2,3,3,3,0),dim=c(2,2,3))
data.mh #list data
mantelhaen.test(data.mh) #carry out MH test
```

Here *data.mh* is an object which contains a three–dimensional array corresponding to Table 6.5. The first argument of the function *array* lists the 12 frequencies and the second argument gives the dimensions of the array. The test itself is performed by the function *mantelhaen.test*. The output gives:

```
          Mantel-Haenszel chi-square test with continuity correction

    data: data.mh
    Mantel–Haenszel chi-square = 4.8992, df = 1, p value = 0.0269
```

This is the same answer as obtained with SAS®.

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

7

Some Special Designs

7.1 THE SCOPE OF THIS CHAPTER

In this chapter we shall consider some special types of cross-over trials. These will be of four kinds: factorial designs, incomplete block designs, '*n* of 1' designs and bioequivalence studies. We shall cover the first two in some detail but the second two only briefly.

Factorial designs occur whenever the treatments which are given to patients in particular periods may be expressed in terms of combinations of more 'primitive' factors. We discussed their use in investigating combination therapies in Section 5.1.3 but there are other applications for which such designs may be useful. *Incomplete block designs* may be used if we wish to investigate a number of treatments using fewer periods so that each patient receives a subset of the treatments. Trials where a single patient is treated are known as *n of 1* designs. The patient is given each treatment (usually there are only two) a number of times. Effectively a controlled clinical trial is run using one patient only with *episodes* providing the replication. In *bioequivalence studies* two formulations of the same treatment are compared in terms of pharmacokinetics to see if they are equivalent.

7.2 FACTORIAL DESIGNS

We introduce this topic by considering an example.

Example 7.1 This concerns a single-dose cross-over design in four periods with the object of comparing two formulations of formoterol aerosol (a suspension formulation and a solution formulation) given at two doses ($12\,\mu$g and $24\,\mu$g). For this design we can define two *factors*: formulation, with *levels* suspension and solution, and dose, with levels high and low. The suspension and solution canisters are indistinguishable and so do not require dummies to maintain blinding as long as the comparison is at one dose only. However, the canisters of suspension and solution formoterol deliver $12\,\mu$g per puff and thus to take $24\,\mu$g two puffs are needed. To maintain blinding, for each 'treatment' the patient was provided with two canisters and instructed to take one puff from each. For the two treatments at 24 μg the two canisters were filled identically with suspension or solution (as the case was); for the two 12 μg treatments one of the canisters was a matching placebo. Thus, the four treatments were:

A: one puff 12 μg formoterol suspension + one puff placebo = *SUS12*;
B: one puff 12 μg formoterol suspension + one puff formoterol suspension = *SUS24*;
C: one puff 12 μg formoterol solution + one puff placebo = *SOL12*;
D: one puff 12 μg formoterol solution + one puff formoterol solution = *SOL24*.

There were 16 patients, nine men and seven women, aged between 30 and 63 years. They were given their treatment at 8.00 a.m. of a particular treatment day and observed for 14 hours. The interval between test days varied from 4 to 16 days. Various lung function measures were obtained throughout the day including FEV_1, *PEF* and specific airways resistance (*sRAW*).

Remark The purpose of this study was to demonstrate the equivalence of the two formulations. Studies with this aim are frequently carried out by manufacturers to allow them to use the experience with one formulation for evaluating another. Manufacturers of generic drugs also use them to claim equivalence to brand name drugs. The usual approach is to compare the serum concentration of the two formulations of the drug in the blood using a so-called 'bioequivalence' study. In this example, however, the formulations are inhaled, and this approach is not relevant. Instead a parallel assay has been performed (Finney, 1978; Govindarajulu, 2001). This particular aspect of the study will not be addressed here. Bioequivalence is discussed later in this chapter.

The 16 patients were allocated in blocks of four to one of four sequences of a Latin square. That is to say, for each set of four patients recruited to the trial each of the four sequences was used by one patient. The square used was square number III2 of Table 5.1, namely:

$$\begin{array}{cccc} A & B & D & C \\ B & C & A & D \\ C & D & B & A \\ D & A & C & B \end{array}$$

Remark This design (for which I bear some responsibility) has two features I do not like. First, it was deliberately chosen to be a Williams square (each treatment follows each other treatment exactly once) and secondly, the block size is extremely small. In the analysis which follows no use will be made of either of these features and the trial will be treated as if patients had been allocated at random in equal numbers to the four sequences of a square chosen at random. (A particular analysis commonly associated with Williams squares will be illustrated using this example in Chapter 10.)

7.2.1 Analysis using basic estimators

Table 7.1 presents the FEV_1 results 6 hours after treatment for each patient for each of the four treatment days as well as three basic estimators. The first, labelled *Form.* (for formulation), gives the difference between the mean reading for suspension (i.e. over both doses) and the mean reading for solution. The second, labelled *Dose*, gives the difference between the mean reading at 24 μg (over both formulations) and the mean reading at 12 μg. In the statistical vocabulary of factorial designs these two contrasts measure *main effects*. The third, labelled *Int.* for *interaction*, gives the difference between the mean of the suspension 24 μg and solution 12 μg reading and the mean of the suspension 12 μg and solution 24 μg reading.

This last contrast is worth explaining further. It can be regarded as measuring either the difference between formulations of the dose effect or the difference between doses of the formulation effect. For suspension we can represent the difference between the higher and lower dose, symbolically as

$$SUS24 - SUS12,$$

whereas for solution it is

$$SOL24 - SOL12.$$

Subtracting the latter from the former and dividing by 2 (this last step is a convenient convention which is usually adhered to in analysing factorial designs) we obtain

$$(SUS24 + SOL12)/2 - (SUS12 + SOL24)/2,$$

which thus corresponds to the interaction of dose and formulation.

Once the contrasts of interest have been established the calculations may proceed as outlined for the basic estimator approach in Section 5.4.1. First, the estimators are averaged within sequences and then the results are in turn averaged over sequences. The results are given at the foot of Table 7.1 where it may be seen that the estimated difference due to formulations is $-0.12\ \ell$, whereas that due to dose is $0.08\ \ell$. The estimate for the interactive effect is $-0.04\ \ell$.

Table 7.2 presents the further calculations necessary to establish 95% confidence intervals for the treatment estimates. Corrected sums of squares are obtained for each contrast for each sequence. They are summed over sequences and then divided by the total degrees of freedom (12 in this case) to obtain an estimate of the variance of the individual basic estimators. Using the rules for

Table 7.1 (Example 7.1) Four-period cross-over trial comparing two formulations of formoterol at two doses. FEV_1 measurements 6 hours after treatment.

		FEV_1						
		Treatment						
		A	*B*	*C*	*D*		Basic estimator	
Sequence	Patient	*SUS12*	*SUS24*	*SOL12*	*SOL24*	Form.	Dose.	Int.
ABDC	3	2.7	1.7	2.2	2.6	−0.2	−0.3	−0.7
	5	2.5	2.4	2.4	2.4	0.05	−0.05	−0.05
	12	2.6	2.5	2.4	2.5	0.1	0.0	−0.1
	13	2.0	2.2	2.6	2.6	−0.5	0.1	0.1
	mean	2.45	2.2	2.4	2.525	−0.1375	−0.0625	−0.1875
BCAD	4	3.7	3.6	3.7	3.6	0.0	−0.1	0.0
	6	0.9	1.4	2.4	1.1	−0.6	−0.4	0.9
	10	2.5	2.6	2.6	2.4	0.05	−0.05	0.15
	16	2.0	2.5	2.2	2.7	−0.2	0.5	0.0
	mean	2.275	2.525	2.725	2.45	−0.1875	−0.0125	0.2625
CDBA	2	1.3	1.3	1.4	1.3	−0.05	−0.05	0.05
	8	2.2	2.2	2.3	2.3	−0.1	0.0	0.0
	9	1.8	1.9	1.0	2.7	0.0	0.9	−0.8
	14	1.9	2.2	2.2	2.1	−0.1	0.1	0.2
	mean	1.8	1.9	1.725	2.1	−0.0625	0.2375	−0.1375
DACB	1	1.7	1.7	1.6	2.0	−0.1	0.2	−0.2
	7	2.2	1.9	1.8	2.6	−0.15	0.25	−0.55
	11	3.3	3.7	3.6	3.3	0.05	0.05	0.35
	15	2.2	2.3	2.4	2.5	−0.2	0.1	0.0
	mean	2.35	2.4	2.35	2.6	−0.1	0.15	−0.1
	mean	2.21875	2.25625	2.3	2.41875	−0.1219	0.0781	−0.0406

Table 7.2 (Example 7.1) Various calculations.

	Corrected sums of squares (ℓ^2) Contrast		
Sequence	Formulation	Dose	Interaction
ABDC	0.226 875	0.086 875	0.371 875
BCAD	0.261 875	0.421 875	0.556 875
CDBA	0.006 875	0.596 875	0.606 875
DACB	0.035 000	0.025 000	0.425 000
Total	0.530 625	1.130 625	1.960 625
variance of basic estimator (ℓ^2)	0.044 219	0.094 219	0.163 385
variance of treatment est (ℓ^2)	0.002 764	0.005 888 7	0.010 211 6
standard error (ℓ)	0.0526	0.0767	0.1011
$\lvert t \rvert$	2.32	1.02	0.40
$t_{12,\,0.025} = 2.179$			
Confidence intervals	−0.24 to −0.01 ℓ	−0.09 to 0.25 ℓ	−0.26 to 0.18 ℓ

linear combinations given in Section 2.2.1 we see that the repeated averaging means that the treatment estimate has variance $\sigma^2(\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4})/16 = \sigma^2/16$, where σ^2 is the variance of the individual basic estimators. Substituting the relevant estimates of σ^2 for σ^2, dividing by 16 and taking the square root gives the estimated standard errors. These are then multiplied by 2.179, which is the value cutting off the top $2\frac{1}{2}\%$ of a t distribution with 12 degrees of freedom. The product is then added and subtracted from the treatment estimate.

Remark From the fact that the 95% confidence interval for the difference between the two formulations does not overlap zero, and also (equivalently) because the ratio of the estimate to its standard error exceeds the critical value for a two-sided test, it may be deduced from Table 7.2 that the formulations are by a standard statistical convention 'significantly' different, although the observed effect is very small. On the other hand the fact that of the two confidence limits the one which is furthest away from zero is less than a $\frac{1}{4}\,\ell$ away in FEV_1, which is quite a modest amount, suggests (as far as this measure is concerned) that the formulations might be clinically equivalent. This is in any case, however, always a matter of rather arbitrary judgement. Even if a difference of $\frac{1}{4}\,\ell$ had universal assent amongst physicians as constituting the minimum clinically relevant difference for FEV_1 the conclusion that the formulations were equivalent would still not be convincing due to a lack of evidence of a dose response. This trial is thus an experiment for which nothing much happened (at least as regards comparison of the treatments with each other using the 6 hour FEV_1 values). The results are difficult to interpret in isolation. A more convincing demonstration of equivalence would have been to have observed a strong dose response together with an absence of any important difference between formulations or of any interactive effect. The solution would be to express the difference between formulations in terms of the dose scale. However, the difference in doses observed is too slight to make this practical. With the benefit of hindsight it would have been preferable to use a wider dose range to compare the formulations and a larger sample size for the trial.

7.2.2 Analysis using ordinary least squares

A more conventional way to analyse this experiment would be to use ordinary least squares. This can be done quite simply using *proc glm* of SAS®. As usual, we first represent the data as a single outcome variable, say Y, consisting of $4 \times 16 = 64$ observations. We also create a variable *PATIENT* with 16 values, one called *PERIOD* with four values, I, II, III and IV and one called *TREAT* with four values A, B, C, D. For a given observation on a given patient a data line now records under each of the variable headings the FEV_1 reading, the patient it was taken from, the period it was measured in and the treatment he was then receiving. We may then use the following code:

```
proc glm;
  class TREAT PERIOD PATIENT;
  model Y = PATIENT PERIOD TREAT;
  estimate "formulation" TREAT 0.5 0.5 - 0.5 - 0.5;
  estimate "dose" TREAT -0.5 0.5 -0.5 0.5;
  estimate "interact" -0.5 0.5 0.5 -0.5;
run;
```

If the *estimate* statements are studied carefully then, bearing in mind that the order of the weights reflects the alphabetical orders of the treatment labels A, B, C and D and that these correspond to suspension 12μg, suspension 24μg, solution 12μg and solution 24μg, it will be seen that these produce the contrasts we have already estimated. Indeed, in this balanced design the factor estimates are identical using OLS or the basic estimator approach for amongst the SAS® output we find the following:

Source	DF	Sum of squares	Mean square	F value	Pr > F
Model	21	22.719 531 25	1.081 882 44	10.54	0.0001
Error	42	4.310 312 50	0.102 626 49		
Corrected Total	63	27.029 843 75			

Parameter	Estimate	T for H_0 Parameter = 0	Pr > $\|T\|$	Std error of estimate
Formulation	−0.121 875 00	−1.52	0.1356	0.080 088 42
Dose	0.078 125 00	0.98	0.3349	0.080 088 42
Interact	−0.040 625 00	−0.51	0.6146	0.080 088 42

The estimates are the same as we obtained before. It will be noted, however, that the standard errors are now identical to each other and hence not the same as those we obtained before. The reason is that these are based upon a sum of squares for error which has been pooled from the experiment as a whole. In fact the mean square error has been obtained by dividing the error sum of squares of $4.310\ 312\ 50\,\ell^2$ by its degrees of freedom, 42 to get $0.102\ 626\,\ell^2$. The further step of dividing by 16 and taking the square root of the result is identical to that undertaken above and will be found to produce the value of $0.080\ 088\ 42\,\ell$.

Remark Using this method of analysis the difference between formulations is not 'significant' and so there are not, by the conventional logic of the significance test, grounds for concluding that the formulations are different. On the other hand, the 95% confidence limits for the formulation effect are $-0.28\,\ell$ and $0.04\,\ell$ and so, using the same standard of $0.25\,\ell$ we proposed before, we cannot conclude that the treatments are clinically equivalent. Thus our conclusions should be exactly the opposite of those we reached in Section 7.2.1 before. This is, of course, disturbing. Often conclusions are robust to the method of analysis but not always. As was mentioned in Section 7.2.1, however, the main problem with this example is that we do not have a convincing dose-response and since we do not have a placebo either, it is difficult to know whether or not the trial was competent to find a difference if it existed. We shall pick this theme up again in Section 7.5.

7.3 INCOMPLETE BLOCK DESIGNS

There are practical limits to the number of treatments one can give a patient. The physician cannot expect unbounded cooperation. The longer he studies the patient the greater the risk that the patient may discontinue the trial prematurely. Also, if the total time under treatment for a given patient is increased it may delay the time in which a conclusion regarding the efficacy of the treatments may be reached (but see Chapter 9 for a discussion of this issue). If, therefore, it is considered desirable to study a number of treatments it may be necessary to use incomplete block designs which, as has been noted above, are designs where patients receive subsets of the total treatments in the trial.

Remark The design issues here are not as simple as is sometimes supposed. Unlike agriculture, where a growing season may constitute a natural treatment period, there is no such thing as a natural treatment period for a patient. Although for some diseases it may be a part of a clinical investigation to establish for how long a patient must be treated to effect a cure this is not the case where cross-over trials are used since these are, of course, only suitable for chronic disease. For the purposes of the clinical trial a judgement must be made as to how long he is to receive a given treatment for its effect to be capable of being estimated. Consequently, for example, where four treatments are to be investigated in one multiple-dose design one investigator may choose to give patients two treatments in two months in an incomplete blocks cross-over, another may choose to give them four treatments in four months and yet another may choose to give them four treatments in two months and so on (Senn and Lambrou, 1998). Furthermore, the time under treatment may be only a small part of the total time for which the trial runs, since patients are not recruited simultaneously and the time between recruitments may be considerable. Since to have the same efficiency as a cross-over with complete blocks an incomplete cross-over will have to have more patients, the latter may well take longer to run than the former. Where single-dose cross-over trials are concerned, however, and extensive investigations are carried out on a given treatment day, these treatment days constitute the major inconvenience which the patient suffers by being on the trial and there may be interest in keeping these to a minimum.

7.3.1 An example

Example 7.2 This concerns a double-blind placebo controlled cross-over trial designed to measure the onset of action of two doses of formoterol solution aerosol: 12 μg and 24 μg. For practical reasons it was decided that the patients could only be studied during four visits. Since each treatment day was to be preceded by a general medical evaluation this meant that only two treatment days were possible. It was decided to allocate the patients in equal numbers to one of the six possible sequences of two treatments at a time. Blindness was maintained using dummy loading and each patient used two aerosols at each of visits 2 and 4, taking one puff from each. The aerosols were matched and depending on whether both aerosols were formoterol solution 12 μg, both placebo or one of each, the patient received 24, 0 or 12 μg formoterol. The wash-out period between visits 2 and 3 was approximately one week.

The treatments were as follows:

$$F_{24} = \text{two puffs of formoterol } 12\,\mu\text{g solution aerosol;}$$
$$P = \text{two puffs of placebo;} \qquad .$$
$$F_{12} = \text{one puff of formoterol}12\,\mu\text{g solution aerosol} + \text{one puff of placebo}$$

Six sequences of administration were used: $F_{12}P$, PF_{12}, $F_{24}F_{12}$, $F_{12}F_{24}$, $F_{24}P$, PF_{24}, where (for example) PF_{12} means that the patient had placebo at visit 2 and formoterol 12 μg at visit 4. It was planned to allocate the 24 patients at random in equal numbers to the 6 sequences (i.e. 4 per sequence). (The block size was 6 and there were 2 centres. Both of these features will be ignored in the analysis.) An administrative error led to two reserve packs of medication being used for patients 23 and 24. One of these had the wrong sequence and the net result was that there were 5 patients allocated to sequence $F_{12}P$ and 3 to sequence $F_{12}F_{24}$.

Remark This sort of regrettable accident can always occur in a clinical trial. In this particular case the consequence is not serious. There is a slight loss in efficiency from the unbalanced allocation. The estimation of treatment contrasts becomes slightly more complex but since this is likely to be done in any case with the help of a computer package this latter problem is not serious. Of course, the credibility of a trial may be adversely affected by this sort of incident and a purist who was an extreme devotee of randomization approaches to inference might insist that the patients should be analysed using the observed results, but

Table 7.3 (Example 7.2) Incomplete blocks cross-over comparing 12 μg and 24 μg formoterol solution aerosol to placebo.

		FEV$_1$ Readings 3 Minutes after Treatment (ℓ) Treatment		
Sequence	Patient	P	F_{12}	F_{24}
$F_{12}P$	4	2.500	3.400	
	11	1.925	2.250	
	14	1.260	1.460	
	21	0.880	1.480	
	35	2.100	2.050	
	mean	1.733	2.128	
PF_{12}	5	2.500	3.500	
	9	1.600	2.650	
	16	1.750	2.190	
	19	0.640	0.840	
	mean	1.6225	2.295	
$F_{24}F_{12}$	2		2.250	2.700
	12		0.925	0.900
	13		1.010	1.270
	36		2.100	2.150
	mean		1.57125	1.755
$F_{12}F_{24}$	6		2.500	2.450
	10		1.750	1.725
	15		1.370	1.120
	mean		1.87333	1.765
$F_{24}P$	3	1.350		1.750
	7	2.150		2.525
	18	0.840		1.080
	22	2.310		3.120
	mean	1.6625		2.11875
PF_{24}	1	2.100		3.100
	8	2.300		2.700
	17	1.030		1.870
	20	0.810		0.940
	mean	1.560		2.1525

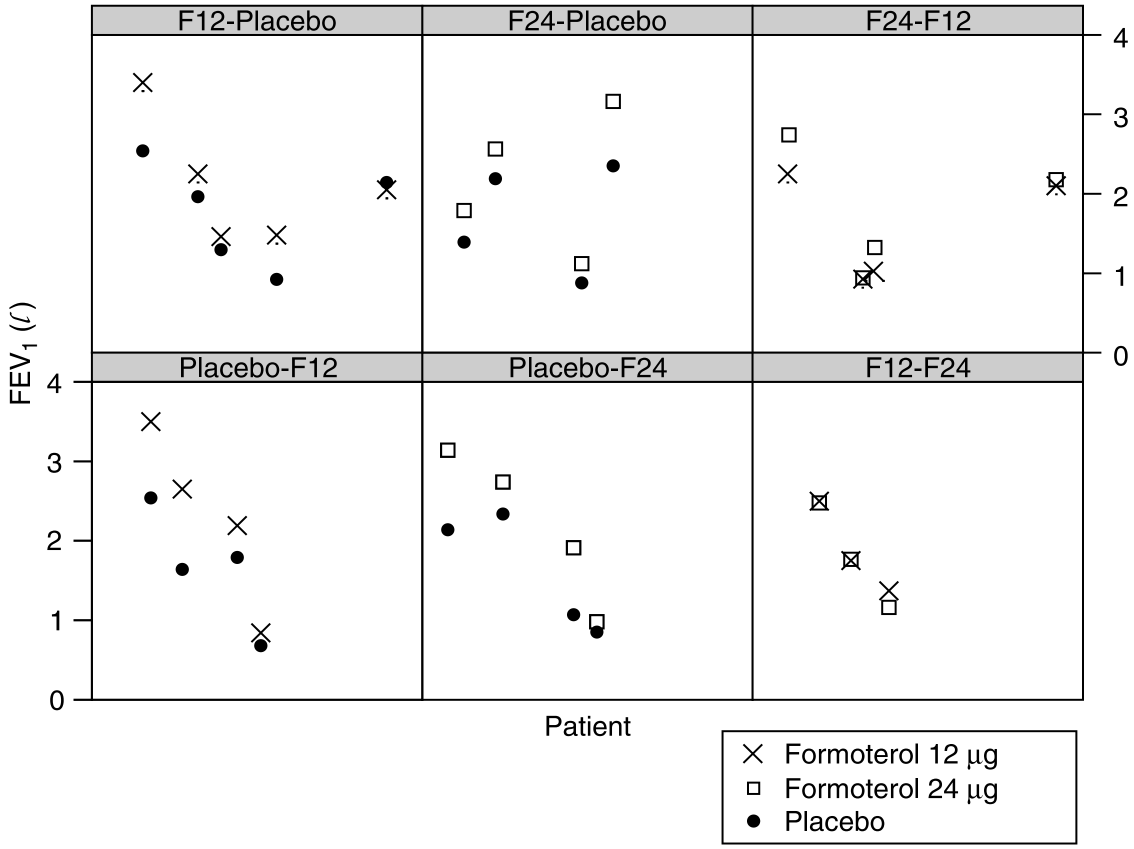


Figure 7.1 (Example 7.2) Forced expiratory volume in one second (FEV1) for an incomplete blocks design. Two doses of formoterol compared to placebo in two periods and six sequences.

according to the treatments they would have received had the randomization proceeded correctly. A more reasonable measure might be to produce two analyses: one with and one without the patient who received the wrong pack but using the treatments actually received.

The results of FEV$_1$ readings taken 3 minutes after treatment are given in Table 7.3. The figures are given to three decimal places of a litre but this is obviously not the precision to which they have been recorded. Of the 48 readings, in the third place of decimals there are 44 '0s' and four '5s' so that even if they were being recorded to the nearest 5 ml there is a strong preference for 0. The 4 readings with a '5' in the third place of decimals have a '2' in the second so that possibly they are being recorded to the nearest 25 ml. In the second place of decimals of the 44 readings whose third place is 0 there are 14 '0s' and 11 '5s'. Thus it appears that some of these measurements are recorded to the nearest 5 ml or possibly 25 ml, some to the nearest 10 ml, some to the nearest 50 ml and some to the nearest 100 ml. This represents a regrettable loss of precision.

The data are plotted in Figure 7.1 from which it is quite clear that there is an important difference between each dose and placebo and also that there is apparently no evidence of a difference between doses. A more formal analysis is considered below.

7.3.2 An illustrative analysis

We now illustrate a possible analysis of these data. This is done merely to give some feel for the analysis of incomplete blocks and to illustrate some features of such designs. It may prove useful as an exploratory way of looking at this particular incomplete blocks design but I make no particular claims for it. In practice I would always use an ordinary least squares analysis via a computer package. Incomplete block designs require rather stronger assumptions for efficient analysis and there seems to be little point, having taken the plunge and designed such a trial, not to analyse them as if such assumptions were justified.

The first step we take is to note that this particular design may be divided into three *AB*/*BA* cross-overs: a comparison of F_{12} and P using 9 patients, a comparison of F_{24} and F_{12} using 7 patients and a comparison of F_{24} and P using 8 patients. We thus use the standard CROS analysis of Section 3.6 on each of these three cross-overs. The necessary calculations are given in Table 7.4, where each column gives the computations for one of the contrasts. The calculations extract what we might refer to as the *direct within-patient information* regarding treatment. Since there is nothing different about these calculations to what was already illustrated for the *AB*/*BA* cross-over in Section 3.6 the reader is invited to check them for himself without benefit of further explanation as to how they are performed.

Remark A number of points are worth noting about these results. First, that the two doses differ significantly from placebo. Second, that the two doses do not appear to differ from each other. Third, that the variances of the basic estimators are similar when the two active doses are compared to placebo but not when compared to each other. This last point is extremely interesting. If patients did not have a uniform response when treated with formoterol (i.e. some had their FEV_1 increased more than others) then we should expect to see exactly this sort of effect on the variances if the two doses showed similar important differences to placebo.

The direct within-patient information is not the only information regarding treatment effects which we have. Consider the difference between formoterol 12 μg and placebo. The direct within-patient estimate is $0.5338\,\ell$. By subtracting the treatment effects of formoterol 24 compared to formoterol 12 from that of formoterol 24 compared to placebo we may obtain a further, and independent estimate (being based on quite different patients), of the effect of treatment. Thus we may calculate $0.5244\,\ell - 0.0377\,\ell = 0.4867\,\ell$. This estimate is also within-patient, being formed itself from two within-patient estimates, but it is indirect. We refer to it, therefore, as an *indirect within-patient estimate*.

We now consider how the two estimates may be combined efficiently. Suppose it is the case that variances of the individual basic estimators are the same irrespective of the contrast being examined. In that case the variances of the

Table 7.4 (Example 7.2) Various calculations.

	Basic estimators (ℓ FEV$_1$)			
	$F_{12}-P$	$F_{24}-F_{12}$	$F_{24}-P$	
	0.900	0.450	0.400	
	0.325	−0.025	0.375	
	0.200	0.260	0.240	
	0.600	0.050	0.810	
	−0.050			
Mean	0.39500	0.18375	0.45625	
CSS (ℓ^2)	0.5380	0.1382	0.1817	
	1.000	−0.050	1.000	
	1.050	−0.025	0.400	
	0.440	−0.250	0.840	
	0.200		0.130	
Mean	0.67250	−0.10833	0.59250	
CSS (ℓ^2)	0.5271	0.0304	0.4783	
Mean	0.5338	0.0377	0.5244	Total
CSS (ℓ^2)	1.0651	0.1686	0.6600	1.8937
DF (ϕ)	7	5	6	18
Var est. (ℓ^2)	0.1522	0.0337	0.1100	0.1052
Varparam. (ℓ^2)	0.01712	0.00492	0.01375	
Stand err.	0.1308	0.0701	0.1173	
t statistic	4.08	0.54	4.47	
$t_{\phi,0.025}$	2.365	2.571	2.447	

direct and indirect within-patient treatment estimates are proportional to the same (unknown) within-patient variance. By our rules for linear combinations (Section 2.2.1) the first has a variance proportional to $(1/5 + 1/4)/4 = 0.1125$ and the second proportional to $(1/4 + 1/3)/4 + (1/4 + 1/4)/4 = 0.2708$. Using the same rules we see that if we combine the two estimates using weights w_1 and $(1 - w_1)$ (i.e. by multiplying the two estimates by these weights and adding) then the result has a variance of

$$\{0.1125w_1^2 + 0.2708(1 - w_1)^2\}\sigma^2 = \{0.3833w_1^2 - 0.5416w_1 + 0.2708\}\sigma^2, \tag{7.1}$$

where σ^2 is the variance of a basic estimator.

Dividing by 0.3833 we find that the variance of the combined estimate is proportional to

$$w_1^2 - 1.413w_1 + 0.707.$$

Completing the square we obtain

$$(w_1 - 0.706)^2 + 0.707 - 0.499.$$

The square of the term in brackets is never negative and hence the whole expression is minimized when w_1 is set equal to 0.706.

Remark A general rule which may be used to find w_1 is as follows. If the direct estimate has variance proportional to a and the indirect proportional to b then the direct estimate should be given the weight $w_1 = b/(a + b)$ and the indirect estimate the weight $1 - w_1 = a/(a + b)$. The reader may check for himself that in this example this produces the same result we have already obtained.

Applying the weights to our direct and indirect estimate we get

$$\hat{\tau} = 0.706 \times 0.5338\,\ell + 0.294 \times 0.4867\,\ell = 0.52\,\ell.$$

The variance of the estimate is found by substituting 0.706 for w_1 in (7.1). Hence

$$\text{var}(\hat{\tau}) = 0.0795\sigma^2. \tag{7.2}$$

To estimate σ^2 we use the total corrected sums of squares from Table 7.4 and divide by the total degrees of freedom. Hence $\hat{\sigma}^2 = 1.8937\,\ell^2/18 = 0.1052\,\ell^2$. Substituting in (7.2) we obtain an estimated treatment variance of $0.008\,36\,\ell^2$. Taking the square root we have a standard error of $0.09\,\ell$.

Remark We have combined direct and indirect within-patient information to produce a single estimate. There is, in fact, a third form of information which may be recovered from incomplete blocks cross-over trials, namely between-block or between-patient information. Such information exists because different patients receive different treatments. If we regard patient effects as random then the differences between observations obtained on patients in different sequences reflects differences between treatments and random variation. Such differences are then amenable to statistical analysis. Recovering such between-patient information is not easy and *usually* makes little difference to the final result. The use of random effect models to recover between-patient information is considered briefly in Section 7.3.5 below. Chi (1991) also covers this in some detail.

7.3.3 The analysis using ordinary least squares

The analysis using ordinary least squares may be implemented quite simply using *proc glm* in SAS®. As usual we gather all the observations under one variable Y. (There are 24 patients, each of whom is measured twice so that there are 48 observations.) Each observation's data line must record under separate variables the *PATIENT* (with values 1 to 22 and 35 and 36) whose observation it is, the *PERIOD* (with values i and ii, say) in which it was taken and the *TREAT*ment (with values *F*12, *F*24, *P*, say) which the patient was receiving. For example, the data line for the first observation on the first patient might read

Y	*PATIENT*	*PERIOD*	*TREAT*
2.100	1	*i*	*P*

We may then use the standard code we have used in Chapters 3 and 5 for analysing cross-over trials:

```
proc glm;
  class PATIENT PERIOD TREAT;
  model Y = PATIENT PERIOD TREAT;
  estimate "F12-Placebo" TREAT 1 0 − 1;
  estimate "F24-Placebo" TREAT 0 1 − 1;
  estimate "F24-F12" TREAT − 1 1 0;
run;
```

The output includes the following:

```
Source            DF   Sum of squares   Mean square    F value   Pr > F

Model             26   24.791 554 48    0.953 521 33   17.94     0.0001
Error             21    1.116 193 44    0.053 152 07
Corrected Total   47   25.907 747 92

                                     T for H0                    Std error of
Parameter        Estimate         Parameter = 0     Pr > |T|      estimate

F12-Placebo      0.504 068 52         5.51          0.0001       0.091 402 75
F24-Placebo      0.544 299 47         5.76          0.0001       0.094 531 74
F24–F12          0.040 230 95         0.41          0.6835       0.097 321 60
```

The results are very similar to those which we obtained before. As before there is clear evidence of a difference between the two active doses and placebo but little evidence of a difference between doses. The degrees of freedom for error are somewhat increased from this analysis. Here they are 21 whereas for the analysis of Section 7.3.2 they were only 18. The reason is that in that analysis 1 degree of freedom for treatment and 1 for period was used up for each of the three cross-over analyses, making a loss of 6 degrees of freedom in total for period and treatment. In the OLS analysis only 2 degrees of freedom for treatment and 1 for period are used. The remaining 3 contribute to the error estimate.

7.3.4 Cell means, weights and ordinary least squares*

This example may be used to make a general point about analysis using ordinary least squares in order to demystify this procedure. The OLS estimates of the treatment effects are simply linear combinations of the 'cell means'. There are $6 \times 2 = 12$ possible cross-classifications of the results according to the sequences and periods they were obtained under. Each of these cross-classifications constitutes a cell. The means from these cells were given in Table 7.3. In Table 7.5 they are reproduced together with the weights used in an OLS analysis. (The derivation of these weights is beyond the scope of this book and need not concern us but a similar case is illustrated in Chapter 10.)

If we take, as an example, the weights for the $F_{12} - P$ contrast, we shall find that when summed over any given sequence, any given period and over the F_{24} treatment, they come to zero. Thus, these factors are eliminated from the estimate of the contrast. On the other hand when summed over F12 they come to 1, and when summed over P to -1, as is necessary if they are to estimate the effect of interest. Indeed, any unbiased estimate of this contrast which allowed for the same factors would have to have this property. The ordinary least squares estimate is simply that linear combination of the observations satisfying this property which has minimum variance. The reader may check for himself that the sum of the product of a given set of weights and the cell means produces the OLS estimates of the contrasts given in Section 7.3.3.

Table 7.5 (Example 7.2) Cell means and weights for an OLS analysis.

				Weights for contrast		
Sequence	Period	Treatment	Cell mean	$F_{12} - P$	$F_{24} - P$	$F_{24} - F_{12}$
$F_{12}P$	1	F_{12}	2.128	0.3855	0.1665	−0.2190
	2	P	1.733	−0.3855	−0.1665	0.2190
PF_{12}	1	P	1.6225	−0.3204	−0.1612	0.1592
	2	F_{12}	2.295	0.3204	0.1612	−0.1592
$F_{24}F_{12}$	1	F_{24}	1.755	−0.1732	0.1752	0.3484
	2	F_{12}	1.57125	0.1732	−0.1752	−0.3484
$F_{12}F_{24}$	1	F_{12}	1.87333	0.1209	−0.1524	−0.2733
	2	F_{24}	1.765	−0.1209	0.1524	0.2733
$F_{24}P$	1	F_{24}	2.11875	0.1408	0.3224	0.1812
	2	P	1.6625	−0.1408	−0.3224	−0.1812
PF_{24}	1	P	1.560	−0.1532	−0.3504	−0.1972
	2	F_{24}	2.1525	0.1532	0.3504	0.1972

7.3.5 Recovering between-patient information using random effect models*

We have, at various points in the book so far, briefly considered the use of random effect models. Their use has had little, if any, impact on the analyses. Indeed, it has generally had no effect at all. In the *AB*/*BA* design, for example, we saw that it made no difference if we fitted the terms due to patients as fixed or random. It is perhaps worth considering briefly again why this is so. To simplify the discussion we shall consider an *AB*/*BA* cross-over trial in an indication for which we consider that the period effect will be negligible and which we have therefore designed by randomizing patients without restriction to the two sequences. This, in the philosophy we have outlined, entitles us to use a simple matched-pairs *t*. We now consider this analysis in terms of fixed and random effect analysis.

In the fixed effect approach, we allow that each patient has his or her own general 'level' and we avoid all speculation as to any distribution that there might be over all levels. We thus introduce these parameters as fixed effects for each patient and, indeed, if we decide to dispense with an overall grand mean we can model the data as follows:

$$Y_{it} = \phi_i + \tau_t + \varepsilon_{it}. \tag{7.3}$$

Here Y_{it} is the response for patient i, $i, = l, \ldots, n$ when given treatment t, $t = A, B$, ϕ_i is the patient effect for patient i, τ_t is the treatment t effect for treatment t and ε_{it} are random independently and identically distributed within-patient disturbance terms. Although we have used two parameters for the treatment effects, it is only the contrast $\tau = \tau_B - \tau_A$ that is estimable. Now, since the ϕ_i are fixed constants, they have to be removed from any unbiased estimate of τ. The only way that any given ϕ_i can be eliminated is to form an individual contrast, a 'basic estimator', of the form $Y_{iB} - Y_{iA}$ for patient i. These thus become the building block for any overall estimator and, since each of these is an equally precise estimate of τ, the optimal overall estimate weights them equally, so that we have

$$\hat{\tau}_f = \sum_{i=1}^{n}(Y_{iB} - Y_{iA})/n, \tag{7.4}$$

where the subscript f in $\hat{\tau}_f$ stands for fixed.

On the other hand, a possible random effects model, is identical to (7.3) but treats the ϕ_i as being independently (both of each other and of the ε_{it} terms) and identically distributed random variables with expectation 0. That being so, any estimator of the form

$$\hat{\tau} = \sum_{i=1}^{n} w_{iA}Y_{iA} + \sum_{i=1}^{n} w_{iB}Y_{iB}, \tag{7.5}$$

where the w_{it} are weights, is an unbiased estimator for τ, provided

$$\sum_{i=1}^{n} w_{iB} = -\sum_{i=1}^{n} w_{iA} = 1.$$

Since any values for the weights, w_{it}, subject only to the constraint above, will make the estimate given by (7.5) unbiased, the question then is what values for the individual weights will minimize its variance. It turns out that the weights for B should all be set equal to $1/n$ and those for A to $-l/n$ so that the solution is exactly as for the fixed effects estimator. The proof of this will not be given here, but its reasonableness can be seen intuitively by considering that if each outcome is measured with equal precision and if patients are completely exchangeable as regards their relevance to determining the effect of treatment, which is implied by our belief that there is no period effect and reflected in our randomization approach, any system which did not weight all observations under treatment A (or B as the case may be) would be completely arbitrary. We thus require that each $w_{iB} = 1/n$ and that each $w_{iA} = -1/n$.

It should be noted, however, that making patient effects random can often give more freedom in estimation. This is because, when patient effects are fixed, each patient introduces a constraint to the estimation process since each patient effect must be eliminated. On the other hand, treating the patient as random replaces this by the single constraint that the sum of these effects should be zero. When, the design is incomplete in some sense, for example if there are missing observations, this can then yield different estimators. The most obvious case is when we have a parallel group trial. If the patient effect is fixed no estimate is possible, since the effects of a given group of patients are confounded with the effect for the treatment for that group. Less extreme cases occur when there are a few missing observations (by chance) from an otherwise complete design. An intermediate case is provided by incomplete block designs where there is a considerable amount of designed missingness.

Consider Example 7.2, and in particular sequences $F_{12}P$ and $P_{12}F$ on the one hand and sequences $F_{24}P$ and PF_{24} on the other. Suppose we compare the patient totals for these two groupings of sequences. Both periods contribute to these totals so that if they differ from one grouping to another it is not because of any period effect. In fact, if we assume that there is no carry-over and no period by treatment interaction they can only differ in a way that reflects the difference between treatment F_{12} (formoterol 12 μg) and treatment F_{24} (formoterol 24 μg) or differences between patients, since different patients are involved. However, if we treat patients as random then differences between these totals vary randomly and we can thus obtain an unbiased estimate of the difference in the treatment estimate. Furthermore, these patient totals are, given suitable assumptions, independent of the within-patient differences, which have formed the basis of our estimation procedures so far. That being so, they provide additional information, which can then be combined with that already obtained. We now illustrate the analysis of this between group information.

Table 7.6 gives summary statistics to be used in recovering inter-block information for Example 7.2. The six sequences of Table 7.3 have been grouped in three pairs according to which two treatments were given, since the order in which they were given is irrelevant to the analysis, which is based on totals. For each of the three groupings the table gives the number of patients, the mean of the totals by patient over the two periods and the corrected sum of squares.

The difference between the mean of grouping III and grouping I is $3.747 - 3.886 = -0.139\,\ell$. This provides an estimate of the treatment contrast F_{24}-F_{12}. By dividing the overall corrected sum of squares of $44.199\,\ell^2$ by the total degrees of freedom, $24 - 3 = 21$, we obtain an estimate of the variance of a patient total of $44.199\,\ell^2/21 = 2.105\,\ell^2$. By our rules for linear combinations, this needs to be multiplied by $(1/8 + 1/9) = 17/72 = 0.236$ to yield $0.497\,\ell^2$ as the variance of this treatment contrast. Taking the square root, we obtain $0.705\,\ell$. Applying the same procedure to the other three contrasts, we have the results given in Table 7.7.

Remark In calculating the standard errors here we have pooled the corrected sum of squares over all three sequence groupings despite the fact that in any contrast only two groupings are involved. An obvious alternative would be to use only the two groupings involved in any sequence. The issues are similar to those that were raised in Chapter 5, and there is no universal recommendation that can be given regarding this. Either strategy can be defended. Pooling is less robust if the purpose is to carry out tests or calculate confidence intervals. In practice, however, such inter-block estimates are very imprecise and, unlike within-patient estimates, never used on their own for inference. In fact, where

Table 7.6 (Example 7.2) Analysis of inter-block information.

Grouping	Number of patients	Mean	Corrected sum of squares
I. $F_{12}P$ & PF_{12}	9	3.886	18.291
II. $F_{24}F_{12}$ & $F_{12}F_{24}$	7	3.460	10.071
III. $F_{24}P$ & PF_{24}	8	3.747	15.837
Total	24		44.199

Table 7.7 Treatment contrasts and standard errors based on inter-patient information.

Contrast	Estimate	Standard error
$F_{12} - P$	−0.2869	0.751
$F_{24} - P$	−0.4261	0.731
$F_{24} - F_{12}$	−0.1392	0.705

they are used, they are combined with inter-patient estimates, as will be discussed in the next section. For this purpose it would be desirable to have a variance estimate based on as many degrees of freedom as possible. I therefore recommend the approach that has been illustrated here.

Remark It is not necessary to perform the calculation in this way. A regression of the patient totals on a set of relevant dummy variables indicating which treatments the patient has taken will achieve the same result. This is illustrated in the SAS® code below, where dummy variables coded 0 and 1 have been used to indicate whether the patient took formoterol 12 μg ($F12$) at any time or formoterol 24 μg ($F24$). It is not necessary to include a further dummy variable for placebo since a coding of 1 for $F12$ and 0 for $F24$, for example, indicates that the other treatment must have been placebo. The variable *FEV1_TOT* carries the total values of forced expiratory volume in one second. The code is:

```
proc glm data=SCHWEIZ2;
  class F12 F24;
  model FEV1_TOT = F12 F24;
  estimate "F12-Placebo" F12 -1 1;
  estimate "F24-Placebo" F24 -1 1;
  estimate "F24-F12" F12 1 -1 F24 -1 1;
run;
```

Here, since the redundant dummy variable for placebo is not included, the effect of the other treatments is measured with respect to placebo. The first two estimate statements produce the contrasts to placebo. The final estimate statement, which compares F_{24} and F_{12}, is formed as the contrast of the two preceding contrasts. The results are as calculated by hand.

Remark In practice the between-patient estimate of the treatment effect from an incomplete blocks experiment is rarely of interest in itself as it generally lacks precision. This is confirmed in this example. The between-patient standard errors are seven to eight times those given in Section 7.3.3.

7.3.6 Combining between-patient and within-patient information*

A superior estimate can be produced by combining within- and between-patient estimates. Given a suitable assumption, the two estimates are uncorrelated. This follows from the fact that the within-patient estimates are based on terms of the general form $Y_2 - Y_1$ whereas the between-patient estimates are based on terms of the form $Y_1 + Y_2$. If we consider the covariance of such terms, we have

$$\begin{aligned}\operatorname{cov}\{(Y_2 - Y_1), (Y_1 + Y_2)\} &= \operatorname{cov}(Y_2, Y_2) - \operatorname{cov}(Y_1, Y_1)\\ &= \operatorname{var}(Y_2) - \operatorname{var}(Y_1) = 0\end{aligned}$$

provided that the variances of first and second period terms are equal.

Now consider a particular contrast, τ. Let us refer to the (unbiased) within-patient estimate of it as $\hat{\tau}_w$ with variance σ_w^2 and to the (unbiased) between-patient estimate as $\hat{\tau}_b$ with variance σ_b^2, and let $\hat{\tau}_c = w\hat{\tau}_w + (1 - w)\hat{\tau}_b$ be a linear combination of them both. We now use our rules for linear combinations given in Section 2.2.1. By (2.2) $\hat{\tau}_c$ is unbiased and has variance $\sigma_c^2 = w^2\sigma_w^2 + (1 - w)^2\ \sigma_b^2$. By expanding this and completing the square, we obtain

$$\begin{aligned}\sigma_c^2 &= w^2\left(\sigma_w^2 + \sigma_b^2\right) - 2w\sigma_b^2 + \sigma_b^2\\ &= \left(\sigma_w^2 + \sigma_b^2\right)\left\{w^2 - 2w\sigma_b^2/\left(\sigma_w^2 + \sigma_b^2\right)\right\} + \sigma_b^2\\ &= \left(\sigma_w^2 + \sigma_b^2\right)\left\{w - \sigma_b^2/\left(\sigma_w^2 + \sigma_b^2\right)\right\}^2 + \sigma_b^2 - \left(\sigma_b^2\right)^4/\left(\sigma_w^2 + \sigma_b^2\right)\\ &= \left(\sigma_w^2 + \sigma_b^2\right)\left\{w - \sigma_b^2/\left(\sigma_w^2 + \sigma_b^2\right)\right\}^2 + \sigma_b^2\sigma_w^2/\left(\sigma_w^2 + \sigma_b^2\right).\end{aligned} \tag{7.6}$$

The second term does not depend on w. The first, being a square, is clearly minimized when zero; this requires setting

$$w = \sigma_b^2/\left(\sigma_w^2 + \sigma_b^2\right), \tag{7.7}$$

which yields a variance at the minimum of

$$\sigma_c^2 = \sigma_b^2\sigma_w^2/\left(\sigma_w^2 + \sigma_b^2\right). \tag{7.8}$$

Remark The weight given by expression (7.7) is the ratio of the variance of the between-patient estimator to the total of the between and the within variances. However, these variances are unknown and that means that w is itself unknown and has to be estimated. Unfortunately, substituting estimated variances in (7.8) yield an estimate of σ_c^2 that is biased downwards: it is on average too low. This can be seen simply by considering that we can rewrite (7.8) as $w(1 - w)(\sigma_b^2 + \sigma_w^2)$. If the ratio w were known we could estimate this as

$$w(1 - w)(\hat{\sigma}_b^2 + \hat{\sigma}_w^2), \tag{7.9}$$

where $\hat{\sigma}_b^2$ and $\hat{\sigma}_w^2$ are the usual unbiased estimators of σ_b^2 and σ_w^2. With w as a known constant, (7.9) would then clearly be unbiased. However, for any value of w other than $\hat{\sigma}_b^2/(\hat{\sigma}_b^2 + \hat{\sigma}_w^2)$, (7.9) would not have reached its minimum, therefore the expected value of that minimum is less than the expected value of the unbiased estimator unless $w = \hat{\sigma}_b^2/(\hat{\sigma}_b^2 + \sigma_w^2)$, which will not happen in

Table 7.8. (Example 7.2) Calculations to combine between- and within-patient information.

	Within-patient			Between-patient		
Contrast	Estimate	SE	Variance	Estimate	SE	Variance
	$\hat{\tau}_w$	$\hat{\sigma}_w$	$\hat{\sigma}_w^2$	$\hat{\tau}_b$	$\hat{\sigma}_b$	$\hat{\sigma}_b^2$
$F_{12} - P$	0.5041	0.0914	0.008 354	−0.2869	0.751	0.564 001
$F_{24} - P$	0.5443	0.0945	0.008 930	−0.4261	0.731	0.534 361
$F_{24} - F_{12}$	0.0402	0.0973	0.009 467	−0.1392	0.705	0.497 025

Contrast	Weight	Combined Estimate	Variance	SE
	$\hat{\sigma}_b^2/(\hat{\sigma}_w^2 + \hat{\sigma}_b^2)$	$w\hat{\tau}_w + (1 - w)\hat{\tau}_b$	$\dfrac{\hat{\sigma}_w^2\hat{\sigma}_b^2}{\hat{\sigma}_w^2 + \hat{\sigma}_b^2}$	
	w	$\hat{\tau}_c$	$\hat{\sigma}_c^2$	$\hat{\sigma}_c$
$F_{12} - P$	0.985	0.493	0.008 23	0.0907
$F_{24} - P$	0.984	0.528	0.008 78	0.0937
$F_{24} - F_{12}$	0.981	0.037	0.009 29	0.0964

general. Hence, if we substitute the variance estimates for parameters in (7.7), the corresponding variance estimate obtained by making the substitution in (7.8) is biased.

Further remark This bias applies to the variance estimate only. It does not apply to the treatment estimate which, being a weighted average of unbiased estimators, with weights w and $(1 - w)$ adding to 1, is unbiased. The net consequence, however, is that confidence intervals for the parameter estimate will be too narrow and significance tests based on comparing point estimate and standard error too liberal. The importance of this should not be over-stressed but, since the amount of between-patient information is often small, it is an argument against recovering such information.

The calculations for combining the between- and within-patient information are given in Table 7.8.

7.3.7 REML analysis*

The combination of within- and between-patient information has been presented rather laboriously above as part of the general didactic aim of this book, wherever possible, illustrating analyses that could be carried out with a pocket calculator as a necessary aid to understanding. However, there is no need in practice to do this. SAS®, GenStat® and S-Plus®, and indeed many other packages, offer facilities, as we have already seen, that permit so-called ‘mixed modelling’ of fixed and random effect factors. The fitting algorithm is known as residual (or restricted) maximum likelihood (REML) estimation and was introduced by Patterson and Thompson (1971). In this section we show, without technical discussion, how this approach can be applied to Example 3.2. The use of GenStat® and S-Plus® is illustrated in the appendices to this chapter.

The SAS® code is as follows:

```
proc mixed;
  class PATIENT PERIOD TREAT;
  model FEV1= PERIOD TREAT;
  random PATIENT;
  estimate "F12-Placebo" TREAT 1 0 −1;
  estimate "F24-Placebo" TREAT 0 1 −1;
  estimate "F24-F12" TREAT −1 1 0;
run;
```

PATIENT, *PERIOD* and *TREAT* are factors for the relevant predictors in the model. *PATIENT* is declared as random and the model statement instructs SAS® to model the response variable *FEV1* as a function of the fixed effects *PERIOD* and *TREAT*.

The output includes:

Parameter	Estimate	Std Error	DF	t	Pr > $\|t\|$
F12–Placebo	0.493 244 59	0.090 716 02	21	5.44	0.0001
F24–Placebo	0.528 949 57	0.093 721 68	21	5.64	0.0001
F24–F12	0.035 704 98	0.096 389 01	21	0.37	0.7148

which can be seen to be very close to that obtained in Section 7.3.6.

7.3.8 Incomplete blocks in general

We have illustrated an incomplete blocks design in which three treatments were studied in two periods. There are many other possible designs. One way of constructing incomplete block designs in three periods and four treatments, for example, is to start with one of the 4 × 4 Latin squares given in Table 5.1 and strike out a column. Suppose we take design I1 of Table 5.1 and strike out the last column. This will leave us with

$$\begin{array}{c} ABC \\ BAD \\ CDA \\ DCB. \end{array}$$

This design is balanced in the following senses: each treatment appears in three sequences, each treatment appears once in each period and each possible pair of treatments appears in two sequences. Such a design could be analysed using ordinary least squares in the way discussed above.

There is an extensive literature on the design and analysis of incomplete blocks. Useful introductions to the topic will be found in Box *et al.* (1978) and Fleiss (1986a). For treatments in more depth the books by Cox (1958), John and Quenouille (1977) and Mead (1988) may be consulted.

7.4 *n* OF 1 TRIALS

A type of within-patient trial which has received a considerable amount of attention in the medical literature recently is the trial in a single patient. Such trials are often referred to as '*n* of 1' trials.

Guyatt *et al.* (1990) describe a simple example of an *n* of 1 trial of amitriptyline in fibrositis. Fibrositis is a condition in which, amongst other ailments, patients suffer from general aches and pains and morning stiffness. A woman aged 47 who had had symptoms for 5 years was treated during three separate periods, each lasting 4 weeks, with 10 mg amitriptyline per day and during 3 similar periods with placebo. Randomization was carried out on paired treatment periods so that during the first, second and third pair of treatment periods the patient always received amitriptyline for 4 weeks and placebo for 4 weeks. The patient was assessed on a weekly basis and rated severity of 7 symptoms each on a 7 point scale. When, at the end of the trial, the results from the periods were compared on a pairwise basis the average score over the 4 weeks was found to be better under amitriptyline than under placebo.

The patient had already been observed to respond to amitriptyline during the course of observation carried out prior to the trial. If a P value is calculated it is appropriate, therefore, to make it one sided. Guyatt *et al.* (1990) point out that it may be calculated as the probability, under the null hypothesis of no difference between treatments, of observing that in each of three pairs the results for the active treatment were better. It may thus be calculated as $P = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = (\frac{1}{2})^3 = 0.125$. Essentially, the same sort of argument may be used as we used in Section 4.3.2 to analyse cross-over trials using the sign test. There, where a number of patients were studied, results were matched by patients. Here one patient is studied but, the randomization having been carried out in terms of pairs of periods, the results are matched by periods.

Guyatt *et al.* (1990) also illustrate how such data may be analysed using the paired t test. The method is then the same as that of Section 3.2 with the difference, as above, that values within the one patient are matched by period as opposed to values within one trial being matched by patient.

7.4.1 Why undertake *n* of 1 trials?

There are two reasons why one may wish to undertake *n* of 1 trials. The first is as a means of saying something in general about treatments (Guyatt *et al.*, 1990). As experimenters we seek to replicate our results. Usually in clinical trials patients form the replications. In cross-over trials periods also provide replication. In an *n* of 1 trial only periods provide the replication. The second reason is to say something specific about an individual patient. The *n* of 1 trial thus represents the extension of the scientific method into the individual doctor–patient relationship (Johannessen, 1991). We shall consider these distinct purposes in our discussion at the end of this section.

7.4.2 Blinding and analysis in *n* of 1 trials

The discussion which follows considers a link between blinding, randomization and inference which is particularly strong in *n* of 1 trials. In the discussion a key role will be given to randomization. My purpose is not to suggest that randomization arguments are the only ones which may be used in considering *n* of 1 trials. I do, however, consider that where blinding is considered to be of the essence randomization arguments should be carefully examined before adopting other types of inferential arguments.

The *P* value of 0.125 which we calculated in Section 7.4 is rather disappointing: the best possible result is observed and yet by conventional standards this is not significant. As we noted in Section 4.3, however, if we consider that blinding is essential then the richness of the randomization performed imposes limits on the *P* value which may be achieved. Because this point is very important, we illustrate it again.

In this trial the possible sequences of amitriptyline (*A*) and placebo (*P*) were

$$\begin{gathered} AP\ AP\ AP \\ AP\ AP\ PA \\ AP\ PA\ AP \\ AP\ PA\ PA \\ PA\ AP\ AP \\ PA\ AP\ PA \\ PA\ PA\ AP \\ PA\ PA\ PA. \end{gathered}$$

These sequences are 8 in number and the *P* value of $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$ may equally well be interpreted as the probability of a lucky guess as to which of the 8 sequences was used (it was, in fact, the sequence *AP AP PA*). Guyatt *et al.* (1990) do not provide the raw data for this example but from the graphs which they do provide, the mean scores for the six periods, in the order in which they were obtained, must be approximately 4.7, 4.2, 5.0, 4.0, 4.2, 5.0. Calculating the pairwise differences ($A - P$) we obtain 0.5, 1.0, 0.8. These in turn have a mean of 0.77 and a sample standard deviation (using the divisor 2) of 0.252. Hence the estimate of the standard error is $0.252/\sqrt{3} = 0.145$ and the t statistic is $5.3 = 0.77/0.145$. (But this degree of precision of the final result is not at all justified by the data values I have given. See Preece, 1981 for a discussion of this problem.) Guyatt *et al.* themselves quote a value for the t statistic of 4.90.

The critical value of t on two degrees of freedom for a one-sided test at the 5% level is 2.92 so that this result, by the logic of the t test and by conventional standards, is clearly significant. Suppose, however, that neither placebo nor treatment has any genuine effect whatsoever and that in the absence of any treatments the values which the lady would have recorded would have been 4.3, 4.6, 4.6, 4.4, 4.6, 4.6. The pairwise differences would then have been −0.3, 0.2, and 0.0. The standard deviation of these values is 0.252 as it was before, but the mean is very close to zero. Suppose, further, that the lady was very suggestible and had a strong prejudice that the treatment would be effective, and that this would cause her to increase her rating by 0.4 when she suspected a treatment was being given and to reduce it by 0.4 when she thought there would be a placebo. She must, naturally, speculate as to what treatment she is being given. If she happens to guess correctly she will produce the values that we did in fact observe. But the probability of her guessing correctly is simply the probability of guessing which of the 8 sequences was actually used, which is 1/8 or 0.125, the probability we obtained in Section 7.4. If the reader compares this with the example given in Section 4.3, where six treatment periods were also used, he will see that where blinding is considered essential the minimum P value obtainable is governed by the richness of the randomization actually used.

This is a rule which applies to blinded trials in general. When blinding the patient is considered important, however, it causes particular problems for n of 1 trials since only one patient is involved. It is also a mistake to think that blinding can be improved by lying to the patient—for example by claiming that she will be given amitriptyline 3 times and placebo 3 times completely at random whilst secretly resolving to randomize in pairs. First, such behaviour is clearly unethical. When we experiment on human beings we hope they will tell us the truth. We should not expect higher standards of truthfulness from our subjects than we are prepared to employ ourselves. Secondly, the success of this strategy relies on the supposed stupidity of the subject: she may be cleverer than we think.

All of these issues were grasped by R. A. Fisher a long time ago but seem to have been forgotten by modern commentators. Fisher considers the problem of designing a ‘psycho-physical experiment’:

> A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup . . . Our experiment

> consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgement in a random order. *The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order*...Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received. (Fisher, 1990b, p. 11, my italics.)

In discussing this example we shall use the symbol *M* to denote 'milk first' and *T* to denote 'tea first'.

The important point here is that Fisher has told the subject how the randomization will be carried out. Had he lied to her and said that he would randomize completely at random, so that sequences like *MTTTTMTT* were possible, whilst secretly resolving to have 4 cups of each kind, he would then be uncertain as to which of two possible *P* values would be appropriate if the lady guessed all cups correctly: 1/70 reflecting the 70 possible sequences actually used or 1/256 reflecting the 256 possible sequences he wished her to believe had been used. Similarly, Fisher will be wise to instruct her carefully as to how the randomization will be carried out. Suppose he obtains the sequence *MTTMTMTM*. This is a sequence which could have been obtained by randomizing in pairs. If the lady wrongly assumed he would randomize in pairs her chance of guessing correctly given this lucky randomization is $(\frac{1}{2})^4 = \frac{1}{16}$. Of course, unconditionally, her probability of guessing correctly is still $\frac{1}{70}$, since only 16 out of the 70 possible sequences of the randomization actually performed will form a sequence which could have arisen randomizing in pairs. His strategy of informing her, however, has the advantage of keeping the conditional and unconditional probabilities in line.

Thus, my advice in analysing *n* of 1 trials is, that if *P* values or confidence limits are to be calculated, and if these calculations are not based on randomization arguments, then if blinding is considered to be important, it would be useful to carry out the randomization analysis as well.

7.4.3 General discussion

We do not have space to cover the analysis of *n* of 1 trials in general. Curiously, however, the analysis of a single *n* of 1 trial in isolation does not require any methodology not already available for the analysis of single centre parallel group trials. This is because the episodes of treatment of the patient may be regarded as replicate observations of the patient in the same way as the patients in the single-centre trial may be used as replicate observations on that centre.

Different issues arise where we either have a series of *n* of 1 trials or we wish to use the results of an *n* of 1 trial in connection with background information available from other trials to make a specific recommendation for a given patient.

In the former case we can usually treat the series of *n* of 1 trials as a single cross-over. This is, in fact, one way of looking at the Cushny and Peebles (1905) data we considered in Chapter 1. The patients had received each treatment a number of times. Cushny and Peebles (1905) reduced these multiple measurements to a single average for each treatment for each patient and these averages formed the basis of Student's (1908) analysis. This is a general and robust option to analysing such data although there may be interest in trying to establish to what extent different patients respond to different treatments and this requires an alternative analysis in terms of the original observations (Preece, 1982). Where we wish to make recommendations for an individual patient, the *P* value from his individual trial is of little interest. Either we have no other information available, in which case we might simply make the pragmatic recommendation that he should take as a future treatment that treatment which has the best response in the trial in question, or we need to combine the 'local' information from his own trial appropriately with the 'global' information from other studies. We cannot pursue this fascinating issue further here.

Further comments on *n* of 1 trials will be found in Spiegelhalter (1988), Lewis (1991) and Senn (1993d).

7.5 BIOEQUIVALENCE STUDIES

In Section 5.1.5 we discussed equivalence studies. Example 7.1 of this chapter might also be described as an equivalence study. A common type of equivalence study, which is carried out almost exclusively as a cross-over, is the so called *bioequivalence* study. The purpose of these studies is to compare two formulations of a treatment as regards their bioavailability. Usually this is determined by the concentration over time of the active ingredient in blood plasma. Normally these studies are carried out in healthy volunteers. The desired object is to be able to declare, at the end of the study, that the two formulations are, as regards their bioavailability, effectively equivalent.

For example, we may have a treatment available in the form of a pill which has already been studied extensively. Some patients, however, may prefer to take the medicine in a dispersible form. As a strategy for developing the new formulation then, rather than repeat the whole experimental programme, we attempt to show that once the new formulation has been ingested it is equivalent to the standard.

As regards the production of treatment estimates and standard errors these types of study present no new features not already covered elsewhere in our discussion of cross-over trials. As regards hypothesis testing, and more generally interpretation, they are quite different from other trials. There are many different opinions regarding appropriate analyses. I list, below, some brief comments regarding various features of these studies. For a more thorough discussion of statistical issues the reader may consult Metzler (1991), Senn (2001c) or Chow and Liu (2000). The first edition of *Cross-over Trials in Clinical Research* drew attention to the papers by Dighe and Adams (1991), McGilveray (1991), Rauws (1991), Salmonson *et al.* (1991) and Walters and Hall (1991) in the book edited by Welling *et al.* (1991) covering regulatory aspects from the perspectives of the USA, Canada, Australia and the Nordic countries, respectively. Since then there has been a considerable development both on the regulatory front and in terms of methodological research. The Food and Drug Administration developed a guideline in the late 1990s (FDA CDER, 2000) to address additional issues of population and individual bioequivalence, concepts that were introduced by Anderson and Hauck (1990; Hanck and Anderson, 1991). These concepts will be discussed briefly in Sections 7.5.5 and 7.5.6 below. The CPMP has also issued a recent guideline (Committee for Proprietary Medicinal Products, 2001), which does not cover population or individual bioequivalence. More up-to-date discussion of various regulatory requirements will be found in Gleiter *et al.* (1998), Hauschke and Steinijans (2000) and Steinijans and Hauschke (1997). Further discussions of pharmacokinetic theories in general will be found in Chapter 10.

7.5.1 Measures

Traditionally three measures have been used to compare treatments for bioequivalence: area under the curve, AUC, maximum concentration, $C_{\max}$, and time to reach maximum concentration, $T_{\max}$.

Area under the curve is usually calculated using the trapezoidal rule. Thus if we have $n+1$ measurements, $Y_0, Y_1, \ldots, Y_n$ taken at times $t_0, t_1, \ldots, t_n$, then

$$\begin{aligned}AUC =&(Y_0+Y_1)(t_1-t_0)/2+(Y_1+Y_2)(t_2-t_1)/2+\cdots\\&+(Y_{n-1}+Y_n)(t_n-t_{n-1})/2.\end{aligned}$$

When the measurements are taken at equal intervals t this reduces to

$$AUC = t\,\{(Y_0+Y_n)/2+Y_1+Y_2+\cdots+Y_{n-1}\},$$

and so is effectively a weighted sum of the observations. The factor t is irrelevant in that it will neither change from subject to subject nor from treatment to treatment. If we standardize the total time of observation, nt, to be 1 then $t=1/n$ and the AUC is almost identical to the arithmetic mean. It is, therefore, a fairly stable measurement which does not suffer too much from random variation.

$C_{\max}$ is simply the highest recorded value for each patient. Other things being equal, the more frequently we measure the greater it will be. There are two reasons: first, the more often we measure the more likely we are to get close to the true peak and secondly, if we have n measurements and add one more, the maximum of the $n+1$ values might be higher but could never be lower than the maximum of the n values.

T_{max} is the time at which the maximum concentration is measured. It is a measure which shows great variability in comparison to its sensitivity. For example, if concentration rises rapidly and then reaches a plateau by the 3rd measurement (say) which it maintains approximately until the 7th measurement (say) it makes little difference to our assessment of the maximum concentration for a given patient whether for him the 3rd, 4th, 5th, 6th or 7th measurement is the highest. Obviously, it affects the T_{max} considerably.

For these reasons I dislike C_{max} and especially T_{max} as individual measures. (As pharmacokinetic *concepts* they are very useful but that is a different issue.) I recommend, if possible, trying to look at concentration curves for individuals in more robust terms: for example by calculating area under the curve for the first and second half of the observation period and so on. (See Section 8.6 for an example of such a calculation.) This is not standard practice, however, and it is likely that C_{max} and T_{max} will continue to be used in bioequivalence studies, although there are signs that various regulatory agencies are becoming more flexible with regard to their attitude to these measures. A good discussion of the use of C_{max} and T_{max} in bioequivalence studies will be found in McGilveray (1992).

7.5.2 Blinding

We have already noted that blinding is a problem for equivalence studies generally (e.g. Section 4.3.11). Its value for bioequivalence studies is somewhat doubtful. It is certainly no protection against fraudulent claims of equivalence. All an unscrupulous investigator needs to do is mix the blood sample from one treatment day from that with the other for each patient and divide them in two again if he wishes to ‘prove’ equivalence. I do not wish to suggest that this is a real danger with most investigators! The blinding issue for equivalence trials, is, however, merely a symptom of another more general problem that affects them. In carrying out such trials we are attempting to ‘prove’ that the observations come from a single distribution. This is the opposite of what we usually attempt with clinical trials. We consider the problem this poses in the next section.

7.5.3 Hypothesis (significance) testing*

The view of R. A. Fisher (with which I agree) was that ‘A test of significance contains no criterion for “accepting” a hypothesis’ (Fisher, 1990c, p. 45). It is not possible, for example, to state that two treatments are equivalent because we fail to 'reject' by a test of significance that they are equivalent.

In the Neyman–Pearson formulation of hypothesis tests a solution to this problem is possible. We accept that our object is to prove equivalence of two treatments to an acceptable degree. If the true treatment effect is τ and the smallest difference not accepted as being clinically equivalent is δ, $\delta > 0$, then we test

$$H_0: |\tau| \geqslant \delta \quad \text{against} \quad H_1: |\tau| < \delta,$$

where, in non-mathematical terms, H_0 represents 'not equivalent' and H_1 'equivalent'. The problem is then one of finding a suitable test statistic and a suitable critical region such that given that H_0 is true our probability of concluding that the treatments are equivalent is never greater than some predetermined level α (the size of the test) and the other error, β, that of concluding that the two treatments are not equivalent when they are, is minimized.

Although in mathematical terms reversing the role of null and alternative hypotheses in this way seems to provide a logical solution to the problem it is my opinion that this solution is nevertheless unacceptable. As an example (amongst many) of the difficulties such a test would pose we may consider that in practice to implement such a test we would assume that the variances under standard and new formulations were equal. In the more usual case where our null hypothesis is that the treatments are equal in their effects, to assume that their variances are equal when testing the equality of their means involves no serious contradiction: the assumption belongs to the hypothesis being tested (Fisher, 1925a, p. 125). On the other hand where the 'null' hypothesis is that the treatments are unequal there is no logical justification for assuming as part of the proof of their equality that the variances are equal.

In my opinion it is far better to regard this problem as an estimation problem and to report estimates and confidence intervals for treatment differences for the measures examined. If any formal requirement for equivalence is to be stated it should be stated in terms of limits of equivalence within which the confidence limits should lie. It should be clearly understood, however, that such demonstrations of 'equivalence' are quite different (and weaker than) demonstrations of differences between an active treatment and placebo (Temple, 1982; Makuch and Johnson, 1989).

When one comes to look at the regulatory requirements for the confidence intervals to be used, there is not universal agreement. The Canadian Health Protection Branch (HPB), for example, prefers 95% limits (McGilveray, 1991) whereas, for example, the Australian Drug Evaluation Committee (ADEC) prefers 90% limits. Using the latter with given limits of equivalence corresponds to applying two one-sided tests at the 5% limit, each adopting as a null hypothesis that the treatment difference is (just beyond) either the lower or upper limit (as the case may be) of equivalence. It should be noted, however, that regulatory agencies usually expect to see two-sided tests at the 5% level when comparing an active treatment with a placebo. Since no regulator would register a sponsor's drug if it were significantly *worse* than placebo, the *regulator's risk* in such cases for a non-effective drug is 2.5%. It might be argued that the equivalent approach in a bioequivalence study is to use two one-sided tests at the 2.5% level which would bring us back to 95% confidence intervals.

In fact, the confidence interval approach does not quite correspond to the most powerful hypothesis test of non-equivalence against equivalence although in most practical cases it is very similar (Berger and Hsu, 1996; Brown *et al.*, 1997; Mehring, 1993). The reason why this is so may be understood by considering the unrealistic but instructive case where the true population variance is known. In that case, for any study, the width of the confidence interval will be known in advance. It is only necessary to carry out the study in order to obtain a point estimate. However, if the study is extremely small, the standard error will be large and it may be that the width of the confidence interval exceeds the width of the interval of equivalence. Thus, even if the point estimate for the difference on the log scale is zero, a declaration of equivalence will be impossible. It follows, therefore, that the Type I error rate is zero. In order to recover (say) a Type I error rate of 5% it will be necessary to permit a declaration of equivalence for 'small' differences from zero of the point estimate. If the standard error is large enough it turns out that 'small' can be greater than the limit of equivalence. Thus a declaration of equivalence becomes possible even when the point estimate is not within the limits of equivalence! Of course, in practice no regulator would accept such a declaration (Senn, 2000a, 2001a, 2001c). As I have said above, however, I prefer to look at this as an estimation problem.

7.5.4 Limits of equivalence

Quite often the results of a bioequivalence study are expressed in terms of ratios of effects new to reference. As we pointed out in Section 4.2.1 this may be achieved by logarithmic transformation of the data, performing the analysis and antilogging the result. I can see no advantages in the alternative approach of working in terms of original observations and using complicated theorems in probability to express resulting treatment differences as ratios. It should be noted, however, that, at the time of writing, some authorities, e.g. the US Foods and Drugs Administration (FDA) prefer analyses on untransformed data (Dighe and Adams, 1991), whereas others, e.g. the Canadian HPB (McGilveray, 1991) prefer log transformations.

It is common to see the limits of equivalence 0.8 to 1.2 or (for more strict equivalence 0.9 to 1.1). This is not particularly logical. It is often a historical accident as to which of the two formulations is standard and which is new. A ratio of new to standard of 0.8 corresponds to a ratio of standard to new of 1.25

and a ratio of 0.9 corresponds to 1.11. To maintain invariance as to which is labelled new and which standard it is thus logical to use the limits 0.8 to 1.25 (or 0.83 to 1.20) or 0.9 to 1.11 (or 0.909 to 1.1) or, indeed, any other two limits which are reciprocals of each other.

7.5.5 Population and individual bioequivalence*

These topics have become fashionable in recent years. There is a vast and rapidly growing literature on this topic, and only a superficial coverage will be possible here. For reasons which will be explained in due course I do not believe, however, that individual bioequivalence is of much practical importance (Senn, 1998b). However, it does relate to the issue of treatment by patient interaction, a topic which ought to be investigated more thoroughly than it commonly is. It is ironic that it is in the field in which it is likely to be least relevant, that of bioequivalence, that most work regarding this is being done. Therefore, because examples of trials that permit estimation of patient by treatment interaction are rare, some space will be taken to look at this topic here.

It is, of course, possible that two treatments could show the same bioavailability on average but rather different variability (Hauck and Anderson, 1991, 1994). This might, for example, be a consequence of difference in manufacturing processes, the one being more tightly controlled than the other. Alternatively, if different routes of administration are involved, as might be the case in comparing an oral formulation to a suppository, it is likely that the dose delivered will have to be different in order to ensure that the same expected dose is absorbed. Thus a different fraction of the drug will be absorbed and at least some different body systems will be involved in the one case or the other, so that variability may plausibly be different. Furthermore, as mentioned in Section 7.5.3 since the null hypothesis is one of a difference rather than equivalence, equality of variance does not correspond to the hypothesis being tested. If the formulations varied in terms of variability, they would not be equally ‘prescribable’, to use the term introduced by Anderson and Hauck (1990).

A less plausible possibility is that two formulations could have the same average bioavailability and the same variability but that a subject by formulation interaction could be present. Thus the two formulations would actually be different for different subjects, the one being more bioavailable for some and less bioavailable for others. As expressed by Anderson and Hauck (1990), the two formulations would not then be switchable.

In order to examine the subject by treatment interaction, it is desirable to carry out a cross-over trial in which subjects have been treated more than once with each treatment. An example of such a trial, with listing of all relevant data, is provided by Shumaker and Metzler (1998). We now discuss this example.

Example 7.3 A cross-over trial in four periods was used to compare two formulations of phenytoin, a test formulation (T) and a reference formulation (R). Twenty-six subjects were randomized to receive one of two treatment sequences: *RTTR* or *TRRT*. Shumaker and Metzler (1998) provide data for both $C_{\max}$ and AUC. Those for AUC (log-transformed) are reproduced in Table 7.9. (Since we are interested in relative bioavailability only, the units are irrelevant.)

We shall not analyse these data by hand but instead will use *proc mixed* of SAS®, using code that is very similar to that provided by Shumaker and Metzler (1998). However, before we do that, we consider a graphical representation of the data that will help illustrate, heuristically, why a design with repeated periods such as this permits us to investigate individual bioequivalence.

Table 7.9 (Example 7.3) Data (log-AUCs) for a bioequivalence study comparing two formulations of phenytoin (Shumaker and Metzler, 1998). Reproduced with permission R. C. Shumaker.

		Period				Estimate	
Subject		1	2	3	4	1	2
	Sequence	T	R	R	T	T–R	T–R
3		4.294	4.285	4.162	4.103	0.009	−0.060
4		3.830	4.037	3.866	3.869	−0.207	0.003
6		3.661	3.730	3.676	3.731	−0.069	0.055
7		4.465	4.503	4.460	4.488	−0.038	0.028
10		4.292	4.344	4.331	4.303	−0.051	−0.028
11		3.800	3.850	3.915	3.919	−0.049	0.003
13		4.161	4.149	4.202	4.191	0.012	−0.010
16		4.149	4.164	4.086	4.131	−0.015	0.046
18		4.149	4.193	4.107	4.296	−0.044	0.189
21		3.598	3.587	3.658	3.664	0.011	0.006
22		4.566	4.654	4.536	4.557	−0.088	0.022
23		3.676	3.728	3.766	3.796	−0.052	0.030
26		3.683	3.743	3.943	3.758	−0.060	−0.185
	Sequence	R	T	T	R	T–R	T–R
1		3.591	3.636	3.608	3.629	0.045	−0.021
2		3.907	3.900	3.849	3.941	−0.007	−0.091
5		3.962	4.036	3.969	3.842	0.075	0.127
8		3.530	3.756	3.628	3.668	0.226	−0.040
9		3.887	3.976	3.865	3.940	0.089	−0.075
12		4.470	4.437	4.463	4.446	−0.032	0.017
14		4.132	4.135	4.159	4.126	0.003	0.034
15		3.691	3.694	3.658	3.709	0.003	−0.050
17		3.740	3.744	3.694	3.781	0.004	−0.087
19		4.114	4.111	4.064	4.085	−0.004	−0.021
20		3.747	3.787	3.806	3.810	0.040	−0.004
24		3.695	3.848	3.771	3.914	0.153	−0.143
25		3.862	3.810	3.812	3.967	−0.052	−0.155

Each subject gives us the possibility of estimating two basic estimators of the difference between the test and reference formulations. These are, in fact, presented in the last two columns of Table 7.9. We assume for the moment that there is no period effect and also no carry-over effect (but will return to this point). If there is no treatment by subject interaction, each subject provides a means of estimating exactly the same quantity and there is no reason why we should expect two estimates to be more similar if they come from the same subject than if they come from a different subject. However, a positive correlation between the two estimates from the same subject would indicate that for different subjects the contrast T–R was estimating a different effect. We can thus plot the result for the second estimate against the first. The result is shown in Figure 7.2.

This scatter-plot exhibits no evidence of positive correlation between the estimated contrasts from the two periods. In fact, when the correlation coefficient is calculated it is found to be negative and equal to −0.18. This is not significantly different from zero and it seems reasonable to assume, in view of the fact that formulation by subject interaction is unlikely *a priori*, that the true correlation is either zero or very close to it.

The point, however, is that what permits us to estimate the interaction is the fact that we have more than one estimate of the treatment effect for each subject.

We now illustrate two possible analyses of these data with SAS®. For the first we suppose that the data are arranged in vectors of length 26 × 4 = 104, so that each of the 26 subjects has four lines of data. We assume that this data set has been given the name *ORIGINAL*. We then use the following code:

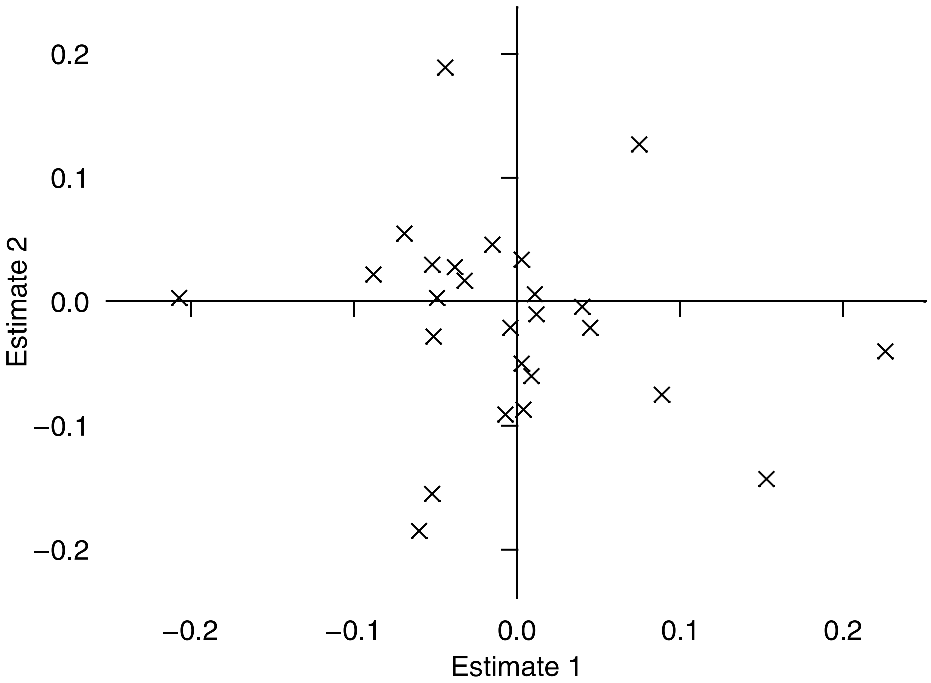


Figure 7.2 (Example 7.3) Scatterplot showing second determination of log-difference in AUC plotted against the first.

```
proc mixed data=ORIGINAL;
  class SUBJECT FORMUL PERIOD;
  model lnAUC = PERIOD FORMUL;
  random SUBJECT SUBJECT*FORMUL;
  estimate 'formulation' FORMUL -1 1;
run;
```

Here *SUBJECT*, *PERIOD* and *FORMUL* are factors with 26, 4 and 2 levels respectively corresponding (obviously) to subjects 1–26, periods 1–4 and formulations (test or reference). By nominating not only *SUBJECT* but also *SUBJECT***FORMUL* as a random effect we inform SAS that the extent to which the formulations differ must be referred to the extent to which this difference varies from subject to subject. The output includes the following:

```
                                Estimates

                                Standard
   Label           Estimate       Error      DF    t Value      Pr > |t|
   formulation     -0.00986     0.01091      24     -0.90       0.3750
```

Note that the degrees of freedom for error are fewer than the number of subjects. The output also includes

```
                          Covariance Parameter
                                Estimates

                   Cov Parm                 Estimate
                   SUBJECT                   0.07670
                   SUBJECT*FORMUL                  0
                   Residual                 0.003095
```

Note that the estimate of the variance of the subject by formulation interaction is zero. This is because the correlation coefficient is negative but the lowest value that can realistically be imagined is 0, which would in fact be the value corresponding to a random effect variance for subject by formulation interaction of zero. The default fitting approach for SAS® *proc mixed* constrains the value to be non-negative.

For the second, suppose that the data are arranged into vectors of length 26 as in Table 7.9. We calculate for each subject two basic estimators of the formulation effect just as we did to produce Figure 7.2. We now calculate the mean of these two for each subject, *BASICM*, and the semi-difference, *BASICD* (which we shall use later), and store these in the data set *BASIC*. We than carry out an analysis using the following code:

```
proc glm data=BASIC;
  class SEQ;
  model BASICM=SEQ;
  estimate ‘formulation’ INTERCEPT 1;
run;
```

Included in the output is the following

```
Dependent Variable: BASICM
-----------------------------------------------------------------------
                                Sum of
Source             DF          Squares    Mean Square   F Value   Pr > F
-----------------------------------------------------------------------
Model               1       0.00319935     0.00319935      1.19   0.2868
Error              24       0.06469908     0.00269580
Corrected Total    25       0.06789843
-----------------------------------------------------------------------

                                    Standard
Parameter           Estimate           Error    t Value   Pr > |t|
formulation       0.00986220      0.01018256       0.97     0.3424
------------------------------------------------------------------
```

But for a change of sign the estimate is exactly as we had it before. The standard error is a little lower. This is because this method, being constructed empirically from the observed variance of the mean of the two basic estimators, takes no account of the fact that this variation is slightly less than one would expect when looking at the variability of the data as a whole due to the observed (no doubt fortuitous) negative correlation. On the other hand, our previous fitting method constrained the variance associated with the interaction to be no less than zero. In fact a slight modification of our previous fitting method, to remove this restriction makes the two procedures equivalent. We need to write instead:

```
proc mixed data=ORIGINAL method=type1;.
```

Then we obtain the same results as using basic estimators.

If in our basic estimator analysis we use the semi-difference rather than the mean, we obtain a result, which, to the extent that it reflects random variation, should give the same residual variance as using the basic estimator. The resulting ANOVA table is as follows:

```
Dependent Variable: BASICD
-----------------------------------------------------------------------
                                Sum of
Source             DF          Squares    Mean Square   F Value   Pr > F
-----------------------------------------------------------------------
Model               1       0.03090851     0.03090851     11.21   0.0027
Error              24       0.06618533     0.00275772
Corrected Total    25       0.09709384
-----------------------------------------------------------------------
```

Note that our mean square error is now 0.00276 (to three significant figures), where previously it was 0.00270. It is thus slightly higher. However, the semi-difference must eliminate any subject by treatment interaction whereas the mean reflects it. Thus the fact that this figure is now higher, rather than lower, is yet another illustration of the negative correlation amongst the basic estimators.

This general topic would almost be worth a book on its own. However, we shall not labour the point any further apart from making two remarks in summary.

1. I am firmly convinced that this topic, in the context of bioequivalence, is of no practical importance (Senn, 1998b, 2001a, 2001c). The main reasons for this are (a) Subject by formulation is likely to be unimportant. (b) Even if important, the risk taken by patients switching from one formulation to another is likely to be much less than patients taking the drug in either formulation for the first time. (c) It is not unknown for rival manufacturers to register different formulations of the same drug using free-standing dossiers. Under such circumstances regulators do not even check whether the drugs are bioequivalent in a mean sense. However, patients can be and are switched from one formulation to another. It would be perverse to demand a protection for this under one circumstance (generic competition) and not the other (brand name competition).
2. However, in other contexts, the issue of patient by treatment interaction is an important and sadly neglected one (Senn, 2001b). Such repeated period cross-overs represent an ideal tool for investigating this phenomenon. It is unfortunate that they are rarely used.

7.5.6 Further information

There is a very extensive literature on bioequivalence trials and a rather surprising degree of disagreement upon what is the right approach even among authors apparently working in a frequentist framework (see for example Westlake, 1976, 1981; and Kirkwood, 1981). Various rules are presented by Mandallaz and Mau (1981) who include a discussion from a Bayesian point of view. Bayesian approaches have also been proposed by Fluehler *et al.* (1981), Selwyn *et al.* (1981), Racine-Poon *et al.* (1987) and Lindley (1998), the latter including an explicit consideration of loss functions. A comparison of this and various frequentist approaches is made by Senn (2001c). Binary equivalence, which has particular difficulties, is considered by Holmgren (1999). Hauschke *et al.* (1990) describe some non-parametric approaches. An extremely thorough and rigorous examination of hypothesis testing for equivalence, from a frequentist point of view, is given by Mehring (1992) as part of a general investigation into testing for interval hypotheses. A very useful review is that of O'Quigley and Baudoin (1988). There is also a book by Chow and Liu (2000) devoted exclusively to this topic.

APPENDIX 7.1 ANALYSIS WITH GENSTAT®

We illustrate the analysis of Example 7.2, omitting details of data entry for brevity. All variables and factors are vectors of length 48 corresponding to the two periods of measurement on each of 24 patients. Obvious names, *Patient*, *Period*, *Treat* and *FEV1*, are used for effect factors and the outcome variable. The code (making much use of GenStat® default settings) is as follows.

```
"Fixed effects analysis"
MODEL FEV1
TERMS [FACT=1] Patient+Period+Treat
FIT [TPROB=yes; FACT=1] Patient+Period+Treat

"Random effects analysis"
VCOMPONENTS [FIXED=Period+Treat; FACTORIAL=1]\
RANDOM=Patient; INITIAL=1; CONSTRAINTS=positive
REML FEV1
VDISPLAY [PRINT=effects; PSE=alldifferences]
```

The output for the fixed effects analysis includes the following:

```
*** Estimates of parameters ***
                        estimate      s.e.    t(21)     t pr.
. . .                      . . . . .    . . . .    . . . .    . . . . .
Treat F24                 0.0402    0.0973     0.41     0.684
Treat P                  −0.5041    0.0914    −5.51    < .001
```

The output for the random-effect analysis includes the following:

```
*** Table of effects for Treat ***

Treat               F12          F24            P

                 0.0000       0.0357      −0.4932
```

Standard errors of differences between pairs

```
Treat   F12   1        *
Treat   F24   2   0.0964         *
Treat     P   3   0.0907    0.0937        *
```

The results are the same as those we obtained with SAS® and with S-Plus® (below).

APPENDIX 7.2 ANALYSIS WITH S-PLUS®

We illustrate the analysis of Example 4.2, omitting details of data entry for brevity. All variables and factors are vectors of length 48 corresponding to the two periods of measurement on each of 24 patients. Obvious names, *patient*, *period*, *treat* and *FEV1*, are used for effect factors and the outcome variable. The code is as follows:

```
#Define contrasts
options(contrasts=c("contr.treatment", "contr.poly"))

# Fixed effects analysis
fit1<-lm(FEV1~patient+period+treat)
summary(fit1,corr=F)

# Random effects analysis
fit2<-lme(FEV1~period+treat, random=~ 1 | patient)
summary(fit2)
```

For the fixed effects analysis the output includes:

```
Coefficients:
                 Value Std.      Error     t value     Pr( > |t|)
...              ......         .....      ......      ......
treatF24         0.0402         0.0973     0.4134      0.6835
treatP          −0.5041         0.0914    −5.5148      0.0000
```

For the random effects analysis the output includes:

```
Fixed effects: FEV1 ~ period + treat
                  Value     Std. Error   DF    t-value    p-value
(Intercept)    2.006055     0.1607574    23   12.47877    < .0001
period         0.030373     0.0666605    21    0.45564     0.6533
treatF24       0.035705     0.0963888    21    0.37043     0.7148
treatP        −0.493245     0.0907158    21   −5.43725    < .0001
```

These are the same results obtained with SAS® and GenStat®.

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

8

Graphical and Tabular Presentation of Cross-over Trials

8.1 BASIC PRINCIPLES

An account of various basic principles which ought to influence the preparation of graphs in general will be found in Tufte (1983). Two which he mentions are worth repeating. First, 'to maximise the data–ink ratio, within reason' (p. 96). In other words graphical *display* should be limited as much as is possible to what may be *shown* regarding the data and ornament and embellishment should be avoided. Second, that the number of 'information-carrying (variable) dimensions should not exceed the number of dimensions in the data' (p. 71). We should thus, for example, avoid 'solid' bars in an ordinary histogram: we would then be using three dimensions to represent two (value and frequency). Cleveland and McGill (1987) consider the implications of theories of perception for graphical comprehension. A useful discussion of various points concerning the relationship between graphical methods and statistics will be found in Cox (1978). A general issue which affects all clinical trials and which might seem trivially obvious but is often overlooked, is that graphical representations should attempt to illustrate the *effects* of treatments (Senn and Auclair, 1990). We now consider some particular features which specifically affect tabular and graphical representations from cross-over trials.

The first is that they should show the within-patient structure of the trial and therefore either eliminate between-patient variability or reflect it in such a way that it may be taken into account appropriately. We carry out cross-over trials in order to eliminate between patient variability from our estimates. We should also do the same in our graphs, the reason being that of all the statistical devices we use, graphs are those which have the greatest impact and also (potentially) the widest reception: more persons are able to understand the graphs we use than can understand the analyses we produce. As an example of eliminating between-patient variation, the simplest approach is to use differences between treatments for each patient. As an example of appropriately reflecting between patient variability, if we plot measurements on a number of treatments taken on a number of patients we should use a graph which enables the viewer to link measurements by patient on the graph. This is also true of tabular presentations.

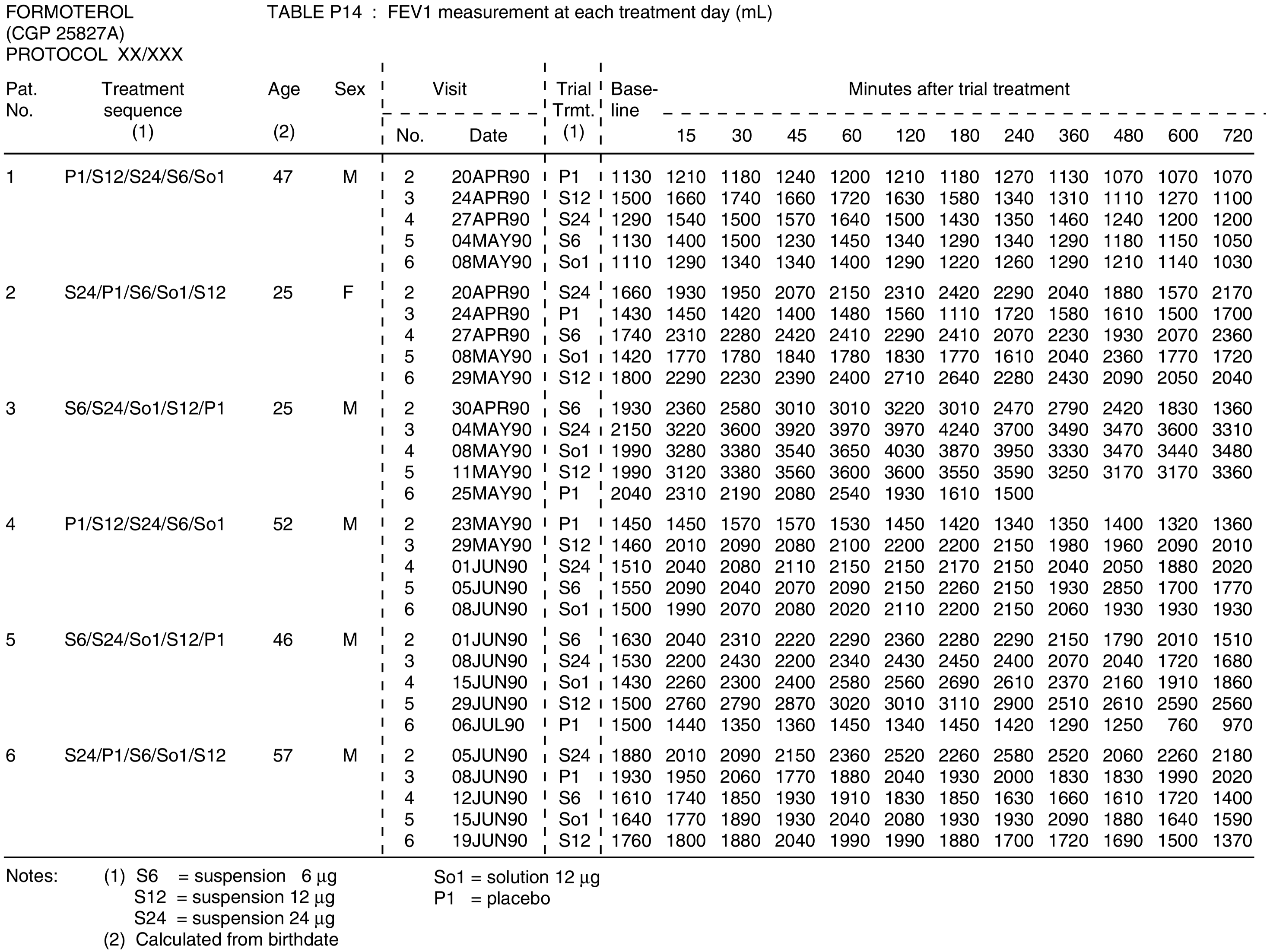
A second point is that the effect of different periods should be taken into account by any given graph unless it is considered to be *a priori* reasonable that this effect will be unimportant or the additional representation of this feature makes the graph unhelpfully cluttered. This second point is thus less important than the first. We would not carry not cross-over trials unless we believed between-patient variation was important. We *must* deal with it. We may or may not consider that period effects will be important.

We now consider various tabular and graphical representations which put these principles into practice.

8.2 PATIENT LISTINGS

Figure 8.1 is a facsimile of part of the patient listing for a clinical trial report of a cross-over trial (Example 8.1) of five treatments in five periods and 15 patients comparing three doses of formoterol suspension aerosol (6, 12 and 24 μg) to formoterol solution aerosol (12 μg) and placebo. The table was prepared according to a CIBA-GEIGY standard (Heath, 1991). Such reports are prepared for regulatory purposes and in addition to presenting the sponsor's analysis and interpretation of results, must provide, in considerable detail, a description of the trial (including the protocol) and all the data collected from the case record forms and/or patient diaries. These data are presented in the form of 'patient listings'. The purpose of these is to present faithfully the data obtained during the course of the trial so that, if necessary, the applicant's analyses and interpretations may be checked by the regulator.

As a result, the data are presented in a 'neutral' fashion. The purpose of the presentation is not to reflect the eventual analytic conclusions to which the data will lead (for this purpose a reorganization by treatment might be appropriate) but (roughly) the order in which the data were collected, so that here we have the data sorted by patient and for each patient by visit with the treatment received being recorded in addition. Also in this case, for the particular variable (FEV_1), the repeated measurements taken during the day are shown. Despite (or because of) their limitations such presentations do reflect the basic structure of the cross-over trial in terms of patients, period and treatments and (in this case) repeated measures. They thus provide the reader with all the information necessary to repeat the analysis. The same standard is not always observed in publications of the clinical trials in the medical literature. Of course, limitations on journal space make the listing of data in this degree of detail impossible. Nevertheless, where key data are presented in such publications, more care



FORMOTEROL
(CGP 25827A)
PROTOCOL XX/XXX

TABLE P14 : FEV1 measurement at each treatment day (mL)

Pat. No.	Treatment sequence (1)	Age (2)	Sex	Visit No.	Visit Date	Trial Trmt. (1)	Base-line	15	30	45	60	120	180	240	360	480	600	720
1	P1/S12/S24/S6/So1	47	M	2	20APR90	P1	1130	1210	1180	1240	1200	1210	1180	1270	1130	1070	1070	1070
				3	24APR90	S12	1500	1660	1740	1660	1720	1630	1580	1340	1310	1110	1270	1100
				4	27APR90	S24	1290	1540	1500	1570	1640	1500	1430	1350	1460	1240	1200	1200
				5	04MAY90	S6	1130	1400	1500	1230	1450	1340	1290	1340	1290	1180	1150	1050
				6	08MAY90	So1	1110	1290	1340	1340	1400	1290	1220	1260	1290	1210	1140	1030
2	S24/P1/S6/So1/S12	25	F	2	20APR90	S24	1660	1930	1950	2070	2150	2310	2420	2290	2040	1880	1570	2170
				3	24APR90	P1	1430	1450	1420	1400	1480	1560	1110	1720	1580	1610	1500	1700
				4	27APR90	S6	1740	2310	2280	2420	2410	2290	2410	2070	2230	1930	2070	2360
				5	08MAY90	So1	1420	1770	1780	1840	1780	1830	1770	1610	2040	2360	1770	1720
				6	29MAY90	S12	1800	2290	2230	2390	2400	2710	2640	2280	2430	2090	2050	2040
3	S6/S24/So1/S12/P1	25	M	2	30APR90	S6	1930	2360	2580	3010	3010	3220	3010	2470	2790	2420	1830	1360
				3	04MAY90	S24	2150	3220	3600	3920	3970	3970	4240	3700	3490	3470	3600	3310
				4	08MAY90	So1	1990	3280	3380	3540	3650	4030	3870	3950	3330	3470	3440	3480
				5	11MAY90	S12	1990	3120	3380	3560	3600	3600	3550	3590	3250	3170	3170	3360
				6	25MAY90	P1	2040	2310	2190	2080	2540	1930	1610	1500				
4	P1/S12/S24/S6/So1	52	M	2	23MAY90	P1	1450	1450	1570	1570	1530	1450	1420	1340	1350	1400	1320	1360
				3	29MAY90	S12	1460	2010	2090	2080	2100	2200	2200	2150	1980	1960	2090	2010
				4	01JUN90	S24	1510	2040	2080	2110	2150	2150	2170	2150	2040	2050	1880	2020
				5	05JUN90	S6	1550	2090	2040	2070	2090	2150	2260	2150	1930	2850	1700	1770
				6	08JUN90	So1	1500	1990	2070	2080	2020	2110	2200	2150	2060	1930	1930	1930
5	S6/S24/So1/S12/P1	46	M	2	01JUN90	S6	1630	2040	2310	2220	2290	2360	2280	2290	2150	1790	2010	1510
				3	08JUN90	S24	1530	2200	2430	2200	2340	2430	2450	2400	2070	2040	1720	1680
				4	15JUN90	So1	1430	2260	2300	2400	2580	2560	2690	2610	2370	2160	1910	1860
				5	29JUN90	S12	1500	2760	2790	2870	3020	3010	3110	2900	2510	2610	2590	2560
				6	06JUL90	P1	1500	1440	1350	1360	1450	1340	1450	1420	1290	1250	760	970
6	S24/P1/S6/So1/S12	57	M	2	05JUN90	S24	1880	2010	2090	2150	2360	2520	2260	2580	2520	2060	2260	2180
				3	08JUN90	P1	1930	1950	2060	1770	1880	2040	1930	2000	1830	1830	1990	2020
				4	12JUN90	S6	1610	1740	1850	1930	1910	1830	1850	1630	1660	1610	1720	1400
				5	15JUN90	So1	1640	1770	1890	1930	2040	2080	1930	1930	2090	1880	1640	1590
				6	19JUN90	S12	1760	1800	1880	2040	1990	1990	1880	1700	1720	1690	1500	1370

Minutes after trial treatment: 15 to 720.

Notes: (1) S6 = suspension 6 µg
S12 = suspension 12 µg
S24 = suspension 24 µg
So1 = solution 12 µg
P1 = placebo
(2) Calculated from birthdate

Figure 8.1 (Example 8.1) Extract from the ‘patient listings’ of a clinical trial report.

could be taken to present data in a form which will permit the reader to carry out analyses for himself.

When, as is the case in this trial and in many cross-over trials, repeated measures are taken within periods, a plot which corresponds roughly to a graphical representation of the patient listing is sometimes useful (Matthews *et al.*, 1990). This consists of plotting on one sheet of paper for each patient the response for each period over time. Figure 8.2 represents the plot over time of FEV_1 for patient 1 for Example 8.1. Such plots sometimes provide useful preliminary indications in choosing summary measures prior to data analysis. However, caution should be exercised since the shape of individual plots reflects not only the evolution of the effect of treatment over time, which is directly relevant to the choice of summary measure, but also any secular trend effects, which are not. This is because, to the extent that trends are common to all treatments, they play no part in their comparison and, to the extent that they are common to all patients, they contribute nothing to random variation either (Senn et al., 2000).

Sometimes it is desirable to be able to plot a number of such profiles on one sheet of paper. So-called 'trellis graphs' are then useful. Figure 8.3 shows a trellis plot of all six patients whose data are represented in Figure 8.1. In order to reduce the number of profiles the placebo results have been omitted and all results have been expressed in terms of difference with respect to placebo. This has the further advantage of controlling for secular time trends within periods (as discussed above), although it does not control for any possible period effects.

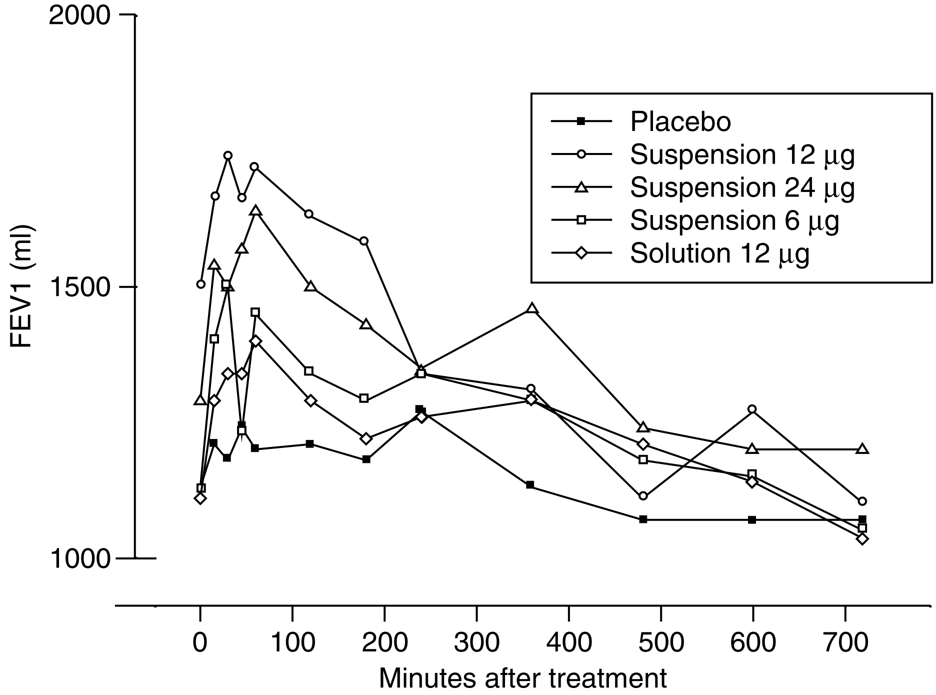


Figure 8.2 (Example 8.1) Forced expiratory volume in one second (FEV_1) for five treatments measured over 12 hours for patient 1.

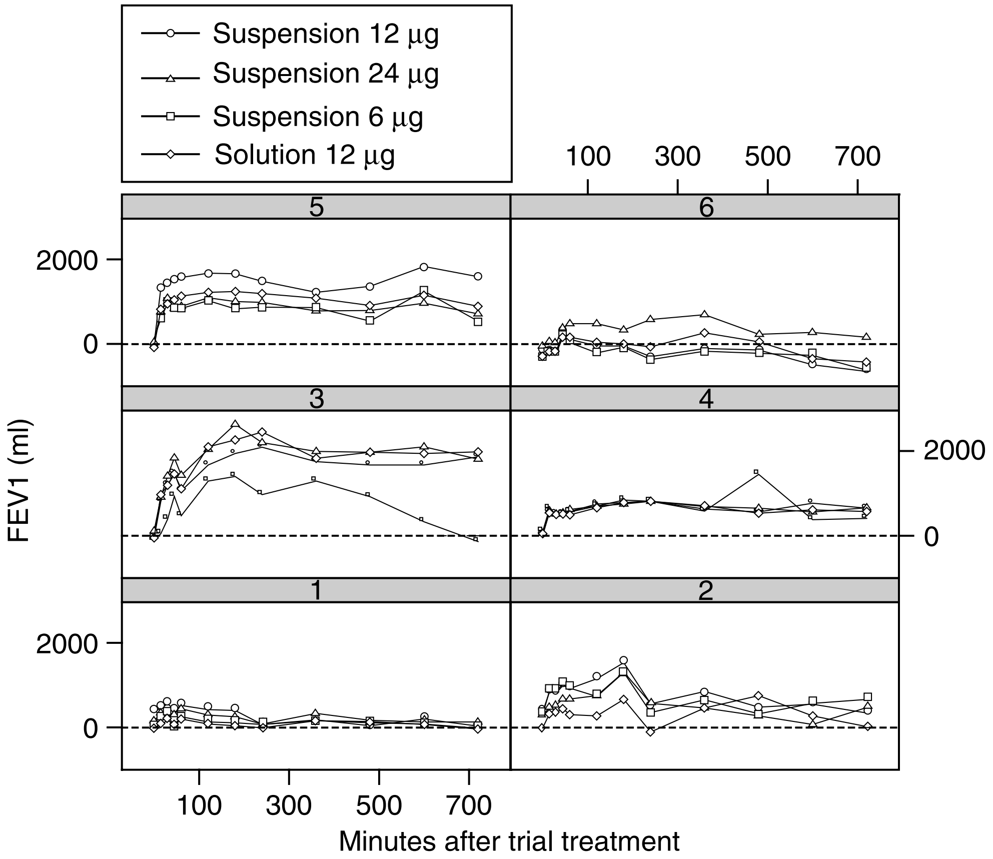


Figure 8.3

A potential disadvantage is that missing values under one treatment may make the differences impossible to calculate. This is, indeed, the case here for patient 3, for whom no values under placebo were recorded after four hours. This difficulty was dealt with here, for the purpose of producing the plots, by carrying forward the four-hour value for patient 3. Presentation of the results in terms of difference with respect to placebo highlights the rather unusual pattern for patient 6 who has several measurements under active treatment lower than the corresponding measurement under placebo.

8.3 RESPONSE BY PATIENT SCATTER-PLOTS

The reader has already encountered on a number of occasions throughout this book a type of graph which I have found very useful when presenting the results of a cross-over trial in a lecture. An example which we now consider is Figure 5.1 of Chapter 5. If raw data are to be presented with the purpose of showing a general picture rapidly rather than documenting in fine detail the individual data, then the patient listing is obviously not suitable. For this purpose the *response by patient* scatter-plot is useful.

For example, in Figure 5.1, the main outcome variable for the trial of Example 5.1 is represented in such a way that (1) the response is shown (the vertical axis of the graph) for each treatment (labelled by a symbol), (2) the responses have been collected together in such a way that they are sorted (a) by patient and (b) by sequence group. The graph thus shows response, treatment, sequence group and patient. Since the combination of sequence and treatment defines a period also, then for this variable all the features of the cross-over trial are illustrated.

In order to maximize the data–ink ratio the patient numbers are not shown since they are not really informative. The points will be seen to be arranged in columns of three corresponding to the results under the three treatments for a given patient. It can be seen quite clearly from this graph that for almost every patient, regardless of sequence, the formoterol reading is the highest and the placebo is the lowest. This graph, simply by organizing the data appropriately, shows much of what a formal analysis would show whilst still retaining the individual points. Of course *patient* is not a true dimension in the same sense as response: the patient number is purely arbitrary. It is used here as a pseudo-dimension simply in order to provide a visual separation of points which enables the viewer to group them efficiently.

8.4 TREATMENT BY TREATMENT SCATTER-PLOTS

The response by patient scatter-plot enables one to show a given response for all treatments from a cross-over trial. If only two treatments are being considered, as will be the case if an *AB*/*BA* design has been used or if we are illustrating one pairwise contrast only from a design with three or more treatments, then we can use each treatment response as a dimension and plot a value for each patient. We may refer to this sort of plot as a *treatment by treatment* scatter-plot.

Figure 8.4 represents a treatment by treatment scatter-plot for formoterol and salbutamol for Example 5.1, the three-treatment cross-over in six sequences shown in Figure 5.1 and discussed above. Here the symbols for the points are used to represent the six sequences whereas the formoterol response is measured on the vertical axis and the salbutamol response on the horizontal axis. A line has been drawn at 45° from the origin. This is the line of exact equality of the two treatments and we may regard a *law of the diagonal* as applying. Any patient plotted to the left of the line had a higher response under formoterol; any patient plotted to the right had a higher response under salbutamol. For many trials the sequence to which a patient belongs will be totally irrelevant and it may be adequate to simplify the graph removing any reference to it. This is probably the case here. However, the sequence identification has been left; not because it helps to interpret this particular trial but because it illustrates the technique.

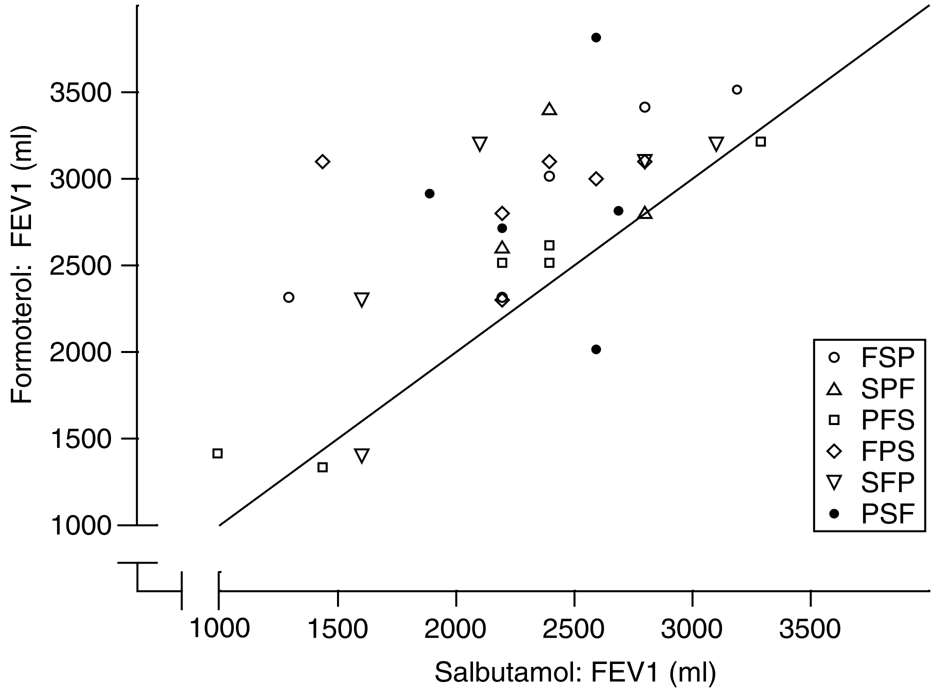


Figure 8.4 (Example 5.1) Forced expiratory volume in one second (FEV_1) under formoterol and salbutamol.

8.5 TREATMENT BY TREATMENT HISTOGRAMS

For ordered categorical data the equivalent representation to the treatment by treatment plot is a *bivariate treatment by treatment* histogram showing the frequency for each bivariate category defining the response under each of two treatments. Figure 8.5 is such a bivariate histogram for Example 8.2 (which has not been previously presented) which was a three-treatment cross-over in six sequences (three patients per sequence) comparing a single dose of two formulations of formoterol aerosol 12 μg (suspension and solution) to placebo. The figure represents the patient's overall evaluation of therapeutic effect at the end of the trial and shows the cross-classification for formoterol solution aerosol and placebo. Again the law of the diagonal applies. Patients to the left of the diagonal prefer one treatment; those to the right prefer the other. It may be seen that only one patient gave a higher rating to the placebo than to the active treatment. He rated the active treatment 'good' and the placebo 'very good'. Five patients (one 'poor–poor', three 'good–good' and one 'very good–very good') rated the two treatments equally. The other 12 patients preferred the active treatment. Distinctions between sequences have been ignored in the presentation.

Although solid bars have been used, we have not flouted Tufte's dimensional dictum if we allow that frequency is a legitimate dimension of the graph. Nevertheless, Tufte (1983) makes the point that some graphs are more usefully replaced by tables and this is perhaps a case. A bivariate table would probably

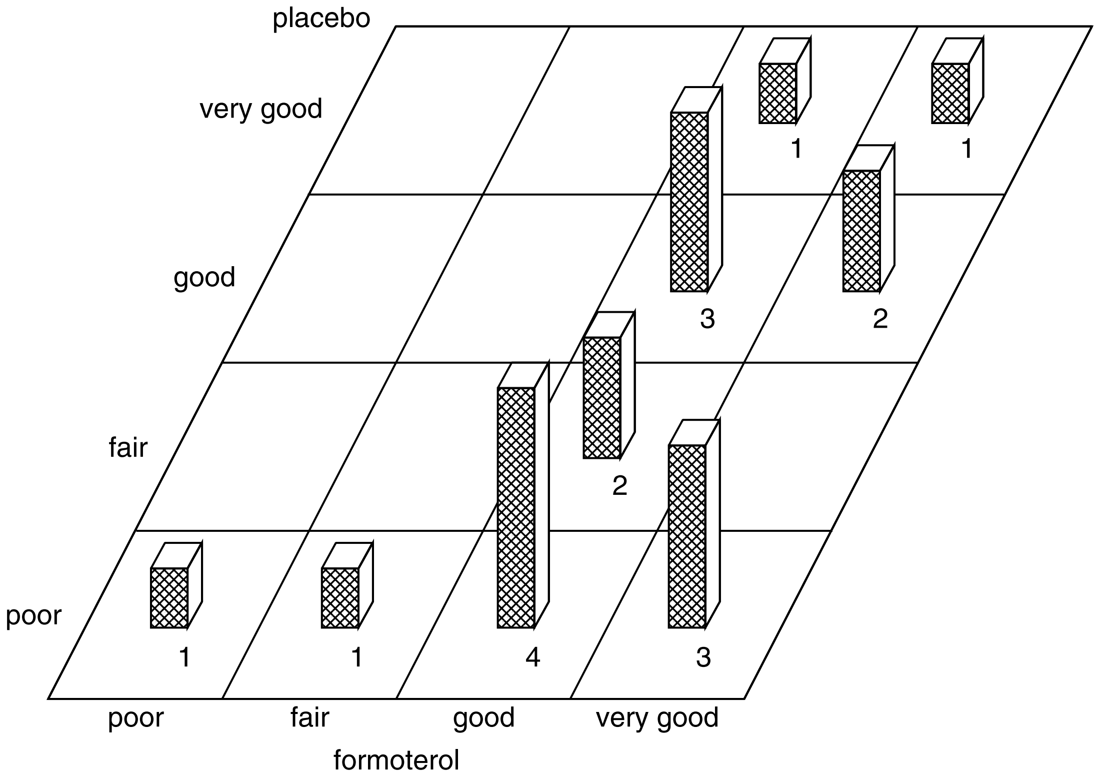


Figure 8.5 (Example 8.2) Bivariate histogram of overall evaluation of patients' opinion of therapeutic effect for two treatments from a three-treatment cross-over.

do as well. My only justification in presenting this diagram is that I have found it useful in replacing the pie charts or histograms which are commonly employed to represent such data from cross-over trials, and it does at least have the virtue, which they do not, of showing the within-patient structure of the trial.

8.6 BASIC ESTIMATOR SCATTER-PLOTS

On a number of occasions throughout the book we have calculated differences between results for two treatments for each patient and plotted the results. As we have already noted for the simple *AB*/*BA* design, such differences are often referred to as cross-over differences. We have used the term 'basic estimator' as a general term to describe such differences both for the *AB*/*BA* design and for other designs in general. Such basic estimators eliminate the between-patient differences and, for a given treatment contrast of interest, we then have the advantage that the within-patient nature of the cross-over trial can be represented using one dimension, freeing a dimension for other purposes.

Thus in Figure 8.4, in order to show the within-patient structure for Example 5.1, we plotted formoterol readings against salbutamol readings. For Figure 3.4, however, we had already calculated basic estimators (formoterol reading–salbutamol reading) for PEF at 8 hours and reduced the treatment difference to a one-dimensional feature. We then used another dimension to do the same for the baseline and so were able to plot one against the other, bringing out some further apects of the data, namely that for Example 3.1 the baseline differences were highly predictive of the differences at 8 hours but that nevertheless there appeared to be an important treatment effect. This is by no means the only sort of basic estimator scatter-plot which may be of use and to illustrate another possibility we consider this example in more detail.

Example 3.1 presented data for PEF 8 hours after treatment measured in an *AB*/*BA* cross-over in 13 patients (Graff-Lonnevig and Browaldh, 1990). This was the most important time point for this trial. However, measurements were made at a number of time points over a 12 hour period. Table 8.1 reproduces the data for PEF over 8 hours, this being the period during which patients were present in the clinic (they then travelled home and continued measurements there). The values marked with an asterisk have been calculated as the average of readings before and after a missing value which they replace. These data were given by Senn and Auclair (1990) and we shall refer to this fuller data set as Example 8.3.

Figure 8.6 presents a basic estimator scatter-plot for area under the curve (*AUC*) and trend *AUC* for Example 8.3. The necessary statistics were calculated as follows. First, for each patient for each treatment the *AUC* for the first four hours (AUC_1) and the last four hours (AUC_2) were calculated (using the trapezoidal rule). Second, for each patient for each treatment these two figures were combined as follows: (a) they were summed to produce $AUC = AUC_1 + AUC_2$; (b) the difference was calculated to give trend $AUC = AUC_2 - AUC_1$. Finally, basic estimators were calculated for both of

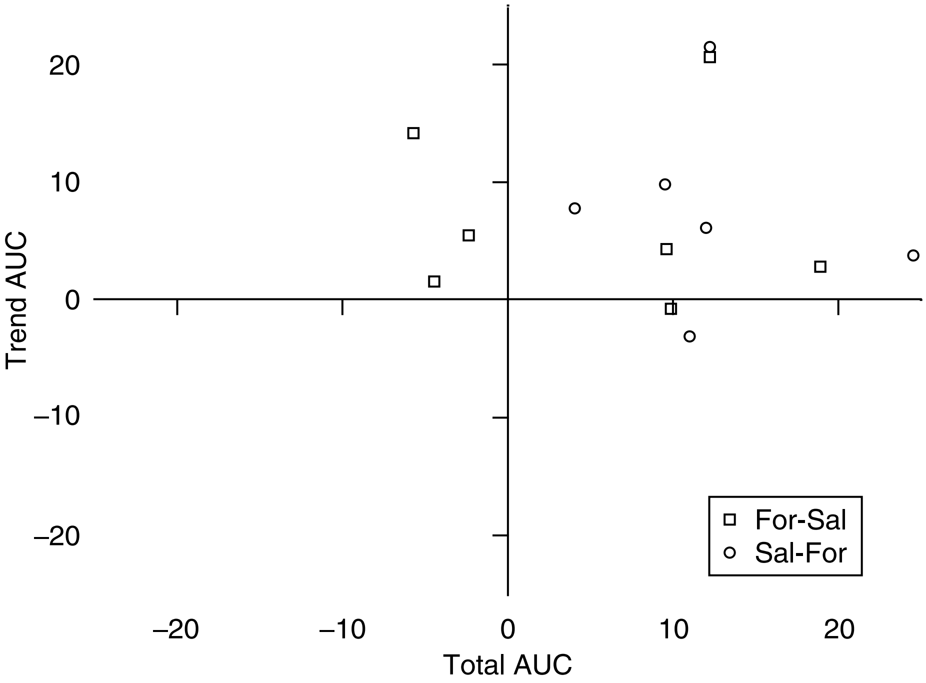


Figure 8.6 (Example 8.3) Area under the curve (*AUC*) and trend-*AUC* for peak expiratory flow (PEF) for an *AB*/*BA* cross-over.

Table 8.1 (Example 8.3) Peak expiratory flow PEF (1/min) in a two-period cross-over trial of formoterol 12 μg (for) and salbutamol 200 μg (sal) in children with asthma.

Sequence Patient Treatment	Time after dose (minutes)											
	Base	20	40	60	90	120	180	240	300	360	420	480
For-Sal												
1												
for	290	305	315	310	340	320	285	280	285	300	280	310
sal	270	330	310	320	330	320	320	300	300	300	300	270
4												
for	300	320	320	340	340	320	330	320	310	320	310	310
sal	270	320	320	330	310	330	310	300	300	270	290	260
6												
for	250	350	350	370	370	380	390	370	360	370	380	370*
sal	210	370	390	370	370	360	300	280	340	320	350	300*
7												
for	390	410	410	410	410	430	410	410	420	410	410	410
sal	390	430	440	430	420	440	430	420	400	410	410	390
10												
for	250	250	260	265*	270	280	270	250	250	250	250	250
sal	240	250	245	260	250	270	230	200	240	240	240	210
11												
for	365	440	410	400	420	410	340	390	400	400	390	380
sal	380	435	450	450	440	400	450	420	390	360	370	350
14												
for	190	315	320	325	330	355	345	370	305	340	370	330
sal	260	405	410	420	405	380	300	260	310	210	260	365

Sal-For												
2												
for	345	365	390	375	390	380	395	390	395	395	385	385
sal	350	385	390	395	400	390	390	400	365	365	350	370
3												
for	370	370	360	360	380	360	380	350	360	360	380	400
sal	350	370	380	375	380	365	330	320	320	350	340	310
5												
for	360	400	400	410	420	420	420	430	430	410	410	410
sal	350	420	425	420	420	430	410	400	370	380	370	380
9												
for	290	340	340	340	330	340	350	340	340	330	350	320
sal	280	310	300	310	310	310	320	300	330	350	310	290
12												
for	310	340	350	340	350	350	350	350	340	360	350	340
sal	270	320	310	300	310	300	300	300	300	300	280	260
13												
for	220	250	240	240	240	250	250	250	230	220	235	220
sal	220	270	285	270	290	270	270	210	190	180	120	90

these measures for each patient by subtracting the salbutamol readings from the formoterol ones.

It can be seen in Figure 8.6 that, on the whole, the points lie in the upper right quadrant. The fact that they are in the right half of the graph shows that the area under the curve as a whole is higher for formoterol. The fact that the points lie in the top half of the graph shows that the contribution of the second four hours is relatively more important for formoterol than salbutamol.

There are many other possibilities for plots involving basic estimators for repeated measures designs. Thus, for a bioequivalence study, we could plot the area under the curve basic estimator against the $C_{\max}$ basic estimator for a given treatment contrast patient by patient. Or, for a pharmacodynamic study, we could plot one clinical outcome against another: for example in an asthma study the basic estimator for FEV_1 against the basic estimator for PEF. Yet another possibility is to plot the basic estimators at each time point over time.

8.7 EFFECT PLOTS

A very common representation of the results of cross-over trials with repeated measurements over time is to plot for each time point the mean response for each treatment. Thus, for example, for a three-treatment cross-over we might produce three traces on one sheet of paper showing the mean response under each treatment over time. There is nothing particularly objectionable about this sort of plot although it is not very informative. It is true, however, that if the cross-over is balanced for sequence, or alternatively if the period effect is negligible, then the difference between two traces will correspond to a treatment effect. What is objectionable is if standard error bars are added to the mean trace. If the investigator feels that it is appropriate to indicate the variability of the experiment, then in a cross-over trial it is the *within-patient variability* that ought to be shown, not the *between-patient variability*. Standard errors of individual treatment means, however, will have been calculated on a between-patient basis.

A simple solution then, is to calculate at each time point the treatment effect, that is to say the treatment estimate for a contrast of interest, using the particular analytic technique favoured and, also using this technique, the appropriate standard error, or even better a confidence interval (Senn and Auclair, 1990). If the CROS analysis of Section 3.6 is used at each time point for Example 8.3 then the treatment estimate and associated confidence intervals may be plotted against time to produce the graph given by Figure 8.6. A horizontal reference line of exact equality has been included.

This graph can be criticized on the grounds that it encourages a proliferation of analyses since it presents an analysis at every time point. It might be argued

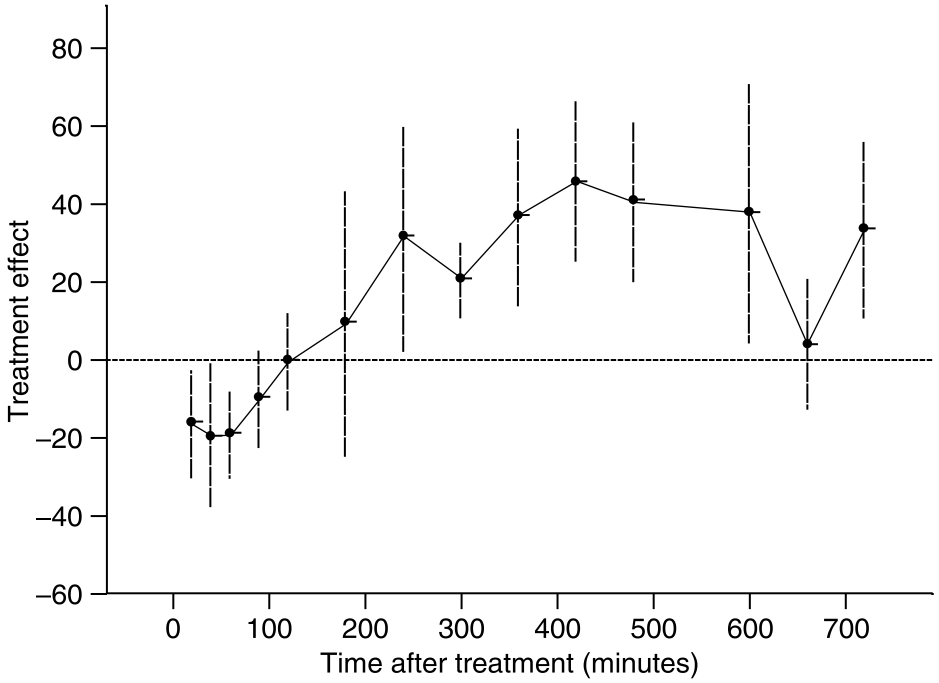


Figure 8.7 (Example 8.3) Treatment estimates and 95% confidence limits for peak expiratory flow (PEF) plotted against time.

that the data should have been reduced by first using some summary measures approach (Matthews *et al.*, 1990). For example, for a formal analysis, it might be preferable to have calculated the area under the curve for each treatment for each patient first (as we did above) and then use a CROS analysis. This example, however, concerns a *duration of action* trial. The time for which the treatment is effective is of primary interest. (The 8 hour reading was of more interest than that at 12 hours simply because it was regarded as more reliable since the patients left the clinic after 8 hours and continued their measurements at home.) Under such circumstances it might be imagined that the summary measure *time to return to baseline* could be used for an analysis. The difficulty with this measure, however, is that it is very vulnerable to any uncontrolled features of the trial. For example, over the day there might be a strong downward trend. As long as the PEF readings under one drug were compared directly to the readings under the other this would not be particularly problematic. If the comparison to baseline were substituted for this, however, and the comparisons between treatments made at one stage remove, this could be very misleading. Under such circumstances, therefore, a comparison of treatments directly at each time point may be of considerable interest. Nevertheless, the fact that this graph has been described here should not be taken as a general endorsement of indiscriminate multiple testing. It is a tool for exploratory analysis which can be extremely useful on occasion and is definitely preferable to the treatment summaries with between-patient standard errors which it was designed to replace.

8.8 TREATMENT CENTRED RESIDUAL PLOTS

In Section 8.6 we covered the plotting of basic estimators. If we consider the factors which a basic estimator reflects they are: (1) the overall treatment effect, (2) that part of the patient's response to treatment which is individual and not explained by the first factor, (3) 'pure' within-patient random variation, (4) a period effect determined by the sequence to which the patient was allocated, (5) any individual trend over periods not expressed by the fourth factor and (6) other effects such as period by treatment interaction and carry-over which we assume to be negligible. A seventh factor which affects the experiment as a whole, namely between-patient variation, is eliminated by the basic estimator.

In an *AB*/*BA* cross-over, the second, third and fifth factors above cannot be separated and are usually referred to as within-patient variation. In designs with more than two measurements per patient such effects can be separated. In fact, one of the uses of *n* of 1 trials is to do just that: we wish to establish something about a patient's response to treatment. The basic estimator approach, however, reduces all measurements on a patient to a single treatment difference reflecting the contrast of interest and for the purpose of comparing any two given treatments all other treatments are ignored. Under such circumstances, therefore, just as for the *AB*/*BA* design, factors two, three and five need not be separated. Hence, if we ignore the sixth factor as being negligible, we may reduce the list to three: treatment effects, period effects and within-patient variation.

If we now go one step further and eliminate the period effect from the basic estimators (how this may be done will be considered in due course) we are then left with values which reflect the treatment effect and the within-patient variation. Such statistics may be referred to as treatment-centred residuals (Senn and Auclair, 1990) since, for an *AB*/*BA* cross-over, they may be equivalently obtained by using the residuals from fitting the full model with patient, period and treatment effects and then adding the overall treatment estimate to the residual.

To illustrate how this may be done, we use Example 8.3, taking the measurements at 8 hours. We already have the basic estimators which we calculated in Table 3.2. Part of this table is reproduced in Table 8.2, which illustrates the calculations which must be done to produce the treatment-centred residuals. The steps are as follows. Having obtained the basic estimators, we average them within each sequence group. The residuals are calculated by subtracting the appropriate group mean from each value. Since all the basic estimators within a group reflect the same period effect, subtracting their mean removes this effect from them. It also, of course, removes the treatment effect. Hence, we are left with residuals from the usual cross-over model which can be seen to sum to zero within each group (barring rounding errors). Since, however, the period effect from the two groups must sum to zero, then the average of the two groups

Table 8.2 (Example 8.3) Peak Expiratory Flow (PEF) in l/min measured 8 hours after treatment: basic estimators and treatment centred residuals.

		PEF		
Sequence	Patient	Basic estimator (a)	Residual ($d = a - b$)	Treatment-centred residual ($e = d + c$)
for/sal	1	40	9.29	55.9
	4	50	19.29	65.9
	6	70	39.29	85.9
	7	20	−10.71	35.9
	10	40	9.29	55.9
	11	30	−0.71	45.9
	14	−35	−65.71	−19.1
	Mean	$30.71^{(b)}$	0.0	46.6
sal/for	2	15	−47.5	−0.9
	3	90	27.5	74.1
	5	30	−32.5	14.1
	9	30	−32.5	14.1
	12	80	17.5	64.1
	13	130	67.5	114.1
	Mean	$62.5^{(b)}$	0.0	46.6
	Treatment estimate	$46.61^{(c)}$		

means, 46.6 ℓ/min, does not reflect the period effect but only the treatment effect (and of course random variation). This figure is, of course, the treatment estimate we obtained in Section 3.6. Finally, adding this figure to the residuals produces the treatment-centred residuals.

As we have already explained, the figures thus produced reflect the treatment effect and also within-patient variation, where this latter term, in this context, covers *all the random influences not eliminated by using each patient as his own control*. The treatment residuals can now be plotted in a suitable form. There are many possible candidates for a suitable plot. Two common plots are the scatter-plot and the box-and-whisker plot (Tukey, 1977; McNeill, 1977). Both of these possibilities are illustrated in Figure 8.8.

The scatter-plot should be self explanatory. The Tukey box-and-whisker plot gives the lower quartile, median and upper quartile (the box) the lowest and highest values not considered an outliers (the whiskers) as well as representing any outliers individually by asterisks. An outlier is defined as any value greater than the sum of the upper quartile and 1.5 times the interquartile range or less than lower quartile minus 1.5 times the interquartile range. For a Normal population this would define 0.7 per cent of all values as outliers. It might be argued that this definition tends to reduce the information in the whiskers since

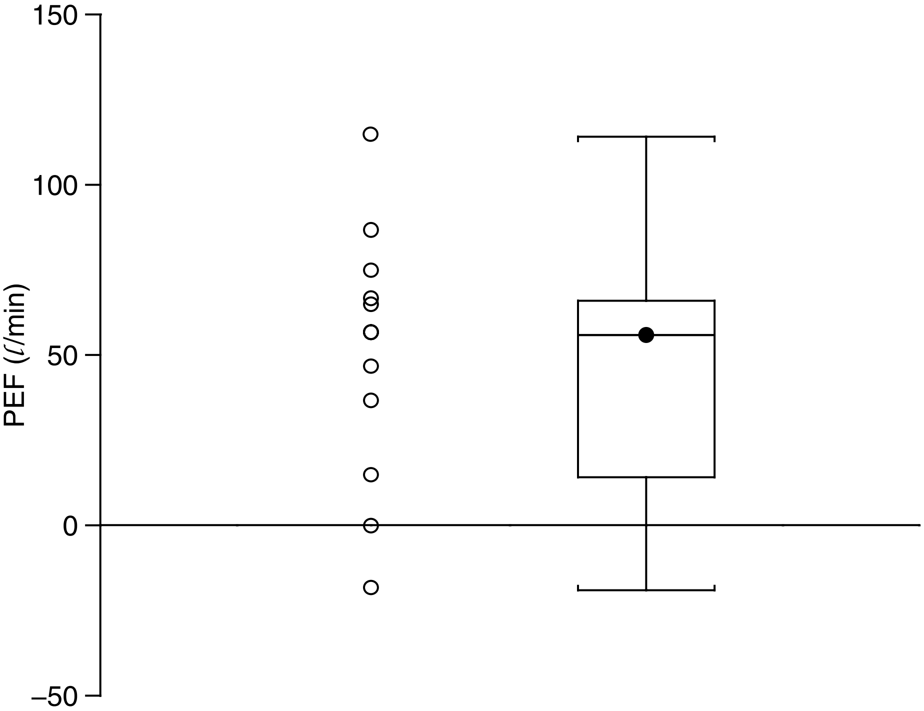


Figure 8.8 (Example 8.3) Scatter-plot and box-plot for treatment centred residuals for peak expiratory flow (PEF) at 8 hours. (See text for explanation.)

they are constrained by the box. Possibly a superior definition would be to use further percentiles for the whiskers. The box has the 25th, 50th and 75th percentiles. Perhaps a further halving of the remaining area would be useful to produce 12.5% and 87.5% marks. This discussion will not be pursued further. An interesting illustration of the uses to which box-plots may be put in presenting data from clinical trials will be found in Dietlein (1981).

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

9

Various Design Issues

9.1 INTRODUCTION

Errors in analysis are serious but, if detected, may be redressed; unless detected before a trial is run, errors in design are irredeemable. If a trial is run with a flawed design it will have to be supplemented by another. In running clinical trials we impose a high price in inconvenience on patients. It is the investigator's responsibility to make the best scientific use of the cooperation which is granted.

Ideally all clinical trials would be planned jointly by a physician and a statistician. This was already the norm within the pharmaceutical industry well before 1993, when the first edition of this book appeared, at which time the pharmaceutical industry, whether in the USA, Europe or in Japan (albeit to a lesser extent), already employed statisticians in considerable numbers. Before the 1990s, however, regulatory agencies—with a very few exceptions, among them the US Food and Drug Administration (FDA)—did not employ statisticians (Pocock *et al.*, 1991). Since then, although there is considerable progress still to be made (EFSPI Working Group, 1999), the influence of statisticians within drug development has grown considerably. In particular, European regulatory agencies, although not matching the FDA in degree of commitment to biostatistics, have gradually been employing statisticians in greater numbers. This has been accompanied by a major international regulatory development, the International Conference on Harmonisation (ICH). This has brought together regulators and sponsors from Europe, America and Japan in order to agree common practice regarding drug development and regulation. At the time of writing, some 40 to 50 guidelines are in various stages of completion. Many of these have considerable statistical content and have benefited from input by statisticians. The most important in this respect is ICH E9 (International Conference on Harmonisation, 1999) which covers statistical principles. Much of that document is taken up with the issue of planning, which forms the subject of this chapter. The document states (section 1.2): 'it is assumed that the actual responsibility for all statistical work associated with clinical trials will lie with an appropriately qualified and experienced statistician, as indicated in ICH E6'. The document to which it refers (International Conference on Harmonisation,

1996) is a guideline on good clinical practice which in turn states (section 5.4.1): 'The sponsor should utilize qualified individuals (e.g. biostatisticians, clinical pharmacologists, and physicians) as appropriate, throughout all stages of the trial process, from designing the protocol and CRFs and planning the analyses to analyzing and preparing interim and final clinical trial reports.' It is thus well enshrined in regulatory requirements that the statistician should be involved in planning clinical trials, and various guidelines of the Statisticians in the Pharmaceutical Industry (PSI), for example, give advice as to how standard operating procedures may be implemented by pharmaceutical sponsors to ensure that statistical standards in planning and analysing trials are maintained (North 1998; Phillips *et al.*, 2000; PSI Professional Standards Working Party, 1994).

However, the academic or hospital-based life scientist may not find it so easy to find statistical collaborators. Furthermore, the statistical standards required by journals for publication being generally considerably lower than those of drug regulatory agencies, there is not the same pressure from statisticians to become involved (Senn, 2000d). Also, academic statisticians generally gain little credit for publications that simply involve applications of statistics and so have little incentive for collaboration. In practice, therefore, some clinical trials outside the pharmaceutical industry will continue to be carried out with very little or even no input from a statistician. Where this is so, the investigator must learn to think like a statistician. Where this is not the case, and the statistician co-operates with the physician (or biologist), the statistician must learn to think like a physician (or biologist). That is to say, he must strive to achieve some minimum understanding of the field of application.

Any cursory survey of the methodological literature on cross-over trials, which consists almost entirely of the publications of statisticians, and the applications literature on cross-over trials, which consists almost exclusively of the publications of life scientists (if this term may be used for physicians, biologists, pharmacists etc.), will show that there is only a minimal consideration of biology in the former and a lack of reference to the methodological literature in the latter. This is a very unsatisfactory state of affairs which is not confined to cross-over trials (Altman and Bland, 1991) but is particularly acute in this field. In my opinion, the responsibility for it is shared between statisticians and life scientists (Senn, 1997c).

The designing of clinical trials presents a continual challenge. Different indications, different treatments and different purposes require different designs. Some of the issues affecting the design of trials have been raised throughout this book. The purpose of this chapter is to present in more depth various design issues which affect cross-over trials. The intention is not so much to cover the field but to give the reader an introduction to some matters which may require consideration when designing such trials.

In a conventional treatment of cross-over trials, two issues would receive particular attention: choosing sequences to deal with carry-over and determining the number of patients needed (e.g. power calculations). In my opinion the approach which has been adopted to the first is fundamentally misguided. It will therefore be critically examined separately in Chapter 10. As regards the second issue, this is not unimportant but it is sometimes given too much weight. I hope to show in this chapter that there are other issues which may be considered as well. On the whole, the design issues which are covered will be those which are not shared with parallel group trials. There are thus many important issues which will be left out. The reader is encouraged to consult Pocock's (1983) excellent book on clinical trials to complete the picture. A useful summary of the most important features of a clinical trial protocol, including a valuable checklist, is given by Johnson (1989). A discussion of many statistical issues that arise in drug development will be found in Senn (1997b).

One other point needs mentioning. In the pharmaceutical industry a clinical trial protocol will usually contain a *detailed* statement of the intended analysis (PSI, 1991). This is a practice which could be imitated elsewhere with profit. It thus follows that the planned analysis is also part of the design. As Johnson (1983, p. 6) puts it, 'the method of analysis should be anticipated in the protocol.' Some of the issues which are treated in this chapter will be concerned with designing the analysis.

9.2 PARALLEL GROUP OR CROSS-OVER?

In Sections 1.2–1.4 we considered reasons as to why cross-over trials are performed. We now look at this issue in more detail. Whereas in many fields of application a cross-over trial is not a reasonable alternative to the parallel group trial, in almost every case where a cross-over trial is performed a parallel group trial might be considered instead. It is thus reasonable to consider first of all, before designing a cross-over trial, whether a parallel group trial might not better serve the purpose. Obviously, in considering this question, the sort of general issues discussed in Sections 1.2–1.4 would be borne in mind. We shall not rehearse these here but instead will cover the issue of the precision of the two types of design for those cases where either is a viable option.

To do this we shall suppose that we wish to compare two treatments, A and B, and would like to choose between two possible designs: first a parallel group design (parallel for short) in which $2p$ patients are assigned at random to treatments A and B, p to each treatment, and secondly an AB/BA cross-over design (cross-over for short) in which $2c$ patients are assigned at random to sequences AB and BA, c to each sequence. In either case we shall assume that randomization takes place just before treatment commences and that each treatment period is t time units long. For the cross-over there is a wash-out period of length w between treatments, where w might be zero. We shall assume that a single measurement is taken at the end of each treatment period and that for the cross-over the effect of the first treatment has disappeared by the time the measurement at the end of the second treatment is taken. Suppose also that the time between recruitment of individual patients is r_1 for the parallel and r_2 for the cross-over. Further suppose that the time spent in the trial before randomization, including, any pre-randomization visits is v. We shall refer to this period as the *run-in*.

Remark This model is a considerable simplification but it is hoped that it is adequate for the purposes of discussion. In many cross-over trials w is not fixed and a minimum value is specified but what constitutes an adequate minimum value for w will in any case be a matter for speculation. Also a matter for speculation in any trial (although this point seems rarely to have been recognized) is what constitutes an appropriate value for t although, of course, the investigator will have to commit himself in advance as to what length of treatment he wishes to study. The recruitment times r_1 and r_2 would not be known in advance of conducting the trial and would in any case be variable within trials. According to Johnson (1989, p. 15). 'The principal difficulty encountered in the design of most clinical trials is the accurate estimation of recruitment rates.' It might be expected that on average r_2 would be greater than r_1 since the total inconvenience to the patient is greater in the cross-over trial and it may be thus more difficult to find willing subjects. There are occasions, however, when the guaranteed opportunity to try each treatment may be a recruitment advantage.

Some relevant times may now be calculated using this model and these are given in Table 9.1. The time that each patient spends in the trial follows simply from the definitions given above. This time may be regarded as an index of the cost to the individual patient for participating in the trial. The total investigation time is the product of the time spent by each patient and the number of patients in the trial. This may be regarded as an index of the work which the investigator(s) may have to undertake to complete a trial. It is not an elapsed time. Finally, the total trial time is the time from beginning the recruitment of the first patient until completing the treatment of the last. It is thus equal to the time until the last patient is recruited, which is the product of the number of patients and the recruitment time per patient, plus the time taken to treat the last patient

Table 9.1 Various times compared for cross-over and parallel group designs.

	Parallel	Cross-over
Time that each patient spends in the trial	$v + t$	$v + 2t + w$
Total investigation time	$2p(v + t)$	$2c(v + 2t + w)$
Total trial time	$2pr_1 + v + t$	$2cr_2 + v + 2t + w$

once recruitment has finished. In a commercial setting this time may be regarded as a contribution towards the total time to registration of a drug and is thus an index of cost in terms of lost sales.

We may draw the following conclusions from Table 9.1.

- The extra time each patient spends in the trial in the cross-over compared to the parallel is $t + w$, or the time of one treatment plus the time of a wash-out.
- The total investigation time for the cross-over will be less than that for the parallel providing

$$p > \{1 + (t + w)/(t + v)\}c. \tag{9.1}$$

In the simple case where the run-in time is the same as the wash-out time (most simply if both are zero) so that $v = w$, then (9.1) reduces to

$$p > 2c. \tag{9.2}$$

- The total trial time for the cross-over will be less than that for the parallel providing

$$p > cr_2/r_1 + (t + w)/(2r_1). \tag{9.3}$$

In the case which most favours the cross-over, where there is no wash-out and $w = 0$, and assuming that $r_1 = r_2 = r$ then this reduces to

$$p > c + t/(2r). \tag{9.4}$$

In order to pursue the issue of the relative efficiencies of these two designs further we now need to consider the precision of estimates from the two designs and therefore to develop a model to do so. Similar arguments will be found in Chassan (1970) and Hills and Armitage (1979).

9.2.1 The precison of cross-over and parallel group trials*

In order to consider the precision of the two trials we need to introduce a model which is based on that which we considered in Section 3.9. We shall assume, however, that there is no carry-over and we shall treat patient effects as random.

The expected responses for the cross-over are then as given in Table 9.2. The expected responses for the parallel are those given for the first period (i.e. the first column) of the table.

As regards these *expected* responses, however, we no longer separately identify patients apart from noting which group they are in. Unlike Table 3.7,

Table 9.2 Expected responses for a cross-over and parallel group design.

Period	
1	2(cross-over only)
A	B
$\mu + \tau$	$\mu + \pi$
B	A
μ	$\mu + \pi + \tau$

therefore, which gave the expected responses for given specific patients i and j, Table 9.2 gives the expected response for a sequence group (in the case of the cross-over) or treatment group (in the case of the parallel) in a given period. If we let the response on patient i in group $h(h = 1$ *or 2) and period* t ($t = 1$ for the parallel, $= 1$ or 2 for the cross-over), be Y_{iht}, then we may write

$$Y_{iht} = E(Y_{ht}) + \beta_{ih} + \varepsilon_{iht}, \tag{9.5}$$

where β_{ih} is the effect due to patient i in group h (a 'between-patient error', now considered random) and ε_{iht} is a disturbance term (a 'within-patient' error). We also assume $E(\beta_{ih}) = 0$, from which $E(\varepsilon_{iht}) = 0$, and that var $(\beta_{ih}) = \sigma_b^2$ and var $(\varepsilon_{iht}) = \sigma_w^2$. We assume in addition that β_{ih} and ε_{iht} are independent, from which it follows that their covariance is zero. We also assume that all ε_{iht} are independent of each other.

Remark This is by no means the only model that might be envisaged, although it is one which is commonly used. The model says that the only effects of any consequence are those due to treatments or periods and that all random variation can be split into two sources: the variation between patients and a variation within patients. It is assumed that all patients are equally variable. The model also assumes that if there are period or treatment effects these do not vary from patient to patient. We shall consider the consequences of relaxing these assumptions in due course.

If we denote the four cell means by

$$\begin{array}{cc} \bar{Y}_{.11} & \bar{Y}_{.12} \\ \bar{Y}_{.21} & \bar{Y}_{.22} \end{array}$$

then the standard estimator of the treatment effect for the AB/BA crossover, the *CROS* estimator (Freeman, 1989) of Section 3.6, is

$$CROS = [\{\bar{Y}_{.11} - \bar{Y}_{.12}\} + \{\bar{Y}_{.22} - \bar{Y}_{.21}\}]/2. \tag{9.6}$$

Noting that if (9.6) is expressed in terms of (9.5) the terms in β_{ih} disappear from this expression and applying our rules for linear combinations from Section 2.2.1 then we obtain the following expression for the variance:

$$\mathrm{var}\,(CROS) = \sigma_{\mathrm{w}}^2/\mathrm{c}. \tag{9.7}$$

On the other hand for the parallel we use the first periods only and hence the estimator of the treatment effect is

$$PAR = \bar{Y}_{.11} - \bar{Y}_{.21}$$

(Freeman, 1989).

Again, applying our rules for linear combinations, we obtain a variance of this estimator of

$$\mathrm{var}\,(PAR) = 2(\sigma_b^2 + \sigma_{\mathrm{w}}^2)/p \tag{9.8}$$

We now consider how many patients we have to recruit to the parallel to have the same variance of the *PAR* estimator as for the *CROS* estimator in the cross-over. We thus set (9.8) = (9.7) so that

$$2(\sigma_b^2 + \sigma_{\mathrm{w}}^2)/p = \sigma_{\mathrm{w}}^2/c$$

from which we have

$$p = 2c(\sigma_b^2 + \sigma_{\mathrm{w}}^2)/\sigma_{\mathrm{w}}^2. \tag{9.9}$$

Remark The expression (9.9) defines the condition under which the treatment estimate from a parallel group trial will have the same variance as the treatment estimate from a cross-over trial. Note that even if the between-patient variance, σ_b^2, is smaller than the within-patient variance, σ_{w}^2, more patients will be required for the parallel. This is because the parallel *is also subject to within-patient variation as is clear from* (9.8) (this point is often overlooked) and because two observations are obtained from every patient in the cross-over.

Substituting (9.9) in (9.1) we see that the condition that the total investigation time is less for the cross-over is

$$2(\sigma_b^2 + \sigma_{\mathrm{w}}^2)/\sigma_{\mathrm{w}}^2 > 1 + (t + w)/(t + v) \tag{9.10}$$

In the special case given by (9.2) where there is no wash-out or run-in this reduces to

$$(\sigma_b^2 + \sigma_{\mathrm{w}}^2)/\sigma_{\mathrm{w}}^2 > 1.$$

Clearly, therefore, in such a case the total investigation time will always be lower for the cross-over. Another interesting case is the trial with a wash-out equal to the treatment period (so that $w = t$) but with no run-in. Then we find

$$(\sigma_b^2 + \sigma_w^2)/\sigma_w^2 > 3/2.$$

This is a condition which will be reached providing the between-patient variation is at least half that of the within-patient variation. In most cases this will be so.

To establish the relative efficiency of the two trials in terms of total trial time we substitute (9.9) in (9.3), obtaining

$$2c(\sigma_b^2 + \sigma_w^2)/\sigma_w^2 > cr_2/r_1 + (t + w)/(2r_1). \tag{9.11}$$

To gain some feel for the practical importance of this formula let us consider two examples.

Example 9.1 Suppose that we have a single-dose trial (rather similar to many we have considered so far) in which the patients are observed during one day, so that $t = 1$ day, and return at weekly intervals, so that $w = 6$ days. Suppose, furthermore, that recruitment is just as easy for a parallel as for a cross-over so that $r_1 = r_2 = r$ and that between-patient variation is of the same size as within-patient variation so that $\sigma_b^2 = \sigma_w^2$. If we substitute these values in (9.11) we get $c > 7$ days/$(6r)$ or, equivalently, $r > 7$ days/$(6c)$. Even for a small cross-over with 7 patients per sequence group (which would correspond to a parallel with 28 patients per treatment), providing the time between recruiting one patient and the next is more than 1/6 days or 4 hours the cross-over will have the shorter trial time. This is, of course, almost certain to be the case.

Example 9.2 For the second example, suppose that recruitment times and variances are again identical but we wish to run a multi-dose trial with individual treatment periods of 12 months, so that $t = 12$ months, and no wash-out, so that $w = 0$. Substituting these values in (9.11) we get $c > 12$ months/$(6r)$, or $r > 2$ months/c. If we had a cross-over with some 30 patients per group, then a correspondingly powerful parallel would have 120 patients per group. Provided we recruit a patient every two days, the total trial time for the parallel will be less than for the cross-over. Since a parallel with 240 patients in total might well have at least 15 centres this sort of target would be achieved providing each centre could recruit a patient a month.

9.2.2 Analysis of covariance or cross-over?

The comparison of cross-over and parallel we have made is unfair to the parallel in this respect. Whereas we have allowed that all of the between patient variation may be removed in the cross-over, by always comparing treatment results within patients, we have assumed that none of this variation may be removed in the case of the parallel. In practice the use of concomitant observations, of which baseline readings are probably the best example, may be used to eliminate some of the variation (Hills and Armitage, 1979).

An efficient way to do this is using analysis of covariance (Senn, 1989a). For example, if we have a baseline measurement. Y_{ih0}, available for patient i in group h at the beginning of a trial and if a corresponding model to (9.5) for such measurements is

$$Y_{ih0} = E(Y_{h0}) + \beta_{ih} + \varepsilon_{ih0}, \tag{9.12}$$

where β_{ih}, ε_{ih0} and ε_{ih1} are independent and the variance of ε_{ih0} is equal to σ_w^2 then, if we use the symbol ρ for $\sigma_b^2/(\sigma_b^2 + \sigma_w^2)$, where $0 \leq \rho \leq 1$, the statistic

$$AOC = \bar{Y}_{.11} - \bar{Y}_{.21} - \rho\{\bar{Y}_{.10} - \bar{Y}_{.20}\} \tag{9.13}$$

will be an unbiased estimate of τ, the treatment effect, since $E(Y_{10}) = E(Y_{20})$, and will have a variance of

$$\text{var}(AOC) = 2(1 - \rho^2)(\sigma_b^2 + \sigma_w^2)/p. \tag{9.14}$$

But for the factor $(1 - \rho^2)$ which must lie between 0 and 1, this is identical to (9.8). Therefore the variance (9.14) will be less than that of (9.8).

Equating (9.14) and (9.7) we obtain

$$\begin{aligned} p &= 2c(1 - \rho^2)(\sigma_b^2 + \sigma_w^2)/\sigma_w^2 \\ &= 2c(1 - \rho^2)/(1 - \rho) \\ &= 2c(1 + \rho). \end{aligned} \tag{9.15}$$

Substituting (9.15) in (9.1) we obtain the condition that the total investigation time is less for the cross-over as

$$2(1 + \rho) > 1 + (t + w)/(t + v), \tag{9.16}$$

and substituting in (9.3) the condition that the total trial time is less for the cross-over is

$$2c(1 + \rho) > cr_2/r_1 + (t + w)/(2r_1). \tag{9.17}$$

The reader may check for himself that, if analysis of covariance were used for a parallel in the case of Example 9.2, it would only be necessary to recruit 90 patients per group to have the same precision as from a cross-over with 30 patients and provided that the recruitment interval was not more than three days the advantage would lie with the parallel.

9.2.3 Discussion

The models we have used have been simplifications. For example we have assumed that the total variance for observations is a sum of between- and within-patient terms. Under this representation it is impossible for the cross-over to be at a disadvantage compared to the parallel in terms of the variance of treatment estimates given a fixed number of patients. It is both theoretically and practically possible, however, for a relatively homogenous set of patients at trial entry to be subject to quite different trends. Some may improve with time and others may deteriorate. Under such circumstances an additional source of variation would be introduced in the cross-over compared to the parallel. (Lehmacher, 1987, treats the cross-over in a general multivariate framework which permits such more general error structures.) Of course, if we had baseline measurements available before each treatment period, then as we showed in Section 3.15 it is possible in a cross-over trial to remove this further source of variation using analysis of covariance.

We have also (implicitly) assumed in discussing analysis of covariance for the parallel that the correlation between baseline readings and outcomes will be the same as that between first and second treatment values. This also might not be the case. For example some patients may be placebo responders and others not. This sort of variation is removed by comparing values obtained under two treatments but not by comparing with a baseline.

Another problem with analysis of covariance, which tends to diminish its efficiency, is that the variance terms have to be estimated. They are not given as we implicitly assumed. Since in practice the baselines tend not to be perfectly balanced between groups the variances of treatment estimates are, in practice, due to this lack of orthogonality, larger than suggested using models which assume that all other parameters are known.

On the other hand, there is no need to limit covariates in an analysis of covariance for a parallel only to a single baseline. Further baseline measurements, or measurements on other variables such as sex, height, weight etc. may also help to reduce variability. Obviously the total elimination of between-patient variation, however, represents an absolute limit which cannot in practice be reached.

For these and other obvious reasons, therefore, the formulae which have been presented, should be treated with caution. In practice, when deciding what sort of a trial to run the general considerations covered in Sections 1.3 to 1.5 will be just as important as, if not more important than, the sort of calculations suggested by the formulae above.

9.3 CARRY-OVER AND THE LENGTH OF THE WASH-OUT PERIOD

Ultimately, the length of wash-out period required is determined by pharmacokinetics and pharmacological response. These, and other associated terms, are defined in Chapter 10, where various issues regarding carry-over are investigated using simple models. Here we limit ourselves to giving some basic advice without either formal justification or explanation of technicalities. The following points should be noted.

- Ethical considerations may make it impossible to have a wash-out period. If so it may be necessary to use a parallel group trial rather than a cross-over.
- Early pharmacokinetic studies should be examined for the clues they can give regarding carry-over. For such studies it is often possible to determine absolutely how much of the active substance persists at given times after administration. Such information may be used to establish the minimum wash-out period necessary.
- For bioequivalence the FDA asks for a wash-out period which is at least three half-lives (Dighe and Adams, 1991). The sampling period itself, however, should also cover at least three half-lives. This implies (if drug elimination follows a half-life law) that by the time the second treatment period starts, the serum concentration should be not more than 1/64 what it was at the beginning of the first period.
- For single-dose studies there will usually be no difficulty in arranging for wash-out periods which are several times as long as the study periods. There is thus little excuse for not designing a trial in that way. As a simple rule of thumb I suggest that the wash-out period should be at least four times as long as the presumed measurable duration of action of a single dose of the treatment.
- Where multiple dose studies are being run a judgement will in any case be made concerning the length of time it will take for a given treatment to reach steady state. (This point is frequently overlooked.) When a number of treatments are being compared and it is desired to study the onset of action of such treatments, and not just the steady state, then the wash-out period for the trial should be no less than the longest presumed time to reach steady state for any treatment.
- When the onset is not being studied, an active wash-out period may be used instead. Measurements should then be limited to that part of the treatment period by which time steady state will have been reached by the slowest acting treatment.
- In designing a drug development programme every opportunity should be taken (within reason) to vary designs by using different sequences,

comparators, and wash-out periods as well as to supplement a programme of cross-over trials with some parallel group trials.

9.4 CHOOSING SEQUENCES FOR A CROSS-OVER

There is an extensive literature on 'optimal' choice of sequences for measuring treatment effects in the presence of carry-over. As I have already stated, I do not find the carry-over model usually employed to be credible. In my opinion, this literature, although possessed of a considerable theoretical fascination, is of no practical use. This point is taken up in Chapter 10. A case in which a choice of sequences can be crucial is incomplete block designs, a topic we considered briefly in Chapter 7. Although we shall consider a practical example below, there is not space in a book like to this to cover the general theory of incomplete blocks. It is covered in a number of texts on experimental design, for example, Cox (1958), John and Quenouille (1977), Mead (1988), Atkinson and Donev (1992) and Cox and Reid (2000). Yet another case where the choice of sequence may be important was covered in the non-parametric analysis of Chapter 6. There we pointed out that certain types of Latin squares were particularly useful. We shall provide a further example, in due course, where choice of sequences was important despite the fact that complete blocks were used and carry-over was assumed absent, but for the moment we divert to consider an example of a design involving incomplete blocks.

Example 9.3 A trial was designed with the object of obtaining evidence that a new formulation of formoterol, a multi-dose dry-powder inhaler, was equivalent to a standard formulation, a single-dose dry-powder inhaler (Senn *et al.*, 1997). Success in this respect would obviate the need to undertake a full development. The purpose of such studies is identical to that of a bioequivalence study, with the added difficulty that it is neither possible nor relevant to measure the concentration of the drug in the blood since it is inhaled and may act locally in the lungs. For similar reasons to those given in connection with Example 7.2, it was felt necessary to undertake a parallel assay in order to achieve the aim of claimed equivalence. The efficiency of such parallel assays depends on the strength of the dose response and it was felt desirable to study at least a quadrupling from 6 μg to 24 μg for each formulation. However, the marketed dose was 12 μg, and it was felt necessary to include these doses as well. Finally, it was also decided to include a placebo. There were thus seven treatments to be given, 6 μg, 12 μg and 24 μg for both test (T) and reference (R) formulations as well as placebo. In discussions with the local investigators who were to run the trial it was discovered that they considered it was unreasonable to treat patients for more than five consecutive periods.

Remark At this point, many standard texts would consider that the work of the statistician would begin. That is to say, the task would be to accept the 'design brief' of seven treatments and five periods and produce a good design. In fact the statisticians involved in planning the trial were also actively involved in constructing the brief, being involved as they were in discussions which led to determining the number of treatments to be compared as well as whether an incomplete blocks design, a full cross-over or a parallel group design should be used. For example, designs considered included a four-treatment cross-over ($6\,\mu$g and $24\,\mu$g for both formulations) and a five-treatment cross-over (with placebo as well). A detailed account of these deliberations will be found in Senn *et al.* (1997), which also includes all the data and an analysis of the trial.

Example 9.3 (continued) It was generally agreed that very high precision was required for this trial. However, an important issue was whether it was necessary to measure all contrasts with equal precision. Note that in order to balance the design for the effect of periods (to make it 'uniform' on the periods, to use design jargon) it is necessary to have multiples of seven sequences. This provides $(7 \times 5) = 35$ episodes per set of such sequences. Since this is divisible by 7 and 5, the seven treatments can be distributed evenly over the five periods. However, there is another sort of balance associated with incomplete block designs, and that is to do with the frequency with which given pairs of treatments are represented within the blocks. The fact that seven treatments are being compared means that $(7 \times 6)/2 = 21$ pairwise contrasts can be estimated. However, each patient, being treated with five treatments, would only provide the means of comparing $(5 \times 4)/2 = 10$ pairs of treatment. To have each pair of treatments represented equally requires 21 sequences since $21 \times 10 = 210$ is the lowest number divisible by both 21 and 10.

Remark A four-period design with seven treatments can be balanced much more easily. This is because each patient provides the means of making $(4 \times 3)/2 = 6$ pairwise comparisons. If seven sequences are used this provides $6 \times 7 = 42$ pairwise within-patient comparisons in total. However, this is divisible by 21 and so each treatment contrasts can be measured in two patients for any set of (suitably chosen) seven sequences. Note, however, that this is *not* a reason for preferring four-period to five-period designs, since, for the latter, even if it is considered impractical to use more than seven sequences, these seven will provide $7 \times 10 = 70$ pairwise within-patient comparisons and this is an average of $70/21 = 10/3 = 3\frac{1}{3}$. In fact things can be arranged so that seven pairs would appear in four sequences and 14 pairs would appear in three. Thus each contrast would be estimated with superior precision to that in the four-period design.

Example 9.3 (concluded) As it turned out, however, it was impossible to agree which contrasts should receive greater emphasis. In the end it was decided to use 21 sequences as given in Table 9.3. This basic design was replicated six times so that the planned number of patients studied was 126.

In Table 9.3 *T* stands for test, *R* for reference, *P* for placebo and the numbers give the dose in micrograms. There is a very strong pattern to these sequences. They are, in fact, based on three 7 × 7 Latin squares, the last two periods being ignored. The first square cycles the randomly chosen sequence *T*6, *P*, *R*24, *T*24, *R*6, *T*12, *R*12, moving each treatment on one period for each subsequent sequence. The second square is based on a sequence constructed from the generating sequence for the first square by placing every second treatment together (thus *T*6, *R*24, *R*6, *R*12, *P*, *T*24, *T*12) and then proceeding as with the first square. The third square takes the first generating sequence and places every third treatment together and then proceeds as before.

Remark This rather *ad hoc* procedure is not ideal in that it imposes a degree of structure on the final design beyond that dictated by the requirements of balancing both in the incomplete blocks sense and on the periods and using (for practical reasons) a minimal set of sequences. My excuse is that I was acting under considerable time pressure when I originally produced this design. Approaches using computer programs are considered later in this chapter.

Table 9.3 (Example 9.3) Sequences for an incomplete blocks cross-over.

			Period		
Sequence	1	2	3	4	5
1	T6	P	R24	T24	R6
2	R12	T6	P	R24	T24
3	T12	R12	T6	P	R24
4	R6	T12	R12	T6	P
5	T24	R6	T12	R12	T6
6	R24	T24	R6	T12	R12
7	P	R24	T24	R6	T12
8	T6	R24	R6	R12	P
9	T12	T6	R24	R6	R12
10	T24	T12	T6	R24	R6
11	P	T24	T12	T6	R24
12	R12	P	T24	T12	T6
13	R6	R12	P	T24	T12
14	R24	R6	R12	P	T24
15	T6	T24	R12	R24	T12
16	R6	T6	T24	R12	R24
17	P	R6	T6	T24	R12
18	T12	P	R6	T6	T24
19	R24	T12	P	R6	T6
20	R12	R24	T12	P	R6
21	T24	R12	R24	T12	P

We now consider a second example of a problem involving choice of sequences. This is rather unusual and is included to illustrate the variety of problems that can arise in practice when designing cross-over trials in clinical research.

Example 9.4 A placebo-controlled trial of formoterol in the prevention of late-phase allergic reactions was designed. Patients were studied on four treatment days. On each day they either received formoterol (*F*) or placebo (*P*) and were either given an active challenge (*A*) with an allergen to which they were known to respond or a dummy challenge (*D*) with saline solution. Each patient received on a separate day each of the four possible combinations of the two factors, treatment and challenge, namely: *FA*, *FD*, *PA*, *PD*. The purpose of the factorial structure was to see whether formoterol had any protective influence in allergic asthma which was not ascribable to its bronchodilating effect. The problem was to choose a suitable set of sequences for a trial. The following considerations applied.

- Although it would not be possible to blind the investigator with respect to challenge (*A* or *D*) he should be blinded regarding treatment.
- The most important consideration regarding wash-out was to keep the time between allergen challenges as long as possible due to a possible hypersensitizing effect. (Thus allergen in itself is held not to represent a problem regarding carry-over but to be subject to a form of latent carry-over which could be activated by a further challenge.)
- The design should be balanced for period.
- It should be completed in a reasonable time.
- It should be as simple as possible using the fewest sequences given the considerations above.

Before revealing the design chosen, it should be pointed out that some standard designs will not be suitable. Consider, for example, the Williams square

$$\begin{array}{cccc} FA & FD & PA & PD \\ FD & PD & FA & PA \\ PA & FA & PD & FD \\ PD & PA & FD & FA. \end{array}$$

This has at least two unsuitable features. It uses four challenge sequences, *ADAD*, *DDAA*, *AADD*, and *DADA*, and four treatment sequences, *FFPP*, *FPFP*, *PFPF* and *PPFF*. Once a challenge sequence is known, so is the treatment sequence, thus the investigator cannot be kept blind. Furthermore the sequences *DDAA* and *AADD* require a longer time between treatments because two allergen challenges are given consecutively.

The design eventually chosen used eight sequences formed as the product of the two challenge sequences, *ADAD* and *DADA*, and four treatment sequences *FPPF*, *PFFP*, *FFPP* and *PPFF*. The sequences were thus

FA PD PA FD
PA FD FA PD
FA FD PA PD
PA PD FA FD
FD PA PD FA
PD FA FD PA
FD FA PD PA
PD PA FD FA.

Here, knowledge of the challenge sequence provides no clue as to the treatment sequence and active challenges are never given consecutively. Thus, if two months were requested between allergen challenges, the patients could still come at monthly intervals.

9.4.1 Using SAS® to find designs

We can use *proc factex* of SAS® to construct designs similar to that of Example 9.3. Some suitable code is given in Table 9.4. First we must establish, using arguments similar to those above, how many Latin squares are needed. In this case the answer is three. We now treat the design as if it were a fractional factorial design with five factors: sequence, period and three factors corresponding to treatment. Each of these factors has seven levels. The purpose of *proc factex* is to produce fractional factorial designs, and by specifying that we want 49 units, by naming the five factors and by stating that they have seven levels (see code) we construct a design that effectively consists of three so-called orthogonal Latin squares. (Or equivalently one hyper-Graeco-Latin square.) That is to say, we construct three squares with the property that when overlaid on each other each letter of one square appears in conjunction with each letter of the other, including itself.

Next we need to drop periods 6 and 7, and then we need to abandon the fiction that the three ‘treatments’ are different. Each of *TREAT1*, *TREAT2* and *TREAT3* is a repetition of the same factor *TREAT*, so we join them together. However, each corresponds to a different set of sequences, so we create the full set. Finally, we note that the same periods are involved. Of course, if the object is simply to find a single Latin square for a complete blocks design, *proc factex* can also be used and the code is considerably simplified. Alternative approaches using GenStat® are discussed in Appendix 9.1.

Table 9.4 Construction of an incomplete blocks design of the sort considered in *Example 9.3*

```
proc factex;
  factors SEQ PERIOD TREAT1 TREAT2 TREAT3/ nlev=7;
  size design=49;
  model resolution=3;
  output out=DESIGN1 randomize (2001)
  TREAT1 cvals=(‘A’ ‘B’ ‘C’ ‘D’ ‘E’ ‘F’ ‘G’)
  TREAT2 cvals=(‘A’ ‘B’ ‘C’ ‘D’ ‘E’ ‘F’ ‘G’)
  TREAT3 cvals=(‘A’ ‘B’ ‘C’ ‘D’ ‘E’ ‘F’ ‘G’)
  SEQ nvals=(1 2 3 4 5 6 7)
  PERIOD nvals=(1 2 3 4 5 6 7);
run;

*Drop periods 6 and 7, join treatments together, repeat periods and
renumber sequences;
data DESIGN2 (keep=period sequence treat);
  set DESIGN1;
  if PERIOD < 6;
  SEQUENCE=SEQ;
  TREAT=TREAT1;
  output;
  SEQUENCE=SEQ+7;
  TREAT=TREAT2;
  output;
  SEQUENCE=SEQ+14;
  TREAT=TREAT3;
  output;
run;
*Dummy is used in the procedure below since at least one numeric;
*Variable is required when using the across option;
proc report data=design2 split=‘|’ nowd;
  column sequence period, (treat sequence=dummy);
    define sequence / group ‘Sequence||’;
    define period / across ‘Period’;
    define treat / width=1";
    define dummy / sum noprint;
run;
```

9.5 SAMPLE-SIZE ISSUES

We now discuss how the investigator who has chosen to carry out a cross-over trial may determine how large his trial should be. Before considering the technical aspects of the problem we look at some general issues.

In many clinical trials the sample size is a paramount ethical issue. If, for example, the patients are suffering from a fatal disease, then it may be ethically unacceptable to continue to treat patients with a treatment which is known to be inferior. It is thus highly desirable to come to a conclusion as to which treatment is superior whilst studying the minimum number of patients. A concept which has been stressed for such trials is the *power* of finding a clinically relevant difference—that is to say, the probability of concluding that the treatments are not equal, given that the true difference in effect between treatments is the least amount considered to be of any practical clinical importance and using a hypothesis test with a chosen level of significance.

Cross-over trials, of course, are not a suitable medium for studying such serious diseases. In a cross-over with complete blocks the patients will in any case have the chance to try all the treatments. Once a drug-development programme is complete future patients may be enrolled on to n of 1 trials. These are all grounds for believing that there is no overriding ethical reason for using the minimum number of patients possible in a cross-over trial. In my opinion, therefore, it is frequently more relevant to design a trial in terms of its ability to measure effects to a given precision rather than in terms of power. Accordingly, we shall consider this issue first.

For both approaches, except where we explicitly state otherwise, we shall assume in the discussion which follows that we are dealing with continuous Normally distributed measurements obtained from an AB/BA cross-over with n patients in total, where n is even, and with $n/2$ patients per sequence, and that we shall use the *CROS* analysis of Section 3.6.

9.5.1 Sample size and precision

We define σ^2 as the variance of a basic estimator for a patient. If the model of Section 9.2 applies then, since a basic estimator is the difference between two observations on a single patient, it follows that $\sigma^2 = 2\sigma_w^2$, where σ_w^2 is the within-patient variance. This relationship is an important one to note since in practice the application of any sample size formula for a cross-over trial depends on using a previously obtained estimate of variance. *It is thus crucial to establish how this estimate of variance was calculated.* Is it an estimate of the variance of a basic estimator or of the within-patient variance as defined in Section 9.2?

As we saw in Section 3.6, the variance of the *CROS* estimator is σ^2/n. (This may be expressed equivalently as $2\sigma_w^2/n$). This simple formula expresses the variability in our treatment estimate as a function of the number of patients in the trial and the variance of a basic estimator. This latter term, however, is unknown and has to be estimated and this introduces a further element of uncertainty. The importance of this latter factor depends on the degrees of freedom with which the variance is estimated. In the case of the *CROS* analysis there are $n - 2$ degrees of freedom for this purpose, one having been used up from each sequence group. There are a number of ways one could express the uncertainty associated with the t distribution. One common method is to compare the critical values associated with a 5% test two-sided to the values of $+1.96$ and -1.96 found from the Normal distribution, since these are the values to which the critical values from the t distribution tend as the number of degrees of freedom increases. A more general approach, less tied to any particular percentage points, is to quote the variance of the t distribution. In general, providing the degrees of freedom, v, exceed 2, then the variance of the t is $v/(v-2)$. Thus the variance of the t associated with the treatment estimate produced using a *CROS* analysis of an *AB*/*BA* cross-over is $(n-2)/(n-4)$. This may be compared with the variance of the standard Normal which is, of course, 1.

Figure 9.1 shows a plot of the variance of the treatment estimate in units of σ^2 (on the left-hand vertical axis) and of t (on the right-hand axis). It will be seen that, for small samples, the two decline rapidly as the sample size increases. As the sample size becomes 'large', however, the variance of t stabilizes. Thus, the influence of the t distribution soon becomes negligible and we see that for even a moderately large cross-over trial the main effect on precision is via the variance of the treatment estimate. One way of combining the two influences together is using Fisher's measure of precision (Fisher, 1990b, pp. 243–4) which in general for an estimator of τ with a variance estimate, $\hat{\sigma}_\tau^2$, based on v degrees of freedom is

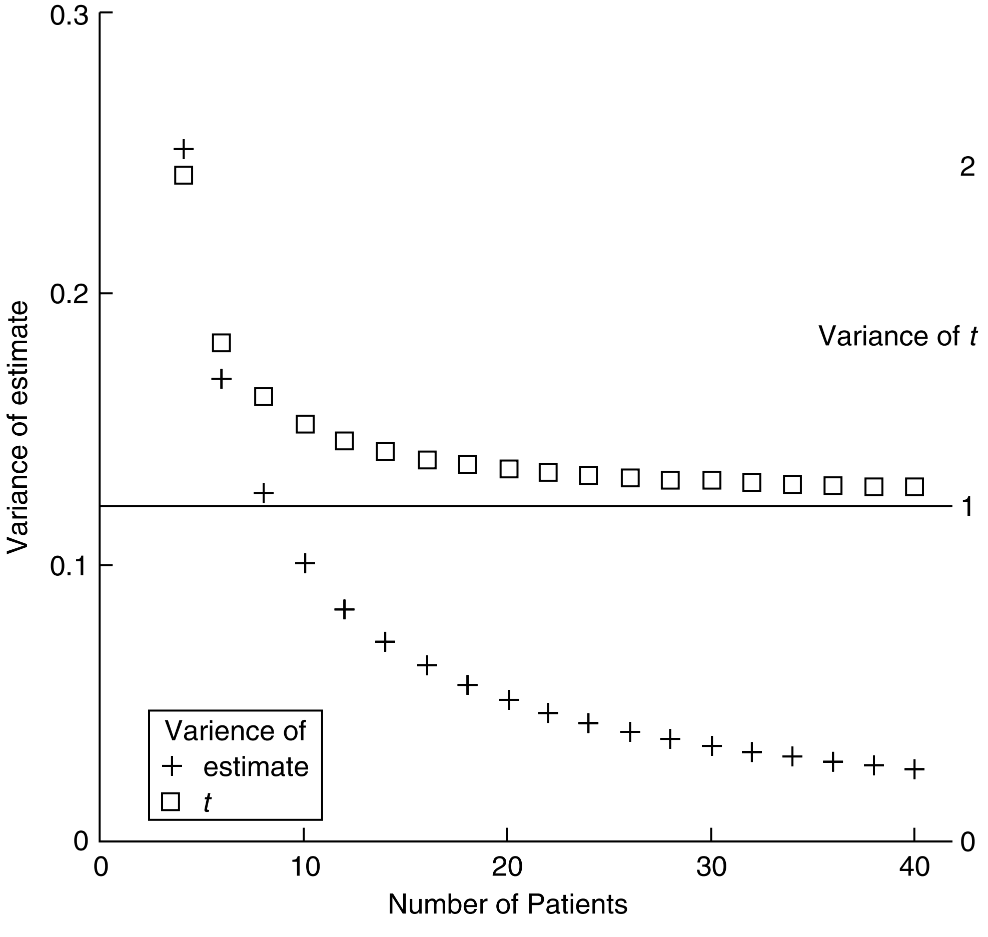


Figure 9.1 Variance of the treatment estimate and of the t distribution for an *AB*/*BA* cross-over as a function of the number of patients.

$$(\nu + 1)/\{\hat{\sigma}_\tau^2(\nu + 3)\}. \tag{9.18}$$

In the case we are considering we substitute $\hat{\sigma}^2/n$ for $\hat{\sigma}_\tau^2$ and $n - 2$ for ν, where $\hat{\sigma}^2$ is the unbiased estimate of $\hat{\sigma}^2$, to obtain

$$(n - 1)/\{(\hat{\sigma}^2/n)(n + 1)\} = n(n - 1)/\{\hat{\sigma}^2(n + 1)\}. \tag{9.19}$$

Obviously, as n increases, (9.19) tends to $n/\hat{\sigma}^2$, which is simply the inverse of the estimated variance of the treatment estimate. The expression (9.19) is not much use as it stands, however, since it depends on the units in which we measure. It is more usefully square-rooted and expressed in terms of the clinically relevant difference, Δ, thus:

$$\text{precision} = \Delta\{n(n - 1)\}^{1/2}/\{\hat{\sigma}(n + 1)^{1/2}\}. \tag{9.20}$$

Remark But for the factor $\{(n - 1)/(n + 1)\}^{1/2}$, (9.20) is simply the ratio of the supposed clinically relevant difference to the estimated standard error. I have never planned a cross-over with fewer than 12 patients (although I have come across designs with as few as 6). In such a case, the adjustment for degrees of freedom in (9.20) is a factor of 0.92 and this is unimportant compared to other uncertainties involved in the use of this formula.

There are two practical difficulties to using (9.20) for planning trials. The first is that the value of $\hat{\sigma}$ will not be determined until the experiment is run. The usual approach here is to substitute for $\hat{\sigma}$ a conservative estimate based on previous clinical trials. The second difficulty is that we need an external standard by which to judge precision in order to apply it in practice. Opinions will differ as to what constitutes a clinically relevant difference and what constitutes adequate precision. Obviously, it would not be possible to produce a table of clinically relevant differences for different diseases and measures! All I can say is that if the precision is defined by (9.20) I should *usually* regard 2 as being the minimum value of interest in clinical trials, 3 as being fair and 4 as being high.

Where cross-over trials other than the *AB*/*BA* are being considered, then my usual approach is to compare the precision with which a given contrast may be estimated compared to the precision of the corresponding estimate from an *AB*/*BA* cross-over. For a balanced design in complete blocks, where every patient receives each of k treatments and the total number of patients n is some integer multiple of k then the variance of a given contrast is σ^2/n, just as for the *AB*/*BA* cross-over.

We now consider an example of how these ideas may be put into practice.

Example 9.5 A single-dose, duration of action, dose finding cross-over is to be run with high precision comparing three doses of a new bronchodilator to a standard treatment and placebo. The outcome variable is FEV_1 12 hours after treatment and the minimum clinically relevant difference is deemed to be $0.25\,\ell$. On the basis of previous studies it is believed that the variance of a basic estimator is unlikely to be more than $0.09\,\ell^2$. How many patients should be enrolled?

Solution We have $\Delta = 0.25\,\ell$, $\hat{\sigma}^2 = 0.09\,\ell^2$, and we are looking for high precision, say 4. Working in terms of the square of (9.20) we require n such that $0.0625\,n(n-1)/\{0.09(n+1)\} = 16$. Hence $n(n-1)/(n+1) = 23$ from which $n^2 - 24n = 23$. Completing the square we have $(n-12)^2 - 144 = 23$, or $n - 12 = 12.9$. Hence $n = 25$. To run an *AB*/*BA* cross-over with the target precision, if we wished to have an even number of patients, we should then recruit 26 patients. For the cross-over in five periods 25 patients is just right.

Remark If ordinary least squares is used to analyse the five period-five treatment cross-over in 25 patients, we shall, in any case, have many more degrees of freedom available for error than the 23 for an *AB*/*BA* cross-over in as many patients. On the other hand if we used a basic estimator approach for analysis we should have three fewer. In practice, of course, all that it is necessary is to note that somewhere in the region of 25 patients are required and to choose an appropriate number for the trial. Another point is that the number makes no allowance for drop-outs.

9.5.2 Sample size and power*

(For further details regarding concepts and mathematics in this section the reader should consult the book by Desu and Raghavarao, 1990 or the tables of Machin and Campbell, 1987.)

A very common approach to sample-size determination is in terms of hypothesis testing. If we wish to test the null hypothesis that the difference between the treatments, as defined by the 'true' unknown treatment effect τ, is zero, against the alternative that it is not, then in the symbolism of hypothesis testing we test

$$H_0\colon \tau = 0 \text{ against } H_1\colon \tau \neq 0.$$

If $\hat{\tau}$ is an estimate of τ and $\hat{\sigma}_\tau$ is its estimated standard error based on v degrees of freedom, and $t_{\alpha/2,\,v}$ is a critical value of the t distribution with v degrees of freedom such that $P(t \geqslant t_{\alpha/2,\,v}) = \alpha/2$, then (given that certain standard assumptions apply) the decision rule,

$$\text{reject } H_0 \text{ unless } - t_{\alpha/2,\,v} < \hat{\tau}/\hat{\sigma}_\tau < t_{\alpha/2,\,v}, \tag{9.21}$$

defines a test of the null hypothesis of *size* α. That is to say, that if the null hypothesis is true, the probability of committing a Type I error and wrongly concluding that there is a difference between treatments is α.

An equivalent formulation of (9.21) is

$$\text{reject } H_0 \text{ unless} - \hat{\sigma}_\tau t_{\alpha/2,\, v} < \hat{\tau} < \hat{\sigma}_\tau t_{\alpha/2,\, v.} \tag{9.22}$$

Now suppose that the null hypothesis is incorrect and that the alternative hypothesis is true. If the estimate, $\hat{\tau}$, falls within the limits given by (9.22) we shall not reject H_0 despite the fact that it is false. This is called a *Type II error* and its probability, given that H_1 is true, is usually designated β. (Since, in fact, there is a range of values of τ which satisfies the alternative hypothesis there is a corresponding range of values of β.) Given any value of τ, δ, not equal to 0, we may calculate the value of β as

$$P(-\hat{\sigma}_\tau t_{\alpha/2,\, v} < \hat{\tau} < \hat{\sigma}_\tau t_{\alpha/2,\, v}) = P\left(-t_{\alpha/2,\, v} < \frac{\hat{\tau}/\sigma_\tau}{\hat{\sigma}_\tau/\sigma_\tau} < t_{\alpha/2,\, v}\right). \tag{9.23}$$

Now if we look at (9.23) we shall see that the divisor, $\hat{\sigma}_\tau/\sigma_\tau$, for the middle term in the inequality is the square root of a chi-square divided by its degrees of freedom, v, whereas the numerator $\hat{\tau}/\sigma_\tau$ has a Normal distribution with mean δ/σ_τ and variance 1. Since numerator and denominator are independent then, by Section 2.2.4, were δ equal to 0, the middle term in (9.23) would have a t distribution with v degrees of freedom and the associated probability would be α, as it should be. However, under H_1, $\delta \neq 0$ and so instead of having a t distribution the middle term has a *non-central t* distribution with degrees of freedom, v, and with non-centrality parameter, $\gamma = \delta/\sigma_\tau$. If we designated such a random variable by $T(\gamma, v)$, then we may write (9.23) as

$$P\{-t_{\alpha/2,\, v} < T(\gamma, v) < t_{\alpha/2,\, v}\}. \tag{9.24}$$

In practice we do not use (9.24) to calculate β but (assuming that we are interested in positive values of γ) we use

$$P\{T(\gamma, v) < t_{\alpha/2,\, v}\}. \tag{9.25}$$

There are two reasons. First, it will be found in practice that there is little difference between (9.24) and (9.25). The latter will be slightly larger but is simpler. Secondly, the difference between the two is the probability of correctly concluding that there is a difference between treatments precisely in those cases in which the treatment which is truly worse appears to perform better. Thus, although formally, in terms of the formulation of the hypothesis testing problem, we should not make a type II error, in practice we should nevertheless make a very serious error.

If, instead of focusing on the type II error we consider the probability of making a correct decision given that the alternative hypothesis is true then the relevant probability is $1 - \beta$. This is referred to as the *power* of the test. Thus we have the following expression for the power:

$$1 - \beta = P\left\{T(\gamma, v) > t_{\alpha/2, v}\right\}. \tag{9.26}$$

A very common approach to choosing the number of patients for a clinical trial is to design a trial with a given power to detect a given treatment difference given a particular size. In order to do this for a cross-over trial we go through the following steps.

(1) Decide upon the values of the size, α, and power, $1 - \beta$, for the trial.
(2) Establish a value of the target treatment difference, δ. Usually this will be the clinically relevant difference, Δ, but this, of course, is itself an ill-defined concept.
(3) Adopt a reasonable value of the variance σ^2 of a basic estimator to be used for the power calculation. This will usually be based on results obtained from earlier trials.
(4) Using the values α, $1 - \beta$, δ, σ and noting that $v = n - 2$, $\sigma_\tau = \sigma/n^{1/2}$ and hence $\gamma = n^{1/2}\delta/\sigma$, solve (9.26) for n and v.

This last step is not particularly straightforward but is facilitated by the fact that an approximate solution to (9.26) may be obtained using the Normal distribution and

$$(1 - \beta) \simeq P\{Z(n^{1/2}\delta/\sigma, 1) > z_{\alpha/2}\}. \tag{9.27}$$

where Z (a, b) is a random Normal variate with mean a and variance b, and z_c is a value of the Normal distribution such that $P\{Z(0, 1) > z_c\} = c$. We may rewrite (9.27) as

$$(1 - \beta) \simeq P\left\{Z(0, 1) > z_{\alpha/2} - n^{1/2}\delta/\sigma\right\}$$

from which

$$-z_\beta \simeq z_{\alpha/2} - n^{1/2}\delta/\sigma$$

and hence

$$n \simeq (z_{\alpha/2} + z_\beta)^2(\sigma/\delta)^2. \tag{9.28}$$

We now consider an example of how these formulae may be used.

Example 9.6 Suppose that we wish to run an *AB*/*BA* single-dose cross-over to compare two bronchodilators (as in Example 3.1) in order to detect a difference of 30 ℓ/min in peak expiratory flow with 80% power for a 5% size given that the standard deviation of a basic estimator is 45 ℓ/min. What size of trial do we require?

Solution We have $\alpha = 0.05$, $\beta = 0.2$, $\delta = 30\,\ell/\text{min}$. $\sigma = 45\,\ell/\text{min}$. From this we may obtain from tables of the Normal distribution $z_{\alpha/2} = 1.9600$, $z_\beta = 0.8416$. Hence $n \simeq (1.96 + 0.84)^2(45/30)^2 \simeq 17.64$. The next highest value of *n* is 18 and this is a multiple of 2 and thus suitable. We now calculate the critical value for a *t* test. The degrees of freedom are $18 - 2 = 16$. From tables of the *t* distribution we find $t_{16,\,0.025} = 2.120$. The noncentrality parameter is $18^{1/2}(30/45) = 2.83$. We now calculate the power of the test using the non-central *t* as $P(t_{2.83,\,16} > 2.120)$. Tables of this are not readily available, and in practice we should have to use a computer package such as SAS® or GenStat® or nQuery®, a package specifically devoted to sample size determination which includes a calculator for certain distributions including non-central *t*, chi-square and *F*. (We illustrate an implementation of the whole calculation on SAS® below.) The power of 0.76 is not high enough, so we try the next highest sample size of 20. For this we find that the power is 0.80. Hence 20 is the number of patients we require.

All of these calculations may be obtained using the SAS® code given in Table 9.5, which also gives the output from the program. Note the use of the control parameter *C*, which permits the user to specify how the standard deviation has been calculated—that is to say, whether it has been calculated as the standard deviation of differences between repeat measures or whether as the standard deviation of within-patient errors. The latter is appropriate if the root mean square error from a linear model analysis of a previous trial has been used. The former is appropriate if, for example, a matched-pairs *t* (as often encountered in the medical literature) is the source. (See remarks in Section 9.5.1).

An alternative approach is to use tables such as those of Machin *et al.* (1997). For example, their Table 4.2 may be used to determine sample size for a matched-pairs *t* test.

But for the fact that their table assumes that the degrees of freedom will be one fewer than the number of patients (a point which in practice makes no difference of any importance) and the fact that we have required an even number of patients for our development, the results should be the same. Their tables are entered using the ratio of effect to standard deviation. For Example 9.6 this ratio is 0.67 and the nearest tabulated ratio of 0.65 yields a sample size of 21 (Machin *et al.* 1997, p. 97). A similar use may be made of nQuery®, which, like the Machin *et al.* (1997) tables, assumes a matched-pairs *t* is being used.

Table 9.5 (Example 9.6) SAS Program for calculating the sample size for an *AB*/*BA* cross-over.

```
CODE

data one;
  label DF=‘degrees of freedom’
         N=‘number of patients’
        T1=‘critical value’ ;
*Set control parameter C
*C=0 if SIGMA = standard deviation of differences repeat measures
C=1 if SIGMA = standard deviation of within patient errors;
C=0;
*Set parameter values;
EFFECT=30;
SIGMA=45;
ALPHA=0.05;
BETA=0.20;

*Approximate sample size;
SIGMA=SIGMA*(sqrt(2))**C;
Z1=probit(1 - ALPHA/2);
Z2=probit(1 - BETA);
N=((Z1+Z2)*(SIGMA/EFFECT))**2;
N=2*(int(N/2+1));
*Calculate the power of this and the next 4 sample sizes using the t
distribution;
do I=1 to 5;
   DF=N-2;
   T1=tinv(1-ALPHA/2,DF);
   NONCENT=EFFECT/(SIGMA/sqrt(N));
   POWER=1-probt(T1,DF,NONCENT);
   output;
   N=N+2;
end;
run;
proc print label;
  var ALPHA BETA EFFECT SIGMA N DF T1 POWER;
run;
```

```
OUTPUT
                                       number   degrees
                                         of        of    critical
OBS  ALPHA  BETA  EFFECT  SIGMA  patients  freedom   value     POWER
 1    0.05   0.2    30      45       18        16     2.11991  0.75665
 2    0.05   0.2    30      45       20        18     2.10092  0.80491
 3    0.05   0.2    30      45       22        20     2.08596  0.84466
 4    0.05   0.2    30      45       24        22     2.07387  0.87708
 5    0.05   0.2    30      45       26        24     2.06390  0.90329
```

9.5.3 Sample-size—a discussion

There is a considerable disparity between the mathematical sophistication involved in a sample-size determination and the practical basis from which it must proceed, namely the choice of a 'clinically relevant difference' (or some other treatment difference) and a 'guesstimate' of the variability in the experiment to be performed. Obviously the issue of a determination of a sample size is not without its practical importance but my opinion is that it is a topic which is often given too much stress in design and in particular in interpretation of experiments. (I do not approve of power calculations, retrospective or otherwise, as an aid to interpreting non-significant results, for example. The investigator would be better off looking at the likelihood under null and alternative hypotheses . . . but these are topics beyond the scope of this book.) In many cases it is a more sensible approach to sample-size determination to have a look at what sort of trial has been run in the past in a particular area and see what sort of inferences were possible rather than going through some complicated power calculation: often this is no more than a ritual.

Two further issues need mentioning. One is to note that we have only covered sample-size determination for Normal data. We do not have space to consider other types of outcome. A good discussion of sample-size determination in general for cross-over trials, including binary outcomes, will be found in Hills and Armitage (1979). Ezzet and Whitehead (1992) also consider sample size determination for binary outcomes.

A practical issue which affects all sample-size determination is the problem of drop-outs, which we consider in the next section.

9.6 DROP-OUTS, PROTOCOL VIOLATIONS AND MISSING OBSERVATIONS

In practice clinical trials are rarely completed perfectly according to plan. Accidents in the conduct of trials may mean that some measurements go unrecorded. Patients may be included although they do not satisfy all inclusion criteria or may decide to withdraw from the trial before they have completed the full course of treatment. We include a discussion of this topic here because it is a sensible precaution to have thought about this issue when designing the protocol. The strategy to be adopted is thus part of the contingency planning of a trial.

I wish I had some wise advice I could offer the reader as to how to deal with such problems. As far as I am aware, however, there are no perfect solutions and all I can offer are some simple rules of conduct as follows.

- It is a good idea, as far as is practically possible, to reach precise agreement with all parties (e.g. investigator, clinical research associate, trial monitor, statistician) before initiating the trial concerning the rules for handling such eventualities. These should be documented in the trial protocol.

- All such departures from the planned course of the trial should be clearly noted.
- If any missing values are replaced by imputed values for the purpose of analysis (see below) in any tables where these imputed values appear they should be highlighted in some way (e.g. marked with an asterisk).
- In addition to performing the analysis as pre-specified in the protocol if patients drop out or if observations are missing it may be necessary to perform other sorts of analyses as well as to check the robustness of the conclusions obtained under the preferred analysis.

These, I think, are obvious and uncontroversial rules of conduct upon which most statisticians would agree.

The following are rules which I usually observe regarding the planning and analysis of trials.

- I usually plan a trial with an intended number of patients which allows for some discontinuations. I now usually prefer this to the alternative strategy of organizing replacements if and when a discontinuation is noted. Although this latter strategy has the advantage that a replacement with the same treatment sequence can be chosen, this in its turn causes difficulties for blinding and organization of the trial.
- Even though the operation of the 'intention to treat' philosophy (Pocock, 1983) is less obvious in the case of cross-over trials (Matthews, 1990a) I always state in the statistical methods section of the protocol that all patients will be analysed regardless of whether or not they satisfy the inclusion criteria of the trial. The reasons for doing this are (a) it is unethical to put patients to the inconvenience of a clinical trial unless we intend to use their data, (b) it is bad practice to operate quality rectification schemes, (c) usually the analysis incorporating all patients is the least ambiguous and (d) in practice I have usually found that such an analysis produces the lowest standard errors.
- As far as is possible I go for the maximum incorporation of data into the analysis. If a patient drops out I always use data from the periods which he did complete.

In practice, of course, if a fixed effect model is used, then a patient cannot contribute information regarding treatment unless he received at least two treatments and in any case he can hardly contribute much information except indirectly (in an analogous manner to incomplete blocks) regarding any given treatment contrast unless he received both treatments represented in the contrast. There is no practical difficulty, however, in incorporating such patients into an analysis using *proc glm* of SAS®. The analysis is carried out in the ordinary way and the program will extract what information is to be extracted under the fixed effects models using ordinary least squares. If a random effects model is used then some further information is recoverable, as already discussed in Chapters 3 and 7. Usually the amount of information recoverable is extremely small. In the context of drug development, however, it will be important to specify beforehand which approach (fixed or random) will be used. Further treatment of missing values in cross-over trials will be found in Jones and Kenward (1989, pp. 76–80), Patel (1986) and Matthews (1990a).

It may happen, however, that the patient received all treatments but that some observations are missing. This is quite common where repeated measures are being taken during the day. Indeed, it was a feature of Example 3.1, where the patients (who were children) travelled home after 8 hours but their parents were meant to continue taking measurements at home. In some cases some of the measurements between 10 and 12 hours are missing for these patients. Where we have situations of this sort, sometimes the missing values are estimated. Such a process is called *imputation* and is one of a number of possible approaches to dealing with missing observations (Little and Rubin, 1983). The following are some simple rules which I use. I make no great claims for their value.

- If an observation is missing in a series of measurements on a patient but observations before and after have been made I would usually estimate the missing observation by interpolation.
- In a single-dose trial, if a patient had to take rescue medication, I would usually carry forward the last reading before rescue medication was taken to be used for the rest of the observation period. Or I might consider using whichever was 'worse' of the last before rescue medication or the 'current' observation.
- If observations are prematurely discontinued during a treatment day, I would usually carry the last available measurement forward.

9.6.1 Missing data: in conclusion

Whatever approach one uses to dealing with missing data one is open to criticism. The fact that an observation is missing will in most cases be informative. For example a patient may discontinue in an *AB*/*BA* cross-over trial because of particular dissatisfaction with the treatment he was given. If this happens to be the first treatment he tried he will not try the other treatment, which might have proved satisfactory. Clearly this will bias the results.

There is no perfect solution to dealing with the problem. When data go missing information is lost. As Johnson (1989, p. 42) has put it, 'to carry out a sensible and informative analysis it is necessary to have a sufficient quantity of high-quality data. No amount of skilful adjustment or manipulation will compensate for more than a small amount of missing information.' 'Restoring' the information through imputation without adjusting the analysis will give the same analytic result as if the imputed observations were real. This is likely to give a misleading overestimate as to how much information is really to be had. On the other hand, the very fact that an observation is missing may reflect on the treatment. Simply analysing the observations which remain as if nothing has happened throws this potentially vital information away. There is only one rule for reporting experiments with missing data on which we can all agree: and that is full and frank disclosure as to what went missing where and why.

APPENDIX 9.1 PLANNING TRIALS WITH GENSTAT®

GenStat® provides extremely powerful facilities for planning experiments, although the emphasis is more on choosing basic designs than on choosing sample sizes. This reflects its origins in agricultural statistics where the experimental material, say a single field, might be fixed but a complex treatment structure had to be applied to the blocking structure. The concerns of the medical statistician are usually different. The planned structure of trials is often simple and the sample size is of primary concern. GenStat® does have a procedure, *REPLICATION*, for calculating the number of replicates of a design but this does not work directly in terms of target power. Similar code to that given for SAS for calculating sample size will be given below. For the moment we illustrate some other features of GenStat®.

For example, to find a Latin square the AGLATIN procedure can be used. The following is an example for a 5 × 5 square.

```
"To illustrate finding a Latin Square"
AGLATIN[ PRINT=design; ANALYSE=yes] 5; NSQUARES=1;\
TREATMENTFACTORS=!p(treat); ROWS=Patient;
COLUMNS=Periods; Seed=270901
```

This not only prints the design but also gives degrees of freedom for the ANOVA table associated with it. So the output is:

```
              * * * Treatment combinations on each unit of the design * * *


 Periods    1    2    3    4    5
Patient
       1    4    3    5    2    1
       2    1    5    2    4    3
       3    3    2    4    1    5
       4    2    1    3    5    4
       5    5    4    1    3    2
```

```
***** Analysis of variance *****
Source of variation              d.f.
Patient stratum                     4
Periods stratum                     4
Patient.Periods stratum
treat                               4
Residual                           12

Total                              24
```

Here the levels of the treatment factors are indicated by numbers. The ANOVA output is a useful reminder that the treatment effects are confounded with patient by period interactions which, of course, we assume to be random. For the random seed in this program I simply used the date on which it was generated: 27 September 2001 = 270901. (Not necessarily a good practice!)

To produce a design similar to that of Example 9.3, we could ask AGLATIN to produce three orthogonal 7 × 7 Latin squares for us. We then simply drop the last two periods. Suitable code might be:

```
"To illustrate finding a balanced incomplete block design
with seven treatments, five periods and 21 patients"

"Find three 7 × 7 orthogonal Latin Squares"
AGLATIN[ PRINT=Design;ANALYSE=no] NROWS=7; NSQUARES=3;\
TREATMENTFACTORS=Treat; ROWS=Pat; COLUMNS=Period;\
Seed=270901

"Construct incomplete blocks design from the squares"
VARIATE[NVALUES=49] Pat1,Pat2,Pat3
CALCULATE Pat1=Pat+0
        & Pat2=Pat1+7
        & Pat3=Pat2+7
APPEND[NEWVECTOR=Pat4] OLDVECTOR=Pat1,Pat2,Pat3
&[NEWVECTOR=Period] OLDVECTOR=Period,Period,Period
&[NEWVECTOR=Treat2] OLDVECTOR=Treat[1] ,Treat[2] ,Treat[3]
FACTOR[nvalues=147;levels=21] Patient; Values=Pat4
FACTOR[nvalues=147; levels=7; LABELS=!T(A,B,C,D,E,F,G)]\
Treatment; VALUES=Treat2
SUBSET[CONDITION=Period.LT.6] Treatment, Patient, Period;
PRINT ‘Incomplete blocks design: columns are periods,
rows are patients’
PRINT[ORIENTATION=across; RLPRINT=*] Treatment;\
                                         FIELDWIDTH=15
```

The first section of this uses the procedure *AGLATIN* to find three orthogonal Latin squares. The second section appends the three Latin squares, drops the last two periods and replaces the numbers produced by *AGLATIN* by letters. The result is:

Incomplete blocks design: columns are periods, rows are patients

B	D	A	C	G
A	C	G	B	F
G	B	F	A	E
E	G	D	F	C
C	E	B	D	A
F	A	E	G	D
D	F	C	E	B
C	G	A	E	F
B	F	G	D	E
A	E	F	C	D
F	C	D	A	B
D	A	B	F	G
G	D	E	B	C
E	B	C	G	A
D	C	A	G	E
C	B	G	F	D
B	A	F	E	C
G	F	D	C	A
E	D	B	A	F
A	G	E	D	B
F	E	C	B	G

The code, analogous to that for SAS® included in this chapter, for sample size determination for the *AB*/*BA* design given chosen Type I and II error rates, a clinically relevant difference and a posited standard deviation is given below:

```
"To calculate sample sizes for a cross-over"
SCALAR Effect, Sigma, Alpha, Beta, Z1, Z2, Noncent, Power, C
    & DF,T1; EXTRA=‘Deg. Freedom’ , ‘Crit. Value’
"Set control parameter C"
"C=0 if Sigma = standard deviation of differences repeat
measures"
"C=1 if Sigma = standard deviation of within patient errors"
CALCULATE C=1
"Set parameter values"
CALCULATE Effect,Sigma, Alpha, Beta=30,45,0.05,0.20
"Approximate sample size"
```

```
CALCULATE Sigma=Sigma*SQRT(2)**C
        & Z1=EDNORMAL(1-Alpha/2)
        & Z2=EDNORMAL(1-Beta)
        & N=((Z1+Z2)*(Sigma/Effect))**2
        & N=2*(INT(N/2+1))
"Calculate the power of this and
the next 4 sample sizes using the t-distribution"
FOR[NTIMES=5]
  CALCULATE DF=N-2
  & T1=EDT(1-Alpha/2;DF)
  & Noncent=Effect/(Sigma/SQRT(N))
  & Power = CUT(T1;DF;Noncent)
PRINT[IP=Extra] Alpha, Beta, Effect, Sigma, N, DF, TI,\
  Power; FIELDWIDTH=5(7),2(15),7; DECIMALS=4(2),2(0),2(4)
  CALCULATE N=N+2
ENDFOR
```

EDNORMAL stands for 'equivalent deviate Normal' and is the inverse Normal distribution function. *EDT*, for 'equivalent deviate *T*', is the corresponding function for the *t* distribution. *CUT* stands for 'cumulative *t*' and is the cumulative upper probability for a non-central *t* distribution.

APPENDIX 9.2. PLANNING TRIALS WITH S-PLUS®

As far as I am aware, S-Plus® does not support functions which permit one easily to select sequences for Latin square and incomplete block designs. It is, however, a fairly simple matter to write code analogous to that already given for SAS® and GenStat® for the purpose of calculating sample sizes for a cross-over trial. Some code is given below:

```
#Program to calculate sample size for an AB/BA cross-over

#Set parameter values
EFFECT< −30 #Effect size
SIGMA < −45 #Standard deviation of difference between
                                              observations
ALPHA < −0.05 #Type I error rate
BETA < −0.20 #Type II error rate
#Set control parameter C
#C=0 if SIGMA = standard deviation of differences repeat
                                                  measures
#C=1 if SIGMA = standard deviation of within patient errors;
C<-0
```

```
#Approximate sample size
SIGMA<-SIGMA* (sqrt(2))^C
Z1<-qnorm(1-ALPHA/2)
Z2<-qnorm(1-BETA)
N1<-((Z1+Z2)* (SIGMA/EFFECT))^2
N1<-2* floor(N1/2+1)

#Declare variables
n<-5 #Number of sample sizes
N<-numeric(n)
DF<-numeric(n)
T1<-numeric(n)
NONCENT<-numeric(n)
POWER<-numeric(n)

#Calculate the power of this and the next 4 sample sizes using
                                            the F distribution
i< −1:n
    N[i] <-N1+2* (i-1)
    DF[i] <-N[ i] -2
    T1[i] <-qt(1-ALPHA/2,DF[i])
    NONCENT[i] <-EFFECT^2/(SIGMA^2/N[i])
    POWER[i] <-1-pf(T1[i] ^2,1,DF[i] ,NONCENT[i])
                              #Note use of non-central F
sample.frame<-data.frame(ALPHA,BETA,EFFECT,
                                 SIGMA,N,DF,T1,POWER)
sample.frame
```

Note that S-Plus® does not support a function for the non-central *t* distribution but it does have a function for the non-central *F* distribution and by using the fact that square of a *t* with ν degrees of freedom is an *F* with $1,\nu$ degrees of freedom (see Chapter 2) the calculation may proceed.

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

10

*Mathematical Approaches to Carry-over**

10.1 THE PURPOSE OF THIS CHAPTER

This chapter provides an introduction to some standard designs and analyses of cross-over trials. These reflect a very common philosophy regarding the modelling of carry-over. The field is very large and since the topic may be regarded as belonging to the more advanced theory of cross-over trials, there would in any case not be room in a book such as this to cover it in detail. However, as I have already explained, I do not consider that this particular modelling philosophy is useful (it attempts either too much or not enough) and there are thus other practical reasons for limiting discussion of this topic. Four things, therefore, will be attempted in this chapter.

First, the model usually employed for carry-over will be introduced. Second, I shall outline reasons for believing that it is not useful and indeed harmful. Third, for the benefit of those readers who are not persuaded by my arguments, the ways in which the model may be employed will be introduced. Finally, advice as to where further information regarding these methods may be obtained will be given.

References will be made at various points to theories of pharmacology. It is appropriate therefore to quote the following definitions (Martin, 1990, p. 529):

pharmacokinetics *n.* the handling of a drug within the body, which includes its absorption, distribution in the body, metabolism, and excretion.
pharmacodynamics *n.* the interaction of drugs with cells. It includes such factors as the binding of drugs to cells, their uptake, and intracellular metabolism.

In fact, in our usage we shall extend the scope of pharmacodynamics still further to include all physiological expression of pharmacological response so that, for example, improvement in lung function as a result of treatment with a drug will be regarded as belonging to the realm of pharmacodynamics.

Pharmacokinetics is sometimes described as being what the body does to the drug and pharmacodynamics as what the drug does to the body. The theory we shall rely on is covered in the book by Gibaldi and Perrier (1982) and Rowland and Towzer (1995). A particularly useful paper covering pharmacodynamics is that of Holford and Sheiner (1981). A good discussion of carry-over will be found in Sheiner *et al.* (1991).

In much of the discussion which follows it will be assumed implicitly that repeated dose cross-overs are being considered. Of course, single dose studies are particularly suitable to be designed as cross-over studies and, indeed, nearly all of the examples in this book are of this sort. In general, however, they present fewer problems regarding carry-over and it is therefore more profitable, when critically examining models, to look at the more difficult multiple dose case. An important concept in multiple-dose studies, and one we shall make frequent reference to, is that of the *steady state*. If repeated doses of a drug are administered it will usually be found that the total response increases with each administration but approaches some limit. This limit is the steady state.

We shall also concern ourselves exclusively with pharmacological carry-over. This is not, of course, the only form of carry-over possible. Discussions of other forms, for example psychological carry-over, will be found in Hills and Armitage (1979) and Millar (1983). If, however, the usual carry-over model contradicts standard pharmacokinetics and pharmacodynamics (as we shall show it does), there will be no reason for employing it, unless an argument can be produced to show that it is appropriate, *despite the fact* that it cannot apply to pharmacological carry-over.

We now proceed to our examination of modelling carry-over by considering an example.

10.2 AN ILLUSTRATIVE EXAMPLE

Suppose that we require to design a cross-over trial in two treatments, four periods and two sequences. An obvious thing to do is to allocate patients in equal numbers to pairs of ‘dual’ sequences in which each patient gets each treatment twice. Thus we would either use an *ABAB*/*BABA* design or an *ABBA*/*BAAB* design or an *AABB*/*BBAA* design. If we ignore the problem of carry-over there is nothing to choose between these three designs. If, however, we wish to adjust for simple carry-over then the second and third designs are more efficient than the first. Indeed, they are the most efficient designs in four periods and two sequences (Jones and Kenward, 1989). For this design, we are able to define eight cell means consisting of the mean response in each of the eight categories defined by the combination of the two sequences and four periods. If we ignore carry-over altogether, then a simple scheme of weights for combining the eight cell means to form an estimate of the difference between *A* and *B* is as given in

Table 10.1. An estimator produced using these weights will be referred to as an *unadjusted estimator*.

These weights have the familiar property that they add to zero over any sequence or any period, thus eliminating patient and period effects, and they add to 1 for *A* and − 1 for *B*, thus providing the necessary contrast. Obviously, they could have been expressed more simply as 1/4, 1/4, −1/4, −1/4 etc. The particular form has been chosen to facilitate comparison with an alternative scheme of weights given in Table 10.2.

The weights in Table 10.2 are those which are appropriate to an estimator which will be referred to as the *adjusted estimator*. (The derivation of such weights will be covered in Section 10.5 and need not concern us here.) If these weights are studied it will be seen that just as in Table 10.1 they sum to zero over sequences and periods and to 1 and −1 respectively for treatments *A* and *B* but in addition, if carry-over is present and *if it is determind entirely by the previous treatment and limited exactly to one period*, then its effect will be eliminated from the treatment estimate. In the literature on cross-over trials this particular form of carry-over is commonly taken to apply (at least to an order of approximation which is assumed to justify its adoption). I shall refer to it as *simple carry-over* in order to keep the terminology consistent with Senn (1992). (In many ways the term *mathematical carry-over* would seem more appropriate.)

Table 10.1 Scheme of weights for combining cell means for the estimate of the treatment effect for an *AABB*/*BBAA* cross-over when carry-over does not apply. (Weights for an unadjusted estimator.)

		Period		
Sequence	I	II	III	IV
AABB	5/20	5/20	− 5/20	− 5/20
BBAA	− 5/20	− 5/20	5/20	5/20

Table 10.2 Scheme of weights for combining cell means for the estimate of the treatment effect for an *AABB*/*BBAA* cross-over when ‘simple’ carry-over applies. (Weights for an adjusted estimator.)

		Period		
Sequence	I	II	III	I‘V
AABB	6/20	4/20 *a*	−7/20 *a*	−3/20 *b*
BBAA	−6/20	−4/20 *b*	7/20 *b*	3/20 *a*

In Table 10.2 the symbols a and b have been used to represent simple carry-over in the table and the weights associated with them add to zero. On the other hand, had we used the weights from Table 10.1 instead, and if simple carry-over were present, then the treatment estimate would be subject to a bias which may easily be seen to be $a/4 - b/4$. At first sight, therefore, there would seem to be no advantage (other than simplicity) in preferring the unadjusted to the adjusted estimator. If there is no carry-over each is unbiased. If there is *simple carry-over*, then the adjusted estimator is unbiased, whereas the unadjusted estimator is biased. If some other form of carry-over applies, then probably both estimators are biased.

The point is, however, that the variances of the estimators are not the same. What the exact variance is in each case depends on the error structure. (I note here in passing that in much of the exceedingly complicated investigations into 'optimal' designs for cross-overs this point is ignored.) On the most optimistic model for the error structure, namely that once between-patient effects have been eliminated the errors are uncorrelated with equal variance, then the rules for linear combinations given in Section 2.2.1 show that the variances of the estimators are proportional to the sum of the squares of the weights so that the variance of the unadjusted estimator equals $0.50\sigma^2/n$, whereas that of the adjusted estimator is $0.55\sigma^2/n$, where σ^2 is the within-patient variance and n is the number of patients per sequence. The variance of the adjusted estimate is thus 1.1 times that of the unadjusted estimate. (This result agrees with Jones and Kenward, 1989, p. 180.) It thus does not follow that the adjusted estimator is automatically superior to the unadjusted one. It will have a higher variance. It may or may not have a lower bias.

10.3 FIVE REASONS FOR BELIEVING THAT THE SIMPLE CARRY-OVER MODEL IS NOT USEFUL

In this section I hope to persuade the reader of the following. First, that in order to justify using the simple carry-over model the investigator either has to know or assume enough about carry-over to be able to design an AB/BA trial for which carry-over is not a problem. Second, that certain aspects of the simple carry-over model are implausible given elementary pharmacokinetic and pharmacodynamic theory. Third, that models which employ simple carry-over are self-contradicting. Fourth, that analyses based on such models are inefficient when compared with simpler methods. Fifth, that designs which are commonly claimed to be optimal on the basis of such models are not. In short, I intend to present a case that there is no place for such models in any practical approach to designing and analysing clinical trials. I shall now proceed to present in turn the five reasons for believing that the simple carry-over model is not useful.

10.3.1 If it applies then the investigator can design a trial which eliminates it

There is an extremely curious feature about the design presented in Section 10.2 above: contrary to appearance, it is really nothing more than an *AB*/*BA* design. All we need to do is measure halfway through each of the treatment periods in an *AB*/*BA* design to obtain four measurements on each patient and hence an *AABB*/*BBAA* design. Furthermore, if we look at the model for simple carry-over employed, we shall see that if it applies then the second of any pair of treatments is only affected by its own carry-over. Thus in the *AABB* sequence (as shown in Table 10.2) the period II value has treatment *A* and carry-over *a* and the period IV value has treatment *B* and carry-over *b*. We can represent this schematically by writing the treatment effects and associated carry-over for the sequences as *A* (*Aa*) (*Ba*) (*Bb*)/*B* (*Bb*) (*Ab*) (*Aa*). Now it is a fact that in multi-dose studies we are most usually interested in comparing the *steady-state effects* of two treatments, that is to say the effects eventually produced once the treatments have reached a steady state. Thus we should be more interested in comparing *Aa* and *Bb* than in comparing *A* and *B* since, whether or not *Aa* and *Bb* represent the steady state responses of the two treatments, they will more closely represent these responses than will *A* and *B* alone. But these responses are to be obtained from the period two and four values. Thus we could base an analysis on these values alone. But if we do that we are doing nothing more than analysing an *AB*/*BA* design using the standard *CROS* analysis which ignores carry-over.

It might be argued that this is unfair. In practice we would not create an *AABB*/*BBAA* design by dividing the periods from an *AB*/*BA* design in two but by doubling the number of periods. This argument provides no escape, however, for if we are able to design such a trial we are equally capable of designing an *AB*/*BA* design with periods which are twice as long.

It might be argued that, even if this is so, an analysis based on adjusted estimators from a four-period design will be more efficient in the sense of having lower variance than an analysis based on the two periods of an *AB*/*BA* design. There are a number of difficulties with this argument. If it really is the case that the variance of a design can be reduced at will simply by repeated measurement, why not, in that case, cut the treatment periods in two yet again to produce the *AAAABBBB*/*BBBBAAAA* design? If the simple carry-over model applies to this design also then we can represent the effects in the sequences as *A* (*Aa*) (*Aa*) (*Ba*) (*Bb*) (*Bb*) (*Bb*)/*etc*. and hence use periods II, III, IV, VI, VII and VIII with weights 1/6, 1/6, 1/6, −1/6, −1/6, −1/6 for the first sequence and the same weight with sign reversed for the second. This estimator then apparently has variance $\sigma^2/(3n)$, which is lower than that provided by the adjusted estimator for the *AABB*/*BBAA* design. In practice, of course, covariances between error terms mean that these sorts of reductions in variance are *not* achieved. If on the other hand we argue that the simple carry-over model does not apply to the

8-period design if the whole design is no longer than the four-period design (i.e. if it has been created by halving the periods) then we admit that we have used our judgement to decide for what period it does apply. Our analysis is not then, as if often claimed. free of assumptions about carry-over. On the contrary it makes important assumptions about it.

In short, the simple carry-over model implies that we know something about the persistence of carry-over. But this is precisely the knowledge which is required to design an *AB*/*BA* cross-over safely. For this reason I am firmly convinced that contrary to what has been claimed by many statisticians the *AB*/*BA* design is as reasonable, if not more reasonable, than any other.

In the field of bioequivalence empirical evidence has been provided for this position. D'Angelo *et al.* (2001) carried out an analysis of a large series of bioequivalence studies. In 324 two-period and 96 three-period designs they found an empirical distribution of the *P* values that was not significantly different from uniform in either case. (This being the distribution expected under the null hypothesis of no effect.) For the target measure AUC, 37 of the two-period designs had significant carry-over at the 10% level where 32.4 trials would have been expected and 4 of the three-period designs had significant carry-over at the 5% level, where 4.8 would have been expected under the null hypothesis of no carry-over in any trial.

10.3.2 It is implausible given elementary pharmacokinetic and pharmacodynamic theory

To make this point we shall need to consider some standard elementary pharmacokinetics and pharmacodynamics. We shall show, in fact, that on the basis of elementary pharmacokinetics alone one aspect of the simple carry-over model is *not* unreasonable. Pharmacological carry-over, that is to say carry-over of active ingredient, might well be mainly dependent on previous treatment (although it would really be dependent on total treatment history) and scarcely at all on the current treatment. However, the limitation of carry-over to one period by the simple carry-over model, as discussed above, is quite arbitrary. It can be defended on the basis of judgement but this judgement is of the same form as that which would be used to claim no carry-over for an *AB*/*BA* design. On the other hand, pharmacodynamic considerations suggest that the assumption that carry-over of effect is not affected by the current treatment is not reasonable.

Suppose we consider that drug disposition may be modelled by a one-compartment pharmacokinetic model and that the treatment is administered by intravenous injection. (The consequence of this is that we need to consider only the elimination of the drug and not its absorption. This assumption simplifies the argument but has no serious bearing on the conclusions.)

Assuming that the instantaneous rate of elimination of the drug is proportional to its concentration in the plasma, then under such circumstances the drug concentration X_t in the plasma at time t after administration is given by

$$X_t = X_0 \mathrm{e}^{-kt}, \tag{10.1}$$

where X_0 is the concentration immediately after injection and k is an elimination constant characteristic of the drug (Gibaldi and Perrier, 1982, pp. 2–3). If we now give repeated administrations of the drug at constant time intervals τ then the local *concentration maximum* after the Nth administration ($C_{\max_N}$) is

$$C_{\max_N} = X_0 \frac{1 - \mathrm{e}^{-Nk\tau}}{1 - \mathrm{e}^{-k\tau}}, \tag{10.2}$$

whereas the local *concentration minimum* just prior to the $(N + 1)$th administration (i.e. corresponding to a so-called *footpoint* measurement) is

$$C_{\min_N} = X_0 \frac{1 - \mathrm{e}^{-Nk\tau}}{1 - \mathrm{e}^{-k\tau}} \mathrm{e}^{-k\tau}$$

and hence

$$C_{\min_N} = C_{\max_N} \mathrm{e}^{-k\tau} \tag{10.3}$$

More generally, if we cease treatment with the Nth administration, then the plasma concentration s time units after the last administration is

$$C_{S_N} = X_0 \frac{1 - \mathrm{e}^{-Nk\tau}}{1 - \mathrm{e}^{-k\tau}} \mathrm{e}^{-ks}. \tag{10.4}$$

Alternatively (10.4) may be used for plasma concentration between the Nth and $(N + 1)$th administration, i.e. where $s < \tau$ (Gibaldi and Perrier, 1982, p. 116).

There are a number of points regarding treatment which may be noted with the help of these formulae.

- (10.5) As N increases the plasma concentrations for $C_{\max_N}$ and $C_{\min_N}$ approach asymptotes,

  $$C_{\max_\infty} = X_0/\left(1 - \mathrm{e}^{-k\tau}\right) \quad \text{and} \quad C_{\min_\infty} = \mathrm{e}^{-k\tau} X_0/\left(1 - \mathrm{e}^{-k\tau}\right)$$

  respectively. (Hence, $C_{\min_\infty} = C_{\max_\infty} \mathrm{e}^{-k\tau}$.)
- The level reached depends on the unit dose, X_0, the rate of elimination k, and the dosing interval τ.
- The more frequently the patient is dosed the less the proportionate difference between the $C_{\max}$ and $C_{\min}$ values.

- Whatever the number of doses given, the decay in plasma concentration is exponential at any time point.

We may note, in passing, that these points have several implications for the design of clinical trials, which are sometimes overlooked. For example, we may see that an effective treatment does not consist of a drug alone but of a drug combined with a treatment regimen, and that to study treatments successfully, both in terms of efficacy and tolerability, we have to choose trials with treatment periods of appropriate length. Such choices are made by exercising judgement based on the results of studies at earlier stages of drug development. If cross-over trials are employed a further judgement is required regarding carry-over but it is not qualitatively different from these.

From (10.3) and (10.5) we can see that if we design a placebo-controlled cross-over trial of an active treatment and we choose the length of a period to be $N\tau$, then the plasma concentration at the end of the first active treatment period is

$$C_{\min_N} = C_{\min_x}\left(1 - \mathrm{e}^{-Nk\tau}\right). \tag{10.6}$$

If we immediately follow it up by another period of active treatment of the same length as the first, then the plasma concentration at the end of the second period will be

$$C_{\min_{2N}} = C_{\min_\infty}\left(1 - \mathrm{e}^{-2Nk\tau}\right). \tag{10.7}$$

The difference between the two is the ‘carry-over’ and equals

$$C_{\min_\infty}\mathrm{e}^{-Nk\tau}\left(1 - \mathrm{e}^{-Nk\tau}\right). \tag{10.8}$$

If, however, we are interested in plasma concentration at time $(N + 1)\tau$ after the administration of the Nth dose, where this was the last dose, as would be the case if we followed the active treatment by a placebo period of equal length, we can substitute $(N + 1)\tau$ for s in (10.4) to get

$$C(N + 1)\tau_N = X_0\frac{1 - \mathrm{e}^{-Nk\tau}}{1 - \mathrm{e}^{-k\tau}}\mathrm{e}^{-k\tau}\mathrm{e}^{-Nk\tau}$$

which, by inspecting (10.5), may be seen to be just the same as (10.8). We thus see that in terms of elementary *pharmacokinetics* one aspect of the simple carry-over model was *reasonable*.

If, however, we looked at the model as it was used in connection with the *AABB/BBAA* design above, we shall see that we assumed implicitly that carry-over into placebo following two periods of active treatment was the same as carry-over into active treatment following one period of active treatment. The latter is given by (10.8) above but the former is given by applying (10.4) to (10.7) to give

$$C_{\min_\infty} \mathrm{e}^{-Nk\tau}\left(1 - \mathrm{e}^{-2Nk\tau}\right), \tag{10.9}$$

and is therefore not the same. We thus see that this aspect of the simple carry-over model is *unreasonable* on the basis of pharmacokinetics. The carry-over must depend on the *total previous history of administration*. Of course, if $\mathrm{e}^{-2Nk\tau} \simeq \mathrm{e}^{-Nk\tau}$ then (10.9) $\simeq$ (10.8), but this condition is satisfied if $\mathrm{e}^{-Nk\tau} \simeq 1$, from which two things follow. First, that by (10.2) and (10.5) the drug has reached steady state by the end of the first period and secondly, that by (10.8) there is no carry-over.

Consequently, we may draw the following conclusion on the basis of elementary pharmacokinetics. *If we have designed a cross-over which is adequate for the purpose of studying steady-state pharmacology we have designed a cross-over for which pharmacological carry-over is eliminated by the end of each subsequent treatment period.* If we wish to eliminate it by the beginning of the next treatment period a wash-out equal in length to the treatment period will suffice.

However, in most clinical trials we do not measure plasma concentration of active treatment (bio-availability and bio-equivalence studies are exceptions), we measure a response. We now consider, therefore, a simple pharmacodynamic model: a model of pharmacological response.

A model which has been used to express the relationship between pharmacological response and plasma concentration is given by the Hill equation (Hill, 1910; Gibaldi and Perrier, 1982, p. 222)

$$R = \frac{R_m C^\alpha}{C_{R50}^\alpha + C^\alpha}. \tag{10.10}$$

R is the pharmacological response, R_m is the maximum response possible, C (as before) is the plasma concentration, and α is a parameter relating response to concentration. As the concentration, C, increases, the response, R, approaches the limit R_m. The parameter, C_{R50}, is the concentration which produces 50% of the maximum response.

Consider two possible treatment sequences: active treatment followed by active treatment and active treatment followed by placebo. Suppose that the concentration at the end of the first active period is C_A and that the carry-over to the end of the following treatment period is D. If the second treatment is an active treatment the total plasma concentration will be $C_A + D$, whereas if it is a placebo the concentration will be D. From (10.10) the response under active treatment, R_A, in the absence of carry-over will be

$$R_A = \frac{R_m C_A^\alpha}{C_{R50}^\alpha + C_A^\alpha}. \tag{10.11}$$

With carry-over the response will be

$$R_{AA} = \frac{R_m (C_A + D)^\alpha}{C_{R50}^\alpha + (C_A + D)^\alpha}. \tag{10.12}$$

The carry-over of effect is given by the difference between the two:

$$CE_{AA} = \frac{R_m (C_A + D)^\alpha}{C_{R50}^\alpha + (C_A + D)^\alpha} - \frac{R_m C_A^\alpha}{C_{R50}^\alpha + C_A^\alpha}. \tag{10.13}$$

On the other hand, in the absence of carry-over, the response under placebo will be $R_P = 0$. Hence the response given that carry-over has occurred is identical to the carry-over of effect and we have

$$CE_{AP} = \frac{R_m D^\alpha}{C_{R50}^\alpha + D^\alpha}. \tag{10.14}$$

Hence, if $CE_{AA} = CE_{AP}$, as the simple carry-over model implies, then

$$\frac{R_m (C_A + D)^\alpha}{C_{R50}^\alpha + (C_A + D)^\alpha} = \frac{R_m C_A^\alpha}{C_{R50}^\alpha + C_A^\alpha} + \frac{R_m D^\alpha}{C_{R50}^\alpha + D^\alpha}. \tag{10.15}$$

Obviously, one trivial solution to (10.14) is if $D = 0$ but this is simply the case in which there is no carry-over. In general, however, (10.15) is not satisfied and (10.13) and (10.14) can yield very different values. For example if $D = C_{R50}$ but C_A^α is large in comparison to C_{R50}^α, then (10.13) $\simeq 0$ whereas (10.14) is $0.5R_m$, i.e. 50% of the maximum response.

As a more realistic example, take the simple case where $\alpha = 1$ and suppose that we have successfully designed a trial so that 75% of the maximum response may be obtained under active treatment. From (10.11) we have $C_A = 3C_{R50}$. Now suppose that 25% of the substance carries over so that $D = 0.25C_A = 0.75C_{R50}$. Substitution into (10.13) and (10.14) yields the widely differing values for carry-over of effect of $CE_{AA} = 0.04R_m$ and $CE_{AP} = 0.43R_m$: a ratio of over 10 to 1!

Thus, to sum up, not only is the simple carry-over model inconsistent with standard theories of pharmacokinetics and pharmacological response, but there is no reason to suppose that it will provide an acceptable approximation to reality unless hardly any carry-over has taken place, in which case it is redundant.

10.3.3 The models which incorporate it are self-contradicting

Consider a factorial cross-over in four periods in which each patient receives either treatment A and B together (AB), A and placebo to B (A*), placebo to A

and B ($*B$), or placebo to A and placebo to B ($**$). 'Optimal' designs assuming simple carry-over have been recommended in connection with such cross-overs as well as similar cross-overs in incomplete blocks (Fletcher *et al.*, 1990). One of the common parameterizations of the standard factorial model will require us to fit parameters for the main effect of A, τ_A, and of B, τ_B, and for interaction, τ_{AB}. The simple carry-over model requires us to fit corresponding carry-over parameters γ_A, γ_B and γ_{AB}. There will also be further parameters for periods and patients but they need not concern us here.

Now consider a patient who is allocated to receive the sequence: $AB^{**}A^{**}B$. The treatment and carry-over effects which apply to him may be schematically expressed as in Table 10.3.

If we look at the parameters for the second period we can see that not only do we allow for the carry-over of A and of B but also for the carry-over of the interaction of A and B, or equivalently, for the interaction of the carry-over of A and the carry-over of B. If we look at the fourth period, however, we see that although we allow for the direct effect of B and the carry-over of A we do not allow for the interaction of the carry-over and the direct effect, yet this must be more important than the interaction of the carry-over of A and the carry-over of B for which we allow.

Of course, only a minority of cross-over trials are factorials but this is not a reasonable defence of the simple carry-over model, for almost any treatment which is suitable for study in a cross-over design could be envisaged as a suitable candidate for combination therapy and almost any combination therapy trial could reasonably be run as a factorial (Pledger, 1989). Hence to defend the use of the simple carry-over model in any trial in which there are two active treatments one would have to argue either that the mere fact that one wishes to use a factorial treatment structure rules out the cross-over design or that there is some special reason why the simple carry-over model does not apply to such designs. But the issue is really whether carry-over and direct treatment will interact and this has to do with the nature of the treatments and the wash-out, *not* with whether the investigator wishes to study such interactions. Similar

Table 10.3 Parameterization of effects for a patient in a factorial cross-over.

	Period			
	I	II	III	IV
Treatmeant combination	AB	$**$	$A*$	$*B$
Treatmeant parameters	τ_A, τ_B, τ_{AB}		τ_A	τ_B
Carry-over parameters		γ_A, γ_B, γ_{AB}		γ_A

inherent contradictions for dose finding cross-overs have been drawn attention to by Senn and Hildebrand (1991).

10.3.4 The estimators based on it are inefficient

To illustrate this point we return to the four-period cross-over in two treatments considered in Section 10.2 and suppose that we have an active treatment, *A*, and a placebo, *P*, so that we may speak of sequences *AAPP* and *PPAA*. Suppose that one quarter of the active substance from *A* carries over to the end of the following period and that the simple one-compartment model considered in Section 10.3.2 applies. It now follows that at the end of two consecutive treatment periods with *A*, the total plasma concentration is 1.25× what it is at the end of one period, so that from (10.6) and (10.7),

$$\frac{C_{\min_{2N}}}{C_{\min_N}} = \frac{C_{\min_\infty}\left(1 - e^{-2Nk\tau}\right)}{C_{\min_\infty}\left(1 - e^{-Nk\tau}\right)} = \frac{5}{4}, \tag{10.16}$$

where N_τ is the length of a treatment period. Solving (10.16) we have $Nk\tau = 1.386$ and $2Nk\tau = 2.772$. If we let the concentration at the end of the first active period be C_A, we already know that the concentration at the end of the second active period is $1.25C_A$. Thus for patients in the second sequence, *PPAA*, the concentrations at the end of periods I, II, III, IV are 0, 0, C_A and $1.25C_A$. For patients in the first sequence, for periods I and II the figures are C_A and $1.25C_A$. The values for periods III and IV are then given by $1.25C_A e^{-1.3863}$ (or $0.25 \times 1.25C_A$) and $1.25C_A e^{-2.7726}$ (or $0.25 \times 0.25 \times 1.25C_A$) and are $0.3125C_A$ and $0.078\ 125C_A$ respectively.

Suppose, further, that the simple model of pharmacological response outlined in (10.10) applies and, as in the example considered at the end of Section 10.3.2, the value of the parameter α is 1 and the response given a concentration C_A is $3R_m/4$. We then have, as before, $C_{R50} = C_A/3$. Suppose, for the moment, that the treatment effect, τ, which we wish to measure is actually that corresponding to the administration of *A* for a single period, then we have $R_m = 4\tau/3$. Substituting in (10.10) we have

$$R = \frac{4\tau C}{C_A + 3C}. \tag{10.17}$$

Substituting the values of 0, C_A, $1.25C_A$, $0.3125C_A$ and $0.078\ 125C_A$ for C in (10.17), we then obtain the relevant responses for the trial. The position is given in Table 10.4.

By multiplying an appropriate set of weights by corresponding responses we can obtain the treatment effect which would be estimated by a given estimator.

Table 10.4 A placebo, P, controlled trial of an active treatment, A, in an $AAPP/PPAA$ cross-over: plasma concentrations ($/C_A$) and responses ($/\tau$) when 25% carry-over applies, as well as weights for unadjusted and adjusted estimators.

		Period			
Sequence		I	II	III	IV
AAPP	Plasma conc.	1	1.25	0.3125	0.078 125
	Response	1	1.052 63	0.645 16	0.253 16
	Weights (unadj.)	5/20	5/20	−5/20	−5/20
	Weights (adj.)	6/20	4/20	−7/20	−3/20
PPAA	Plasma conc.	0	0	1	1.25
	Response	0	0	1	1.052 63
	Weights (unadj.)	−5/20	−5/20	5/20	5/20
	Weights (adj.)	−6/20	−4/20	7/20	3/20

For the unadjusted estimate we have a figure of 0.8017τ and for the adjusted estimate of 0.7546τ. There might well be some argument as to whether the response after one period, τ, or the response after two periods of active treatment, 1.053τ, or even the steady-state response, is what we should be estimating. ($C_{\min_\infty}$ may be shown to be $\frac{4}{3}C_A$, so that the steady state response is 1.067τ.) Whichever of these three is most appropriate it is clear, however, that the bias of the adjusted estimator is worse than that of the unadjusted estimator. Since the variance of the adjusted estimator is also higher than that of the unadjusted estimator, there is no point in using the adjusted estimator.

Of course this is only one example, but others produce similar result. Senn and Lambrou (1998) considered a more general model for placebo controlled cross-over trials. They supposed that the carry-over from an active treatment when followed by an active treatment would be some fraction r of the carry-over when followed by placebo. A value of $r = 1$ corresponds to the simple carry-over model. When $r = 0$, the steady-state model of Fleiss (1986b, 1989), whereby carry-over from a drug into itself is impossible, applies. In general, some value in between might apply and, conceivably, one could have $r > 1$, corresponding to being on a steepening point of the dose–response scale, or $r < 0$, corresponding to tachyphilaxis. Their investigation showed that if you did not know which form of carry-over applied, not adjusting was often a better strategy than adjusting for either the steady-state or simple carry-over model.

In summary, if we adjust our estimates using the simple carry-over model

- we shall increase their variance;
- we may increase their bias; and
- we may misleadingly give the impression that we no longer need assumptions regarding carry-over.

10.3.5 The designs associated with it are not necessarily better than others

In the example of the *AAPP*/*PPAA* cross-over which we have just considered, it turned out that when carry-over was modelled in slightly more realistic terms the adjusted estimator was inferior to the unadjusted one. It does not necessarily follow that the design is inefficient. It might be the case that certain designs which have been developed in association with the simple carry-over model are, nevertheless, the best to use even if the adjusted estimator itself is not.

For example, a common design used in association with the simple carry-over model and adjusted estimators, when an even number of treatments are being studied in the same number of periods, is the *Williams square* (Williams, 1949). We mentioned this design briefly in Chapter 5. It has the property that every treatment follows every other exactly once. Thus if we have four treatments, *A*, *B*, *C* and *D* and arrange them as follows:

	Period			
	I	II	III	IV
	A	*B*	*C*	*D*
Sequence	*B*	*D*	*A*	*C*
	C	*A*	*D*	*B*
	D	*C*	*B*	*A*

the sequences form a Williams square. If simple carry-over applies, such designs are optimal. In the case of a dose-finding cross-over, however, in which *A*, *B*, *C* and *D* were different doses of a given drug, it has been shown (Senn, 1992a) that not only may the unadjusted estimator be superior to the adjusted one but also that a design in which each dose is preceded by the immediately higher dose can be preferable. Thus, if the alphabetical order of the treatment labels corresponds to the size of the doses (from lowest to highest) and if the response is logarithmic in the dose of the drug and if the carry-over of dose is proportionate to the dose, then a design of the form

	Period			
	I	II	III	IV
	D	*C*	*B*	*A*
Sequence	*A*	*D*	*C*	*B*
	B	*A*	*D*	*C*
	C	*B*	*A*	*D*

is preferable.

In the case of a placebo-controlled trial of an active treatment in two sequences, consider the *AAPP*/*PPAA* design we have looked at in this chapter. Suppose we simply replaced it by the standard *AP*/*PA* design but used twice the length of periods. This is really no different from using the *AAPP*/*PPAA* design but measuring in periods II and IV only. (We have already discussed the relationship between these two designs is Section 10.3.1.) We may then use the results presented in Table 10.4 but apply the weights $0\ \frac{1}{2}\ 0\ -\frac{1}{2}$ for the first sequence and $0\ -\frac{1}{2}\ 0\ \frac{1}{2}$ for the second sequence. The expected estimated treatment effect is then 0.9260τ and we thus see that the bias, which is 0.0740τ if we are trying to estimate τ, is considerably less than that associated with either of the two estimators we used in connection with the *AAPP*/*PPAA* design. The biases there were 0.1983τ (unadjusted) and 0.2454τ (adjusted).

Of course, the estimator will have a higher variance. If we make the extremely strong assumption that the error terms are all independent, then if the variance of the estimator for the *AP*/*PA* design is σ^2 the variance of the unadjusted estimator will be $0.55\sigma^2$. In practice this assumption would not be satisfied and the disparity is likely to be less. For the sake of argument let us assume that it applies. To compare such estimators we may use the mean square error (the sum of the square of the bias and the variance). The mean square error for the unadjusted estimator for the *AAPP*/*PPAA* design is $0.0393\tau^2 + 0.5\sigma^2$ as compared with $0.0055\tau^2 + \sigma^2$ for the *AP*/*PA* design. We thus see that the former is lower than the latter if $\tau^2 < 14.8\sigma^2$ or if $\tau < 3.85\sigma$. At the point of indifference, where $\tau = 3.85\sigma$, the mean square error is $1.08\sigma^2$ and its square root is 1.04, giving a ratio of parameter to the root mean square error of 3.7, a precision which is good but not exceptional for a cross-over trial. Values of higher precision would tend to favour the *PA*/*PA* design and so would a more realistic modelling of the variance.

All of these calculations are, of course, highly speculative. Obviously quite other proportions of carry-over could be envisaged as could other values for the parameters of the Hill equation, and indeed quite different pharmacokinetic models and models of pharmacological response could apply. The important point to note, however, is this. Different assumptions about carry-over make different designs 'optimal'. The standard theory of 'optimal' design of cross-over designs has been developed almost exclusively by using one mathematically attractive but biologically unrealistic model of carry-over.

For these reasons I am extremely sceptical about the value of such 'optimal' designs and I do not advocate their use. The investigator who designs a cross-over trial should realize that the validity of his conclusions depends on having designed a trial with adequate wash-out. Some doubt must always remain attached to any trial as to whether the wash-out was adequate. I can see no advantage in pretending otherwise.

10.4 IN SUMMARY: THE CASE AGAINST MATHEMATICAL APPROACHES TO CARRY-OVER

This is a summary of the case against adjusting for carry-over.

- The simple carry-over model has been developed without reference to pharmacological or biological models.
- It does not provide a useful approximation to reality.
- It leads to more complicated estimation procedures which are more difficult to describe and understand.
- Usually the adjusted estimators have higher variance than the unadjusted ones.
- Although the adjusted estimators will be unbiased if simple carry-over applies, in practice if carry-over occurs they will be biased and it is perfectly possible that this bias will be larger than it is for unadjusted estimators.
- The most serious objection, however, is that the use of such approaches encourages the erroneous belief that the validity of estimates obtained from cross-over trials does not depend on adequate wash-out having taken place.

I hope I have persuaded the reader that these points are correct. If so, there is no reason for him to read any further. I have now finished outlining what I consider the ordinary practitioner needs to know about planning and analysing cross-over trials.

10.5 USING THE SIMPLE CARRY-OVER MODEL

There is a vast literature on the subject of using the simple carry-over model. There is only space for the briefest introduction here. First, we shall have a look at estimation techniques for some of the 'optimal' two-treatment designs in two sequences. Such designs can be analysed fairly simply using parametric and non-parametric approaches. Secondly, we shall examine Wiliams square designs. Finally, we shall indicate where further help may be sought.

10.5.1 Treatment estimates for dual balanced designs in two sequences

A design consisting of pairs of treatment sequences, where one sequence in a pair, may be obtained simply by changing the labels from the other, is said to be *dual balanced* (Matthews, 1990a, p. 5.4). Amongst such designs, the simplest are those consisting of a single pair. The *ABA*/*BAB* design is an example of such a design and so is the *ABBA*/*BAAB* design. The *AB*/*BA* design is also dual balanced in two sequences.

In general, if we wish to estimate treatment effects for a cross-over trial we can do so by defining cells of observations corresponding to the cross-classification of sequence and period, calculating the means of such cells and forming an appropriate weighted sum. We shall now illustrate how the appropriate weights for such a calculation may be derived under the assumption that the simple carry-over model applies. We shall use two examples, the *ABA*/*BAB* design and the *ABBA*/*BAAB* design. We start with the three-period design.

First we represent the design schematically as:

	Period		
	I	II	III
Sequence			
1	*A*	*Ba*	*Ab*
2	*B*	*Ab*	*Ba*

Capital letters correspond to treatments and small letters to simple carry-over.

The six weights we need to find for the six cell means are set out in a way which parallels this representation:

$$\begin{array}{ccc} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6. \end{array}$$

Now, if we want a treatment effect from which the influence of periods has been eliminated, then in a given column the weights must add to zero. Thus the weights in the second row must simply be the negative of those in the first. Hence we may rewrite the weights as

$$\begin{array}{ccc} w_1 & w_2 & w_3 \\ -w_1 & -w_2 & -w_3. \end{array}$$

Similarly, if we wish differences between sequences (and patients) to be eliminated then the weights must add to zero along any row. Thus we may now write the weights as

$$\begin{array}{ccc} w_1 & w_2 & (-w_1 - w_2) \\ -w_1 & -w_2 & (w_1 + w_2). \end{array}$$

Turning now to simple carry-over, the sum of weights associated with a given lower case letter must also be zero so that $w_2 + (w_1 + w_2) = 0$ and hence $w_2 = -w_1/2$, and we may write the weights

$$\begin{array}{ccc} w_1 & -w_1/2 & -w_1/2 \\ -w_1 & w_1/2 & w_1/2. \end{array}$$

Finally the sum of the weights associated with A must equal 1 and so $w_1 = 1$ and we may now establish the actual values of the weights:

$$\begin{array}{ccc} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -1 & \frac{1}{2} & \frac{1}{2}. \end{array} \tag{10.16}$$

On the assumption that the errors are independent, the variance of the treatment estimate associated with this scheme of weights is, by (2.3), proportional to the sum of the squares. In this case the sum of the squares is 3.

This design is not the optimal one, given the assumption of independent errors and the simple carry-over model. The 'optimal' design is the *ABB/BAA* design. The reader may check for himself, using the method above, that the weights associated with this design are $\frac{1}{2}\ -\frac{1}{4}\ -\frac{1}{4}/-\frac{1}{2}\ \ \frac{1}{4}\ \ \frac{1}{4}$ and that the sum of the squares of the weights is 0.75. Thus, other things being equal, the variance from the *ABA/BAB* design is four times that from the *ABB/BAA* design.

We now illustrate the method for the

$$\begin{array}{cccc} A & Ba & Bb & Ab \\ B & Ab & Aa & Ba \end{array}$$

design. As we shall see an extra step will be required at the end to obtain the weights. Why this is so we shall discuss in due course but first we illustrate the method.

First, using the property that the row and column totals for the weights must be zero, we can write down the following scheme straight away:

$$\begin{array}{cccc} w_1 & w_2 & w_3 & (-w_1 - w_2 - w_3) \\ -w_1 & -w_2 & -w_3 & (w_1 + w_2 + w_3). \end{array}$$

The fact that the weights associated with a given simple carry-over must sum to zero gives the condition $w_2 - w_3 + (w_1 + w_2 + w_3) = 0$, from which $w_2 = -w_1/2$, so that the scheme is

$$\begin{array}{cccc} w_1 & -w_1/2 & w_3 & (-w_1/2 - w_3) \\ -w_1 & w_1/2 & -w_3 & (w_1/2 + w_3). \end{array}$$

Because the sum of the weights associated with A must be 1 we have $w_1 + w_1/2 - w_3 + (-w_1/2 - w_3) = 1$ or $w_3 = w_1/2 - \frac{1}{2}$. Substituting for w_3 we then obtain the scheme of weights

$$\begin{array}{cccc} w_1 & -w_1/2 & (w_1/2 - \frac{1}{2}) & (\frac{1}{2} - w_1) \\ -w_1 & w_1/2 & (-w_1/2 + \frac{1}{2}) & (-\frac{1}{2} + w_1). \end{array} \qquad (10.19)$$

We have now, however, used up all of our constraints and we do not have a unique set of weights. We need to impose a further condition to find a unique set. Since we are searching for the most efficient estimator and we assume that the errors are independent, we must choose the value of w_1 which makes the sum of the squares a minimum. In fact, since the sum of the squares of the first row is the same as the sum of the squares of the second it is sufficient to minimize the former. Multiplying the weights in (10.19) by 2 first to clear fractions, squaring and gathering terms together we need to minimize:

$$S = 10w_1^2 - 6w_1 + 2.$$

Dividing the right-hand side by 10 we have

$$w_1^2 - 6w_1/10 + 2/10.$$

Completing the square we get

$$(w_1 - 3/10)^2 + 2/10 - 9/100.$$

Since the first term is squared it can never be negative and its lowest value is achieved when $w_1 = 3/10$. This is thus the value we are looking for. Writing 3/10 as 6/20 the weights are:

$$\begin{array}{cccc} 6/20 & -3/20 & -7/20 & 4/20 \\ -6/20 & 3/20 & 7/20 & -4/20. \end{array}$$

Although the pattern of weights is not identical to that given in Table 10.2 for the *AABB*/*BBAA* design, the same individual values are used. Hence the sums of the squares of the weights for the two designs are identical (they are equal to

0.55) and under the assumptions of simple carry-over and independent errors the two designs have equal efficiency (Jones and Kenward, 1989, p. 180). Given these assumptions, they are in fact the two most efficient designs in two sequences and four periods. The reader may be interested to try and derive for himself the weights for the *AABB*/*BBAA* cross-over given in Table 10.2.

The reason that an extra step was required at the end of the derivation of weights for the *ABBA*/*BAAB* design compared to that for the *ABA*/*BAB* design, is that the former, by virtue of having more cell means, had more degrees of freedom available for estimating parameters. A number of weighting schemes were thus possible. The three-period design had six cell means. This number is fully accounted for by using 1 degree of freedom for each of grand mean, treatments, sequences (or patients) and carry-over as well as two for periods. The four-period design has eight cell means but the model which we implicitly used had one each for grand mean, treatments, sequences, and carry-over and three periods: a total of only seven. We used the extra degree of freedom to find an 'optimal' weighting scheme amongst those which would estimate the treatment effect unbiasedly in the presence of patient, period and simple carry-over effects.

There are other uses one can make of this degree of freedom. Suppose we allow that we may have second order carry-over. We may represent this schematically as

$$\begin{array}{cccc} A & Ba & Bba' & Abb' \\ B & Ab & Aab' & Baa' \end{array}$$

where a' is the carry-over of A into the next but one period and b' is the equivalent carry-over for B. If we take the weights given by (10.19), then we see that by setting $(w_1/2 - 1/2) + (-1/2 + w_1) = 0$ we can eliminate the influence of a' and b'. This yields the unique solution $w_1 = 2/3$ and the scheme of weights

$$\begin{array}{cccc} 4/6 & -2/6 & -1/6 & -1/6 \\ -4/6 & 2/6 & 1/6 & 1/6. \end{array} \tag{10.20}$$

The sum of the squares of the weights is now 1.222 compared to 0.55 before. In other words, assuming independent errors, the variance of the estimator which uses this further adjustment is 2.22 times larger. Even for a proponent of carry-over adjustment, however, this would seem to be a particularly large price to pay to remove this extra source of bias. The estimator has a strongly counter-intuitive form since one of the weights for A (the last one in the first row) is actually negative.

10.5.2 Variances of estimates for dual balanced designs in two sequences

So far we have covered the way in which treatment effects may be estimated under the simple carry-over model. We also need to be able to estimate the variances of such estimates. We now illustrate how this may be done using an example.

Example 10.1 Table 10.5 gives data presented by Hafner *et al.* (1988) who also describe an approach to analysing dual balanced designs in two sequences which is almost identical to that presented here. The data are taken from a cross-over trial 'conducted to investigate the metabolism of pentobarbitol in young and mature mice exposed to 0.3 p.p.m. of ozone for 3.75 hours' (p. 472). Only the data for young mice are presented here. The treatment days were separated by one week wash-outs and each mouse was allocated at random to one of the two sequences: ozone, air, ozone (*OAO*) or air, ozone, air (*AOA*). Following exposure to treatment (air or ozone) on the given day, the mouse was immediately given an intramuscular injection of pentobarbitol in a dose proportional to body weight. The outcome measured was the observed sleeping time for each mouse. Further details of the experiment may be found in Canada *et al.* (1986).

Table 10.5 (Example 10.1) Minutes of sleep for nine young mice exposed to both ozone and air. Individual values as well as unadjusted and adjusted basic estimators. (Data taken from Hafner *et al.* 1988).

		Original values		Basic estimators	
Sequence	Period 1	Period 2	Period 3	Unadjusted	Adjusted
OAO	38.1	23.8	16.4	3.45	36.0
OAO	19.7	16.6	15.3	0.90	7.5
OAO	39.0	25.3	30.2	9.30	22.5
OAO	31.7	19.1	17.6	5.55	26.7
OAO	16.8	21.8	20.1	− 3.35	− 8.3
mean	29.06	21.32	19.92	3.17	16.88
CSS				90.98	1215.61
AOA	22.3	21.8	15.1	3.10	− 7.7
AOA	24.0	19.5	15.1	− 0.05	− 13.4
AOA	17.8	17.6	21.9	− 2.25	3.9
AOA	15.5	20.0	10.1	7.20	− 0.9
mean	19.9	19.725	15.55	2.00	− 4.525
CSS				50.52	172.97
Mean both sequences				2.585	6.1775
CSS both sequences				141.50	1388.58
Variances of basic estimators				20.21	198.37

Table 10.5 gives the result for each mouse for each treatment day and also two basic estimators. The unadjusted basic estimator is calculated on the assumption that carry-over does not apply and using the weights $\frac{1}{2}$, -1, $\frac{1}{2}$ for the sequence *OAO* and $-\frac{1}{2}$, 1, $-\frac{1}{2}$ for the sequence *AOA*. This scheme simply calculates for each mouse the difference between the average sleeping time, having been treated with ozone, and the average having been treated with air. The adjusted basic estimators are calculated using the weights 2, -1, -1 for the *OAO* sequence and $-$ 2, 1, 1 for the *AOA* sequence. These weights are exactly twice the values we found in Section 10.5.1 and gave in (10.18). There, however, we were talking in terms of weights for cell means. Here we first average the basic estimator for each sequence and then average the means from each sequence. This final step, involving a division by 2, makes the result identical to that we would have obtained by applying the weights given in (10.18) to the cell means. The reader may check this for himself using the six means given for the three periods of the two sequences in Table 10.4. To perform the same check for the unadjusted estimator he should use the cell mean weights $\frac{1}{4}$ $-\frac{1}{2}$ $\frac{1}{4}$ and $-\frac{1}{4}$ $\frac{1}{2}$ $-\frac{1}{4}$.

Once the basic estimators have been defined the analysis may proceed using exactly the same procedure we have illustrated before for basic estimators. The general procedure was discussed in Section 5.4.1 and the Hills–Armitage analysis of the *AB*/*BA* design illustrated in Section 3.6 is an example of its application to a design in two sequences (Hills and Armitage, 1979). Alternatively, if we reverse the sign of the basic estimators for the second sequence group, then the problem is reduced to the extremely common one of comparing two samples. Thus, we can use the alternative representation of the Hills–Armitage approach as a two-sample *t* test (Section 3.5) or even Koch's (1972) adaptation of the Wilcoxon–Mann–Whitney test (Section 4.3.8). These methods of analysis are illustrated in Hafner *et al.* (1988).

Returning, however, to the basic estimator approach, all that has been done is to produce a treatment estimate using the standard steps of averaging within each sequence and then averaging over sequences. Thus, using unadjusted basic estimators, we have an estimated treatment effect (ozone–air) of 2.6 minutes and, using adjusted basic estimators, of 6.2 minutes. To obtain the variances we use our standard rules for linear combinations given in Section 2.2.1. The variances of the means of each sequence are proportional to $\frac{1}{5}$ and $\frac{1}{4}$ respectively and averaging produces an overall mean with variance proportional to $(\frac{1}{5}+\frac{1}{4})/4 = 0.1125$. This is the figure by which the variances of the individual basis estimators must be multiplied to obtain the variance of the treatment estimate. The variances of the individual basic estimators have been calculated by dividing the appropriate corrected sums of squares by the degrees of freedom $(5-1+4-1)=7$. Putting these results together we obtain variance estimates of 2.274 minutes^2 for the unadjusted estimator and of 22.317 minutes^2 for the adjusted estimator. Taking square roots we obtain figures of

1.508 minutes and 4.724 minutes. The critical value of Student's *t* corresponding to a 5% test (two tailed) for 7 degrees of freedom is 2.3646 (Diem and Seldrup, 1982, p. 30; Lindley and Scott, 1984, p. 45). Multiplying the standard errors by this value and adding or subtracting the result from the corresponding treatment estimate we obtain 95% confidence intervals for the treatment effect of:

Unadjusted	– 1.0 minutes to 6.2 minutes
Adjusted	– 5.0 minutes to 17.3 minutes

It is noticeable that the adjusted estimate has much wider confidence limits, reflecting its lesser efficiency.

10.5.3 Williams designs

A design commonly used in connection with the simple carry-over model, and one which we have already encountered in Chapter 5 and again in Chapter 7, is the Williams square (Williams, 1949). A Williams square is a Latin square in which every treatment follows every other treatment once. Such designs are possible if there are an even number of treatments and occasionally for an odd number of treatments. Where not, by using two Latin squares, a design may be used in which each treatment follows every other treatment twice. See Newcombe (1996) for some other possibilities, and also Prescott (1999). We gave the Williams squares for designs in four treatments in Table 5.1. For a three-treatment cross-over the Williams design consists in using all six possible sequences. For a five-treatment design, 10 sequences are necessary and there is a choice of Williams designs. Matthews (1990a, pp. 4.1–2), gives a clear description of a general algorithm, due to Sheehe and Bross (1961), by which such designs may be constructed. A similar description is given by Jones and Kenward (1989, pp. 197–9).

If simple carry-over applies, then Williams squares are optimal designs given a particular restriction on the number of periods, patients and sequences available for study. In practice, however, if other forms of carry-over apply, such designs are not optimal (Senn, 1992). Deliberately choosing a Williams design, therefore, only makes sense if simple carry-over is believed to apply.

We shall illustrate the way in which the simple cross-over model may be used in connection with a Williams square using Example 7.1. Since, however, factorial structures introduce an additional unnecessary complication, for the purpose of illustrating the method we shall assume that instead of estimating the factor contrasts we wish, instead, to compare two treatments, suspension 12 μg and suspension 24 μg. In Example 7.1 these two treatments were given the labels *A* and *B*. The treatments solution 12 μg solution 24 μg were given the labels *C* and *D*. Table 10.6 reproduces the Williams square which was used in Example 7.1 and also the cell mean weights which might be used to calculate a *B*–*A* contrast adjusted for simple carry-over.

If the weights in Table 10.6 are studied it will be seen that they have the property that they add to zero for any given period as well as for any sequence and also for the treatments *C* and *D*, whereas for *A* the total is -40 and for *B* the total is 40. Since the weights we would use for estimation have been multiplied by 40 we thus measure the contrast of interest, namely *B*–*A*. A much more obvious and simpler scheme, which also has these properties, is simply to use −10 for each *A* and 10 for each *B*. The weights given in Table 10.6 also have the property, however, that when totalled over cells with the same lower-case letters they come to zero. Since these lower-case letters represent the treatments in the previous cell the weights have the property that they will eliminate the effect of carry-over providing it is determined by the previous treatment only and lasts exactly for one period. In other words they adjust the treatment contrast for simple carry-over.

In order to use these weights for analysis we may make a simple adaptation of the basic estimator approach. We calculate an adjusted basic estimator for each patient by using these weights and multiplying by 4. (This factor is necessary if we wish to keep the method in line with the alternative approach of calculating a treatment contrast for each patient and obtaining a final estimate by repeated averaging. An alternative approach which comes to the same thing is to use the weights as they are, obtaining a mean result per sequence, but then adding the sequence means rather than averaging again.)

Applying the weights to Table 7.1 we obtain the adjusted basic estimators given in Table 10.7. Once this step has been taken the rest of the analysis

Table 10.6 Cell weights for calculating an adjusted treatment estimate for the *B*–*A* contrast for a Williams square (Senn, 1992a).

		Weights (× 40) for the estimator adjusted for simple carry-over Period			
		1	2	3	4
Sequence	I	*A* −11	*Ba* 7	*Db* 4	*Cd* 0
	II	*B* 11	*Cb* 4	*Ac* −11	*Da* −4
	III	*C* −1	*Dc* −1	*Bd* 10	*Ab* −8
	IV	*D* 1	*Ad* −10	*Ca* −3	*Bc* 12

Table 10.7 (Example 7.1) Adjusted basic estimators for the suspension 24 μg–suspension 12 μg treatment contrast.

Sequence							
ABDC		*BCAD*		*CDBA*		*DACB*	
Patient	Estimator	Patient	Estimator	Patient	Estimator	Patient	Estimator
3	−0.74	4	−0.07	2	−0.01	1	0.06
5	−0.11	6	1.07	8	−0.02	7	−0.20
12	−0.11	10	0.19	9	0.09	11	0.39
13	0.38	16	0.35	14	0.25	15	0.09
mean	−0.145		0.385		0.0775		0.085
CSS	0.6321		0.7155		0.047 075		0.1749

Overall mean $= 0.100\,625\,\ell$

Total $CSS = 1.569\,575\,\ell^2$

$MSE = 1.569\,575/12 = 0.130\,80\,\ell^2$

Variance of treatment estimate $= 0.130\,80(\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4})/16 = 0.008\,175\,\ell^2$

Standard error $= 0.0904\,\ell$

$t = 0.100\,625/0.0904 = 1.11$

Critical value of t (5% two tailed, 12 degrees of freedom) = 2.18

follows exactly the form outlined for the standard basic estimator approach in Section 5.4.1. The various intermediate calculations are shown in Table 10.7.

Remark The method outlined is only approximately valid even if the simple carry-over model applies and even if the error terms are all independent. If this latter condition is met, then the variances of the individual terms used to obtain estimates are correct. For models which do not adjust for carry-over the same weights are used for each patient. For adjusted estimators this is not the case. For example, in Table 10.6 the sums of the squares of the weights for the four sequences are proportional to: 186, 274, 166, 254. This inequality of variance makes the analysis only approximate. The standard ordinary least squares analysis which we shall illustrate below is valid if the assumptions are guaranteed. I am of the opinion that in practice they will not apply. Correlations among the errors and, in particular, patient by treatment interaction, will mean that ordinary least squares may well be more problematical than the approach described. If, as seems likely to me, the simple carry-over model does not apply, then both approaches are unsatisfactory (Senn and Hildebrand, 1991; Senn, 1992).

We do not have space here to cover the derivation of weights for cell means for cross-over designs in general. Since, however, the weights are easily specified for the three-period three-treatment cross-over in six sequences, and since this design is probably the most commonly encountered of all designs in more than two periods, we specify the weights for the *A*–*B* contrast for a design with treatments *A*, *B* and *C* below. (Since the labels are arbitrary and since all possible sequences are represented it is a simple matter to obtain the other two pairwise contrasts simply by switching labels.) The weights are

A	*Ba*	*Cb*	*C*	*Ac*	*Ba*	*B*	*Cb*	*Ac*
5/24	−2/24	−3/24	−1/24	4/24	−3/24	−4/24	−2/24	6/24
B	*Ab*	*Ca*	*C*	*Bc*	*Ab*	*A*	*Ca*	*Bc*
−5/24	2/24	3/24	1/24	−4/24	3/24	4/24	2/24	−6/24.

It will be seen that these have the property of adding to one when summed over treatment *A* and to minus one over *B* as well as to zero in each period and for each sequence and when summed over treatment *C* and over the simple carry-over labels *a*, *b* and *c*.

10.5.4 Ordinary least squares analysis

The analyses which we have illustrated so far are analogous to those considered in Section 5.4.1: we first reduce all the measurements on a patient to a single measure. If, however, we are prepared to make the strong assumption that all ‘errors’ are independent (and of constant variance), then we can carry out an analysis using ordinary least squares which is analogous to that in Section 5.4.3. If we are performing an analysis using SAS® *proc glm* then we can add an extra factor to reflect the previous treatment.

Thus, for example, one way of analysing Example 7.1 is as follows. As in Chapter 7 we create a data set with the 64 outcomes (4 each for 16 patients) stored under the variable name *Y*. We also have variables *PATIENT*, *PERIOD* and *TREAT*, which record for each of the 64 outcomes for which patient it was recorded and in which period and what treatment the patient was receiving. *PATIENT* will be a variable with 16 levels whilst *PERIOD* and *TREAT* will have four each. We also define a factor *CARRY* with levels *A*, *B*, *C* and *D* for the previous treatment and level Z when there is no previous treatment. (Obviously Z is completely confounded with the first period but SAS® *proc glm* can deal with this.)

The following code will then produce the *OLS* equivalent of the analysis above:

```
proc glm;
  class TREAT PERIOD PATIENT CARRY;
  model Y = PATIENT PERIOD TREAT CARRY;
  estimate ‘suspension’ TREAT -1 1 0 0;
run;
```

As usual, the *estimate* statement is used to define a particular contrast of interest. Amongst the output will be found the following:

```
Dependent Variable Y
Source           DF            Sum of squares   F value   Pr > F
Model            24            23.061 343 75    9.44      0.0001
Error            39              3.968 500 00
Corrected total  63            27.029 843 75
                               T for H0         Pr > |T|  Std Error of
Parameter        Estimate      Parameter = 0              estimate
suspension       0.100 625 00  0.85             0.4001    0.118 285 70.
```

The degrees of freedom for error are thus 39 and for the purpose of calculating confidence intervals etc. we would therefore use a critical value of *t* based on 39 degrees of freedom. The estimate is exactly the same as that which we produced above; however, its standard error is estimated under OLS as 0.118 ℓ as opposed to the figure of 0.090 ℓ which we obtained before.

10.6 WHERE TO FIND OUT MORE ABOUT THE SIMPLE CARRY-OVER MODEL

Criticisms of the simple carry-over model will be found in Fleiss (1986b, 1989), Senn (1992), Senn and Lambrou (1998) and Lambrou (2001). There is no shortage of uncritical papers which discuss how the model may be used. We list some below which have appeared since 1980.

Advice on how to use the simple carry-over model is provided by Jones and Kenward (1989), Matthews (1990a) and Ratkowsky *et al.* (1993). Kershner and Federer (1981), Laska *et al.* (1983) and Laska and Meisner (1985) carried out extensive investigations into optimal design. Other important papers are those of Afsarinejad (1990) and Matthews (1990b). Some results regarding two-treatment designs are given in Carriere (1994) and Carriere and Reinsel (1992). Optimal factorial designs are considered by Fletcher *et al.* (1990). One of the best introductions to the subject, which also has many original results, is Bishop and Jones (1984). A useful review, covering many issues in cross-over trials, including simple carry-over, is that of Matthews (1988). Investigations covering simple carry-over and more general correlation structures, will be found in Kunert (1987a), Matthews (1987) and Kushner (1997). A topic that is not essentially related to the use of the simple carry-over model, but which may also appear when using that model, is that of variance estimation. This is considered by Kunert (1987b), Matthews (1989) and Guilbaud (1993). Deheuvals and Derzko (1991) also cover optimality issues for the simple carry-over model in a paper concerned, however, with incomplete block designs for correlated error structures. A rather special incomplete blocks design in two periods for comparing two active treatments is considered by Koch *et al.* (1989). Use of a random effects (or mixed) model is covered by Laird *et al.* (1992) and Putt and Chinchilli (1999). Hafner *et al.* (1988) cover robust parametric and nonparametric analyses for two treatment designs in two sequences, and Putt and Chinchilli (2000) also cover robust methods. An application of the simple carry-over model to the analysis of binary data from the *ABB*/*BAA* design is given by Morrey (1989). Various designs for trials with more than two treatments are considered by Newcombe (1992, 1996) and Prescott (1999). An example of an application of the simple carry-over model in the medical literature is Guyenne *et al.* (1989).

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

References

Afsarinejad, K. (1990) Repeated measurements designs—a review. *Communications in Statistics: Theory and Methods* **19**, 3985–4028.

Aitken, R. C. B. (1969) Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine* **62**, 989–93.

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

Altman, D. G. and Bland, J. M. (1991) Improving doctors' understanding of statistics (with discussion). *Journal of the Royal Statistical Society A* **154**, 223–68.

Anderson, S. and Hauck, W. W. (1990) Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* **18**, 259–73.

Armitage, P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 223–68.

Armitage, P. and Berry, G. (1987) *Statistical Methods in Medical Research*. Blackwell Scientific Publications, Oxford.

Armitage, P. and Hills, M. (1982) The two-period cross-over trial. *The Statistician* **31**, 119–31.

Atkinson, A. C. and Donev, A. N. (1992) *Optimum Experimental Designs*. Oxford University Press, Oxford.

Balaam, L. N. (1968) A two period design with t^2 experimental units. *Biometrics* **24**, 61–73.

Barnard, G. (1990) Must clinical trials be large? The interpretation of *P*-values and the combination of test results. *Statistics in Medicine* **9**, 601–14.

Bartow, R. A. and Brogden, R. N. (1998) Formoterol. An update of its pharmacological properties and therapeutic efficacy in the management of asthma. *Drugs* **55**, 303–22. Erratum (1998) *Drugs* **55**, 517.

Berger, R. L. and Hsu, J. C. (1996) Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* **11**, 283–302.

Bergmann, R., Ludbrook, J. and Spooren, W. (2000) Different outcomes of the Wilcoxon–Mann–Whitney test from different statistics packages. *American Statistician* **54**, 72–7.

Berk, K. N. and Carey, P. (2000) *Data Analysis with Microsoft Excel*. Duxbury, Pacific Grove, CA.

Bishop, S. H. and Jones, B. (1984) A review of higher order cross-over designs. *Journal of Applied Statistics* **11**, 29–50.

Bounds, W., Molloy, S. and Guillebaud, J. (2002) Pilot study of short-term acceptability and breakage and slippage rates for the loose-fitting polyurethane male condom eZ-ON bi-directional, a randomised cross-over trial. *European Journal of Contraception and Reproductive Health* **7**.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978) *Statistics for Experimenters*. Wiley, New York.

Brown, B. (1980) The cross-over experiment for clinical trials. *Biometrics* **36**, 69–79.

Brown, H. and Prescott, R. (1999) *Applied Mixed Models in Medicine*. Wiley, Chichester.

Brown, L. D., Hwang, J. T. G. and Munk, A. (1997) An unbiased test for the bioequivalence problem. *Annals of Statistics* **25**, 2345–67.

Campbell, M. and Machin, D. (1990) *Medical Statistics: A Commonsense Approach*. Wiley, New York.

Canada, A. T., Calabrese, E. J. and Leonard, D. L. (1986) Age-dependent inhibition of pentobarbitol sleeping time by ozone in mice and rats. *Journal of Gerontology* **41**, 587–9.

Carriere, K. C. (1994) Crossover designs for clinical trials. *Statistics in Medicine* **13**, 1063–9.

Carriere, K. C. and Reinsel, G. C. (1992) Investigation of dual-balanced crossover designs for 2 treatments. *Biometrics* **48**, 1157–64.

Chassan, J. B. (1970) A note on relative efficiency in clinical trials. *Journal of Clinical Pharmacology and the Journal of New Drugs* **10**, 359–60.

Chi, E. M. (1991) Recovery of inter-block information in cross-over trials. *Statistics in Medicine* **10**, 1115–22.

Chow, S. C. and Liu, J. P. (2000) *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker, New York.

Clayton, D. and Hills, M. (1987) A two-period cross-over trial, in D. J. Hand and B. S. Everitt (eds), *The Statistical Consultant in Action*, pp. 42–57. Cambridge University Press, Cambridge.

Cleveland, W. S. and McGill, R. (1987) Graphical perception: the visual decoding of quantitative information on graphical displays of data (with discussion). *Journal of the Royal Statistical Society A* **150**, 192–229.

Cochran, W. G. and Cox, G. (1957) *Experimental Designs*. Wiley, New York.

Cody, R. P. and Smith, J. K. (1997) *Applied Statistics and the SAS Programming Language*. Prentice Hall, Upper Saddle River, NJ.

Committee for Proprietary Medicinal Products (2001) *Note for Guidance on the Investigation of Bioavailability and Bioequivalence*. European Medicines Evaluation Agency, London. http://www.emea.eu.int/pdfs/human/ewp/140198en.pdf.

Cox, D. R. (1958) *Planning of Experiments*. Wiley, New York.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B* **34**, 187–220.

Cox, D. R. (1978) Some remarks on the role in statistics of graphical methods. *Applied Statistics* **27**, 4–9.

Cox, D. R. and Reid, N. (2000) *The Theory of the Design of Experiments*. Chapman & Hall/CRC, Boca Raton, FL.

CPMP, Working Party on Efficacy of Medicinal Products (1990) *Good Clinical Practice for Trials of Medicinal Products in the European Community*. Commission of the European Communities, Brussels.

Cushny, A. R. and Peebles, A. R. (1905) The action of optimal isomers. II. Hyoscines. *Journal of Physiology* **32**, 501–10.

Dahlof, C. and Bjorkman, R. (1993) Diclofenac-K (50 and 100 mg) and placebo in the acute treatment of migraine. *Cephalalgia* **13**, 117–23.

D'Angelo, G., Potvin, D. and Turgeon, J. (2001) Carryover effects in bioequivalence studies. *Journal of Biopharmaceutical Statistics* **11**, 27–36.

Deheuvals, P. and Derzko, G. (1991) Block designs for early-stage clinical trials. International Society for Clinical Biostatistics, Brussels.

Der, G. and Everitt, B. S. (2002) *A Handbook of Statistical Analyses using SAS*. Chapman & Hall/CRC, Boca Raton, FL.

Desu, M. M. and Rhagavarao, D. (1990) *Sample Size Methodology*. Academic Press, Boston.

Diem, K. and Seldrup, J. (eds) (1982) *Introduction to Statistics, Statistical Tables, Mathematical Tables*. CIBA-Geigy Ltd, Basle.

Dietlein, G. (1981) Schematic plots—eine Alternative zur Darstellung von mittleren Verlaufskurven. *Statistical Software Newsletter* **7**, 100–3.

Dighe, S. V. and Adams, P. (1991) Bioequivalence: a United States regulatory perspective, in P. G. Welling, F. L. S. Tse and S. V. Dighe (eds), *Pharmaceutical Bioequivalence*. Marcel Dekker, New York.

Dobson, A. J. (1983) *An Introduction to Statistical Modelling*. Chapman & Hall, London.

Dubey, S. D. (1986) Current thoughts on crossover designs. *Clinical Research Practices and Regulatory Affairs* **4**, 127–42.

Ebbutt, A. F. (1984) Three-period cross-over designs for two treatments. *Biometrics* **40**, 219–24.

EFSPI Working Group (1999) Qualified statisticians in the European pharmaceutical industry: report of a European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) working group. *Drug Information Journal* **33**, 407–15.

Elteren, P. H. van (1960) On the combination of independent two-sample tests of Wilcoxon. *Bulletin de l'Institut International de Statistique* **37**, 351–61.

Ezzet, F. and Whitehead, J. (1991) A random effects model for ordinal response from a cross-over trial. *Statistics in Medicine* **10**, 901–7.

Ezzet, F. and Whitehead, J. (1992) A random effects model for binary data from crossover trials. *Applied Statistics* **41**, 117–26.

Farewell, V. T. (1985) Some remarks on the analysis of crossover trials with a binary response. *Applied Statistics* **34**, 121–8.

Faulds, D., Hollingshead, L. M. and Goa, K. L. (1991) Formoterol: a review of its pharmacological properties and therapeutic potential in reversible obstructive airways disease. *Drugs* **42**, 115–37.

FDA CDER (2000) *Guidance for Industry: Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations*. Center for Drug Evaluation and Research, FDA, Rockville, MD. http://www.fda.gov/cder/guidance/3615fnl.htm.

Feingold, M. and Gillespie, B. W. (1996) Cross-over trials with censored data. *Statistics in Medicine* **15**, 953–67.

Fidler, V. (1984) Change-over trials with binary data: mixed model comparison of tests. *Biometrics* **40**, 1063–70.

Fidler, V. (1986) Change-over trials with binary data: estimation. *Statistica Neerlandica* **40**, 81–6.

Finney, D. J. (1978) *Statistical Methods in Biological Assay*. Griffin, London.

Fisher, R. A. (1925) Applications of 'Student's' distribution. *Metron* **5**, 90–104.

Fisher, R. A. (1990a) [1925] Statistical methods for research workers, in J. H. Bennet (ed.), *Statistical Methods, Experimental Design and Scientific Inference*. Oxford University Press, Oxford.

Fisher, R. A. (1990b) [1935] The design of experiments, in J. H. Bennet (ed.), *Statistical Methods, Experimental Design and Scientific Inference*. Oxford University Press, Oxford.

Fisher, R. A. (1990c) [1956] Statistical methods and scientific inference, in J. H. Bennet (ed.), *Statistical Methods, Experimental Design and Scientific Inference*. Oxford University Press, Oxford.

Fisher, R. A. and Yates, F. (1974) *Statistical Tables for Biological Agricultural and Medical Research*. Longman, Harlow.

Fleiss, J. L. (1986a) *The Design and Analysis of Clinical Experiments*. Wiley, New York.

Fleiss, J. L. (1986b) Letter to the editor. *Biometrics* **42**, 449–50.

Fleiss, J. L. (1989) A critique of recent research on the two-treatment cross-over design. *Controlled Clinical Trials* **10**, 237–43.

Fletcher, D. J., Lewis, S. M. and Matthews, J. N. (1990) Factorial designs for crossover clinical trials. *Statistics in Medicine* **9**, 1121–9.

Fluehler, H., Hirtz, J. and Moser, H. A. (1981) An aid to decision-making in bioequivalence assessment. *Journal of Pharmacokinetics and Biopharmaceutics* **9**, 223–43.

France, L. A., Lewis, J. A. and Kay, R. (1991) The analysis of failure time data in crossover studies. *Statistics in Medicine* **10**, 1099–1161.

Freeman, P. (1989) The performance of the two-stage analysis of two-treatment, two-period cross-over trials. *Statistics in Medicine* **8**, 1421–32.

Freidlin, B. and Gastwirth, J. L. (2000) Should the median test be retired from general use? *American Statistician* **54**, 161– 4.

Gart, J. J. (1969) An exact test for comparing matched proportions in cross-over designs. *Biometrika* **56**, 75–80.

Gehan, E. A. (1965a) A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika* **52**, 650–3.

Gehan, E. A. (1965b) A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203–23.

Gibaldi, M. and Perrier, D. (1982) *Pharmacokinetics*. Marcel Dekker, Basle.

Gibbons, J. D. (1982) Brown–Mood median test, in S. Kotz and N. L. Johnson (eds), *Encyclopedia of Statistics*. Wiley, New York.

Gleiter, C. H., Klotz, U., Kuhlmann, J., Blume, H., Stanislaus, F., Harder, S., Paulus, H., Poethko-Muller, C. and Holz-Slomczyk, M. (1998) When are bioavailability studies required? A German proposal. *Journal of Clinical Pharmacology* **38**, 904–11.

Gough, K. (1993) Book review: Cross-over Trials in Clinical Research. *PSI Newsletter* **15**, 17–19.

Govindarajulu, Z. (2001) *Statistical Techniques in Bioassay*. Karger, Basle.

Graff-Lonnevig, V. and Browaldh, L. (1990) Twelve hours bronchodilating effect of inhaled formoterol in children with asthma: a double-blind cross-over study versus salbutamol. *Clinical and Experimental Allergy* **20**, 429–32.

Grieve, A. P. (1982) The two-period changeover design in clinical trials. *Biometrics* **38**, 517.

Grieve, A. P. (1985) A Bayesian analysis of the two-period crossover design for clinical trials. *Biometrics* **41**, 979–90.

Grieve, A. P. (1987) A note on the analysis of the two-period cross-over design when the period-treatment interaction is significant. *Biometrical Journal* **29**, 771–5.

Grieve, A. P. and Senn, S. J. (1998) Estimating treatment effects in clinical crossover trials. *Journal of Biopharmaceutical Statistics* **8**, 191–233.

Grizzle, J. E. (1965) The two-period change over design and its use in clinical trials. *Biometrics* **21**, 467–80.

Grizzle, J. E. (1974) Correction to Grizzle (1965). *Biometrics* **30**, 727.

Guilbaud, O. (1993) Exact inference about the within-subject variability in 2 × 2 crossover trials. *Journal of the American Statistical Association* **88**, 939– 46.

Guyatt, G. H., Heyting, A. H., Jaeschke, R., Keller, J., Adachi, J. D. and Roberts, R. S. (1990) *N* of 1 trials for investigating new drugs. *Controlled Clinical Trials* **11**, 88–100.

Guyenne, T., Bellet, M., Sassano, P., Serrurier, D., Corvol, P. and Menard, J. (1989) Cross-over design for the dose determination of an angiotensin converting enzyme inhibitor in hypertension. *Journal of Hypertension* **7**, 1005–12.

Hafner, K. B., Koch, G. G. and Canada, A. T. (1988) Some analysis strategies for three-period changeover designs with two treatments. *Statistics in Medicine* **7**, 471–81.

Harding, S., Lane, P. W., Murray, D. and Payne, R. (2000) *GenStat for Windows (5th edition) Introduction*. VSN International, Oxford.

Harrell, F. E. (2001) *Regression Modeling Strategies*. Springer, New York.

Hasselblad, V. and Kong, D. F. (2001) Statistical methods for comparison to placebo in active-control studies. *Drug Information Journal* **35**, 435–49.

Hauck, W. W. and Anderson, S. (1991) Individual bioequivalence—what matters to the patient. *Statistics in Medicine* **10**, 959–60.

Hauck, W. W. and Anderson, S. (1994) Measuring switchability and prescribability: when is average bioequivalence sufficient? *Journal of Pharmacokinetics and Biopharmaceutics* **22**, 551–64.

Hauschke, D., and Steinijans, V. W. (2000) The U.S. draft guidance regarding population and individual bioequivalence approaches: comments by a research-based pharmaceutical company. *Statistics in Medicine* **19**, 2769–74.

Hauschke, D., Steinijans, V. W. and Dlietti, E. (1990) A distribution-free procedure for the statistical analysis of bioequivalence studies. *International Journal of Clinical Pharmacology, Therapy and Toxicology* **28**, 72–8.

Heath, R. (1991) *Guidelines for the Documentation of Data in an Integrated Report*. CIBA-Geigy, Basle.

Hill, A. V. (1910) The possible effect of the aggregation of molecules of haemoglobin on its dissociation curves. *Proceedings of the Physiological Society* **40**, iv–vii.

Hills, M. and Armitage, P. (1979) The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology* **8**, 7–20.

Holford, N. H. G. and Sheiner, L. B. (1981) Understanding the dose effect relationship: clinical application of pharmacokinetic–pharmacodynamic models. *Clinical Pharmacology* **6**, 429–53.

Holmgren, E. B. (1999) Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. *Journal of Biopharmaceutical Statistics* **9**, 651–9.

Hougaard, P. (2000) *Analysis of Multivariate Survival Data*. Springer, New York.

International Conference on Harmonisation (1996) *Guideline for Good Clinical Practice* (E6). International Conference on Harmonisation. http://www.ifpma.org/pdfifpma/e6.pdf.

International Conference on Harmonisation (1999) Statistical principles for clinical trials (ICH E9). *Statistics in Medicine* **18**, 1905–42. See also http://www.ifpma.org/pdfifpma/e9.pdf.

International Conference on Harmonisation (2000) *Choice of Control Group and Related Issues in Clinical Trials* (E10). International Conference on Harmonisation. http://www.ifpma.org/pdfifpma/e10step4.pdf.

Johannessen, T. (1991) Controlled trials in single subjects: 1. Value in clinical medicine. *British Medical Journal* **303**, 173–4.

John, J. A. and Quenouille, M. H. (1977) *Experiments: Design and Analysis*. Griffin, London.

Johnson, A. L. (1983) Clinical trials in psychiatry. *Psychological Medicine* **13**, 1–8.

Johnson, A. L. (1989) Methodology of clinical trials in psychiatry, in G. Freeman and P. Tyrer (eds), *Research Methods in Psychiatry: A Beginner's Guide*. Royal College of Psychiatrists, London.

Jones, B. and Kenward, M. G. (1989) *Design and Analysis of Cross-over Trials*. Chapman & Hall, London.

Jones, B. and Lewis, J. A. (1995) The case for cross-over trials in phase III. *Statistics in Medicine* **14**, 1025–38.

Kenward, M. and Jones, B. (1987a) A log-linear model for binary data. *Applied Statistics* **36**, 192–204.

Kenward, M. and Jones, B. (1987b) The analysis of data from 2 × 2 cross-over trial with baseline measurements. *Statistics in Medicine* **6**, 911–26.

Kenward, M. and Jones, B. (1998) In S. Kotz, C. B. Read, and D. L. Banks (eds), *Encyclopedia of Statistics*, pp. 167–75. Wiley, New York.

Kershner, R. P. and Federer, W. T. (1981) Two-treatment cross-over designs for estimating a variety of effects. *Journal of the American Statistical Association* **76**, 612–19.

Kirkwood, T. B. L. (1981) Bioequivalence testing—a need to rethink. *Biometrics* **37**, 589–91.

Koch, G. G. (1972) The use of non-parametric methods in the statistical analysis of the two-period change-over design. *Biometrics* **28**, 577–84.

Koch, G. G. and Edwards, S. (1988) Clinical efficacy trials with categorical data, in K. G. Peace (ed.), *Biopharmaceutical Statistics for Drug Development*. Marcel Dekker, Basle.

Koch, G. G., Gitomer, S. L., Skalland, L. and Stokes, M. E. (1983) Some non-parametric and categorical data for a change-over design study and discussion of apparent carry-over effects. *Statistics in Medicine* **2**, 397–412.

Koch, G. G., Amara, I. A., Brown, B. W., Colton, T. and Gillings, D. B. (1989) A 2-period crossover design for the comparison of 2 active treatments and placebo. *Statistics in Medicine* **8**, 487–504.

Koch, M. A. (1992) Precision and bias of baseline adjustment estimators in the 2 × 2 cross-over, in *Proceedings of the Biopharmaceutical Section*, pp. 68–73. American Statistical Association, Alexandria, VA.

Krause, A. and Olson, M. (2000) *The Basics of S and S-PLUS*. Springer, New York.

Krzanowski, W. J. (1998) *An Introduction to Statistical Modelling*. Arnold, London.

Kunert, J. (1987a) Nearest neigbour designs for correlated errors. *Biometrics* **74**, 717–24.

Kunert, J. (1987b) On variance estimation in crossover designs. *Biometrics* **43**, 833–45.

Kushner, H. B. (1997) Optimality and efficiency of two-treatment repeated measurements designs. *Biometrika* **84**, 455–68.

Laird, N. M., Skinner, J. and Kenward, M. (1992) An analysis of two-period crossover designs with carry-over effects. *Statistics in Medicine* **11**, 1967–79.

Lambrou, D. (2001) Practical considerations in designing and analysing cross-over clinical trials. PhD thesis, University College London.

Lancaster, H. O. (1961) Significance tests in discrete distributions. *Journal of the American Statistical Association* **56**, 223–34.

Laska, E. and Meisner, M. (1985) A variational approach to optimal two-treatment cross-over designs: applications to carry-over effect models. *Journal of the American Statistical Association* **80**, 704–10.

Laska, E., Meisner, M. and Kushner, H. B. (1983) Optimal crossover designs in the presence of carryover effects. *Biometrics* **39**, 1087–91.

Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models. *Journal of the Royal Statistical Society B* **58**, 619–56.

Lehmacher, W. (1987) *Verlaufskurven und Cross-over*. Springer, Berlin.

Lehmacher, W. (1991) Analysis of cross-over trials in the presence of residual effects. *Statistics in Medicine* **10**, 891–9.

Lehmann, E. L. (1975) *Nonparametrics*. Holden-Day, Oakland, CA.

Leuenberger, P. and Gebre-Michel, I. (1989) Preventative effect of formoterol aerosol in methacholine-induced bronchoconstriction, in R. Davies (ed.), *Formoterol in Asthma—Clinical Profile of a New Long-Acting β-Agonist*, pp. 17–22. Hogrefe and Huber, Toronto.

Lewis, J. A. (1991) Controlled trials in single subjects: 2. Limitations of use. *British Medical Journal* **303**, 175–6.

Lindley, D. V. (1998) Decision analysis and bioequivalence trials. *Statistical Science* **13**, 136–41.

Lindley, D. V. and Scott, W. F. (1984) *New Cambridge Elementary Statistical Tables*. Cambridge University Press, Cambridge.

Lindsey, J. K. (1996) *Parametric Statistical Inference*. Oxford Science Publications, Oxford.

Lindsey, J. K., Jones, B. and Lewis, J. A. (1996) Analysis of cross-over trials for duration data. *Statistics in Medicine* **15**, 527–35.

Lindsey, J. K., Jones, B. and Ebbutt, A. F. (1997) Simple models for repeated ordinal responses with an application to a seasonal rhinitis clinical trial. *Statistics in Medicine* **16**, 2873–82.

Little, R. J. and Rubin, D. B. (1983) Incomplete data, in S. Kotz, and A. L. Johnson (eds), *Encyclopedia of Statistics*. Wiley, New York.

Longford, N. T. (1998) Count data and treatment heterogeneity in 2×2 crossover trials. *Applied Statistics* **47**, 217–29.

Machin, D. and Campbell, M. J. (1987) *Statistical Tables for the Design of Clinical Trials*. Blackwell Scientific Publications, Oxford.

Machin, D., Campbell, M. J., Fayers, P. M. and Pinol, A. P. Y. (1997) *Sample Size Tables for Clinical Studies*. Blackwell Science, Oxford.

Mainland, D. (1963) *Elementary Medical Statistics*. W.B. Saunders, Philadelphia.

Makuch, R. and Johnson, M. (1989) Issues in planning and interpreting active control equivalence studies. *Journal of Clinical Epidemiology* **42**, 503–11.

Mandallaz, D. and Mau, J. (1981) Comparison of different methods for decision-making in bioequivalence assessment. *Biometrics* **37**, 213–22.

Mantel, N. (1963) Chi-square tests with one degree of freedom: extension of the Mantel–Haenszel procedure. *Journal of the American Statistical Association* **58**, 690–700.

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**, 719–48.

Martin, E. (ed.) (1990) *Concise Medical Dictionary*. Oxford University Press, Oxford.

Matthews, J. N. (1987) Optimal cross-over designs for the comparison of two treatments in the presence of carry-over effects and autocorrelated errors. *Biometrika* **74**, 311–20.

Matthews, J. N. (1988) Recent developments in cross-over designs. *International Statistical Review* **56**, 117–27.

Matthews, J. N. (1989) Estimating dispersion parameters in the analysis of data from cross-over trials. *Biometrika* **76**, 239–44.

Matthews, J. N. (1990a) *Cross-over Trials*. Private publication, University of Newcastle upon Tyne.

Matthews, J. N. (1990b) Optimal dual-balanced two-treatment crossover designs. *SankyhāB* **52**, 332–7.

Matthews, J. N. S., Altman, D. G., Campbell, M. J. and Royston, P. (1990) Analysis of serial measurements in medical research. *British Medical Journal* **300**, 230–5.

McConway, K. J., Jones, M. C. and Taylor, P. C. (1999) *Statistical Modelling Using GENSTAT*. Arnold, London.

McCullagh, P. (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society B* **42**, 109–42.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. Chapman & Hall, London.

McGilveray, I. J. (1991) Bioequivalence: a Canadian regulatory perspective, in P. G. Welling, F. L. S. Tse and S. V. Dighe (eds)., *Pharmaceutical Bioequivalence*. Marcel Dekker, New York.

McGilveray, I. J. (1992) Bioavailability testing of medicinal products and harmonization of international testing requirements and standards: the Canadian perspective. *Drug Information Journal* **26**, 365–9.

McNeill, D. (1977) *Interactive Data Analysis*. Wiley, New York.

McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–7.

Mead, R. (1988) *The Design of Experiments*. Cambridge University Press, Cambridge.

Mead, R. (1990) The non-orthogonal design of experiments. *Journal of the Royal Statistical Society A* **153**, 151–201.

Mehring, G. (1993) On optimal tests for general interval hypotheses. *Communications in Statistics: Theory and Methods* **22**, 1257–97.

Mehta, C. R. and Patel, N. (2000) *StatXact 4 for Windows*. Cytel Software Corporation, Cambridge, MA.

Metzler, C. M. (1991) Statistical criteria, in P. G. Welling, F. L. S. Tse and S. V. Dighe (eds), *Pharmaceutical Bioequivalence*. Marcel Dekker, New York.

Millar, K. (1983) Clinical trial design: the neglected problem of asymmetrical transfer in cross-over trials. *Psychological Medicine* **13**, 867–73.

Morrey, G. H. (1989) Binary response and the three-period cross-over. *Biometrical Journal* **31**, 589–98.

National Asthma Educational Program (1991) *Executive Summary: Guidelines for the Diagnosis and Management of Asthma*. US Department of Health and Human Services, National Institutes of Health, Bethesda, MD.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society A* **132**, 107–20.

Newcombe, R. G. (1992) Latin square designs for crossover studies balanced for carry-over effects [letter]. *Statistics in Medicine* **11**, 560.

Newcombe, R. G. (1996) Sequentially balanced three-squares cross-over designs. *Statistics in Medicine* **15**, 2143–7.

North, P. M. (1998) Ensuring good statistical practice in clinical research: guidelines for standard operating procedures (an update). *Drug Information Journal* **32**, 665–82.

O'Hagan, A. (1994) *Kendall's Advanced Theory of Statistics: Bayesian Inference*. Arnold, London.

O'Quigley, J. and Baudoin, C. (1988) General approaches to the problem of bioequivalence. *The Statistician* **37**, 51–8.

Parascandola, J. (1975) Arthur Cushny, optical isomerism and the mechanism of drug action. *Journal of the History of Biology* **8**, 145–65.

Patefield, M. (2000) Conditional and exact tests in crossover trials. *Journal of Biopharmaceutical Statistics* **10**, 109–29.

Patel, H. I. (1983) The use of baselines in the two-period cross-over design. *Communications in Statistics A* **12**, 2693–712.

Patel, H. I. (1986) Analysis of incomplete data in a two-period cross-over design with reference to clinical trials. *Biometrika* **72**, 411–18.

Patterson, S. D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–54.

Phillips, A., Ebbutt, A., France, L. and Morgan, D. (2000) The International Conference on Harmonisation guideline 'Statistical Principles for Clinical Trials': issues in applying the guideline in practice. *Drug Information Journal* **34**, 337–48.

Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

Plackett, R. and Barnard, G. (eds) (1990) *'Student': A Statistical Biography of William Sealy Gosset by E.S. Pearson*. Clarendon Press, Oxford.

Pledger, G. (1989) The role of a placebo-treated control group in combination drug trials. *Controlled Clinical Trials* **10**, 97–107.

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

Pocock, S., Altman, D., Armitage, P., Ashby, D., Bland, M., Chilvers, C., Dawid, P., Ebbutt, A., Evans, S., Finney, D., Gardner, M., Gore, S., Jones, D., Lewis, J., Machin, D., Matthews, J., Spiegelhalter, D., Sutherland, I. and Thompson, S. (1991) Statistics and statisticians in drug regulation in the United Kingdom. *Journal of the Royal Statistical Society A* **154**, 413–19.

Preece, D. A. (1981) Distribution of final digits in data. *The Statistician* **30**, 31–60.

Preece, D. A. (1982) T is for trouble (and textbooks): a critique of some examples of the paired-samples *t*-test. *The Statistician* **31**, 169–95.

Prescott, P. (1999) Construction of sequentially counterbalanced designs formed from two Latin squares. *Utilitas Mathematica* **55**, 135–52.

Prescott, R. J. (1981) The comparison of success rates in cross-over trials in the presence of an order effect. *Applied Statistics* **30**, 9–15.

PSI (1991) *PSI Guidelines for Standard Operating Procedures: Statistical Analysis Plan*. Statisticians in the Pharmaceutical Industry (PSI).

PSI Professional Standards Working Party (1994) Good statistical practice in clinical research: guideline standard operating procedures. *Drug Information Journal* **28**, 615–27.

Putt, M. and Chinchilli, V. M. (1999) A mixed effects model for the analysis of repeated measures cross-over studies. *Statistics in Medicine* **18**, 3037–58.

Putt, M. E. and Chinchilli, V. M. (2000) A robust analysis of crossover designs using multisample generalized L-statistics. *Journal of the American Statistical Association* **95**, 1256–62.

Racine-Poon, A., Grieve, A. P., Fluhler, H. and Smith, A. F. (1987) A two-stage procedure for bioequivalence studies. *Biometrics* **43**, 847–56.

Ratkowsky, D. A., Evans, M. A. and Alldredge, J. R. (1993) *Cross-over Experiments: Design, Analysis, and Application*. Marcel Dekker, New York.

Rauws, A. G. (1991) Bioequivalence: a European Community regulatory perspective, in P. G. Welling, F. L. S. Tse and S. V. Dighe (eds), *Pharmaceutical Equivalence*. Marcel Dekker, New York.

Richardson, W. and Bablok, B. (1992) Clinical experience with formoterol in adults, in S. T. Holgate (ed.), *Formoterol: Fast and Long-Lasting Bronchodilation*. Royal Society of Medicine Services, London.

Rowland, M. and Towzer, T. N. (1995) *Clinical Pharmacokinetics: Concepts and Applications*. Williams and Wilkins, Baltimore, MD.

Royall, R. M. (1986) The effect of sample size on the meaning of significance tests. *American Statistician* **40**, 313–15.

Salmonson, T., Melander, H. and Rane, A. (1991) Bioequivalence: a Nordic regulatory perspective, in P. G. Welling, F. L. S. Tse and S. V. Dighe (eds), *Pharmaceutical Bioequivalence*. Marcel Dekker, New York.

Selwyn, M. R., Dempster, A. P. and Hall, N. R. (1981) A Bayesian approach to bioequivalence for the 2 × 2 changeover design. *Biometrics* **37**, 11–21.

Senn, S. J. (1988) Cross-over trials, carry-over effects and the art of self-delusion. *Statistics in Medicine* **7**, 1099–1101.

Senn, S. J. (1989a) The use of baselines in clinical trials of bronchodilators. *Statistics in Medicine* **8**, 1339–50.

Senn, S. J. (1989b) Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* **8**, 467–75.

Senn, S. J. (1991) Problems with the two stage analysis of crossover trials. *British Journal of Clinical Pharmacology* **32**, 133.

Senn, S. J. (1992) Is the 'simple carry-over' model useful? *Statistics in Medicine* **11**, 715–26. Erratum (1992) *Statistics in Medicine*, **11**, 1619.

Senn, S. J. (1993a) Inherent difficulties with active control equivalence studies. *Statistics in Medicine* **12**, 2367–75.

Senn, S. J. (1993b) A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine* **12**, 2147–51.

Senn, S. J. (1993c) Statistical issues in short term trials in asthma. *Drug Information Journal* **27**, 779–91.

Senn, S. J. (1993d) Suspended judgement: *N* of 1 trials. *Controlled Clinical Trials* **14**, 1–5.

Senn, S. J. (1994) The *AB/BA* crossover: past, present and future? *Statistical Methods in Medical Research* **3**, 303–24.

Senn, S. J. (1995a) Cross-over trials at the cross-roads? *Applied Clinical Trials* **4**, 24–31.

Senn, S. J. (1995b) Some controversies in designing and analysing cross-over trials. *Biocybernetics and Biomedical Engineering* **15**, 27–39.

Senn, S. J. (1996) The *AB/BA* cross-over: how to perform the two-stage analysis if you can't be persuaded that you shouldn't, in B. Hansen and M. de Ridder (eds)., *Liber Amicorum Roel van Strik*, pp. 93–100. Erasmus University, Rotterdam. Privately printed.

Senn, S. J. (1997a) The case for cross-over trials in phase III. *Statistics in Medicine* **16**, 2021–2.

Senn, S. J. (1997b) *Statistical Issues in Drug Development*. Wiley, Chichester.

Senn, S. J. (1997c) Statisticians and pharmacokineticists: what can they learn from each other? in L. Aarons, L. P. Balant, M. Danhof, M. Gex-Fabry, U. A. Gundert-Remy, M. O. Karlsson, F. Mentré, P. L. Morselli, F. Rombout, M. Rowland, J.-L. Steimer and S. Vozeh (eds), *COST B1 Medicine. The Population Approach: Measuring and Managing Variability in Response, Concentration and Dose*, pp. 139–46. European Commission Directorate-General Science, Research and Devolopment, Geneva.

Senn, S. J. (1998a) Cross-over trials, in P. Armitage and T. Colton (eds), *Encyclopedia of Biostatistics*, pp. 1033–49. Wiley, Chichester.

Senn, S. J. (1998b) In the blood: proposed new requirements for registering generic drugs. *Lancet* **352**, 85–6.

Senn, S. J. (2000a) Consensus and controversy in pharmaceutical statistics (with discussion). *The Statistician* **49**, 135–76.

Senn, S. J. (2000b) Crossover design, in S. C. Chow and J. P. Liu (eds), *Encyclopedia of Biopharmaceutical Statistics*, pp. 142–9. Marcel Dekker, New York.

Senn, S. J. (2000c) The many modes of meta. *Drug Information Journal* **34**, 535–49.

Senn, S. J. (2000d) Statistical quality in analysing clinical trials. *Good Clinical Practice Journal* **7**, 22–6.

Senn, S. J. (2001a) Cross-over trials in drug development: theory and practice. *Journal of Statistical Planning and Inference* **96**, 29–40.

Senn, S. J. (2001b) Individual therapy: new dawn or false dawn. *Drug Information Journal* **35**, 1479–94.

Senn, S. J. (2001c) Statistical issues in bioequivalance. *Statistics in Medicine* **20**, 2785–99.

Senn, S. J. (2001d) Within patient studies: Cross-over trials and *n*-of-1 studies, in M. B. Max and J. Lynn (eds), *Interactive Textbook on Clinical Symptom Research*. National Institute of Dental and Craniofacial Research, Bethesda, MD. http://www.symptomresearch.org/chapter_6/index.htm.

Senn, S. J. and Auclair, P. (1990) The graphical representation of clinical trials with particular reference to measurements over time. *Statistics in Medicine* **9**, 1287–302. Erratum (1991) *Statistics in Medicine* **10**, 487.

Senn, S. J. and Hildebrand, H. (1991) Crossover trials, degrees of freedom, the carryover problem and its dual. *Statistics in Medicine* **10**, 1361–74.

Senn, S. J. and Lambrou, D. (1998) Robust and realistic approaches to carry-over. *Statistics in Medicine* **17**, 2849–64.

Senn, S. J. and Richardson, W. (1994) The first *t*-test. *Statistics in Medicine* **13**, 785–803.

Senn, S. J., Lillienthal, J., Patalano, F. and Till, M. D. (1997) An incomplete blocks cross-over in asthma: a case study in collaboration, in J. Vollmar and L. A. Hothorn (eds), *Cross-over Clinical Trials*, pp. 3–26. Fischer, Stuttgart.

Senn, S. J., Stevens, L. and Chaturvedi, N. (2000) Repeated measures in clinical trials: simple strategies for analysis using summary measures. *Statistics in Medicine* **19**, 861–77.

Sheehe, P. R. and Bross, D. J. (1961) Latin squares to balance immediate residual, and other order, effects. *Biometrics* **17**, 405–14.

Sheiner, L. B., Hasimoto, Y. and Beal, S. L. (1991) A simulation study comparing designs for dose-ranging. *Statistics in Medicine* **9**, 1287–302.

Shumaker, R. C. and Metzler, C. M. (1998) The phenytoin trial is a case study of 'individual bioequivalence'. *Drug Information Journal* **32**, 1063–72.

Somes, G. W. and O'Brien, K. F. (1985) Mantel–Haenszel statistic, in S. Kotz and N. L. Johnson (eds), *Encyclopedia of Statistics*. Wiley, New York.

Spiegelhalter, D. J. (1988) Statistical issues in studies of individual response. *Scandinanvian Journal of Gastroenterology Suppl.* **147**, 40–5.

Sprent, P. and Smeeton, N. C. (2001) *Applied Nonparametric Statistical Methods*. Chapman & Hall/CRC, Boca Raton, FL.

Sprott, D. A. and Farewell, V. T. (1993) The difference between 2 normal means. *American Statistician* **47**, 126–8.

Steinijans, V. W. and Diletti, E. (1983) Statistical analysis of bioavailability studies: parametric and nonparametric confidence intervals. *European Journal of Clinical Pharmacology* **24**, 127–36.

Steinijans, V. W. and Hauschke, D. (1997) Individual bioequivalence—a European perspective. *Journal of Biopharmaceutical Statistics* **7**, 31–4.

Student (1908) The probable error of a mean. *Biometrika* **6**, 1–25.

Temple, R. (1982) Government view of clinical trials. *Drug Information Journal* **16**, 10–17.

Tsoy, A. N., Cheltzov, O. V., Zaseyeva, V., Shilinish, L. A. and Yashina, L. A. (1990) *European Respiratory Journal* **3**, 235s.

Tufte, E. R. (1983) *The Visual Display of Quantitative Information*. Graphics Press, Cheshire. CT.

Tukey, J. W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Venables, W. N. and Ripley, B. (1999) *Modern Applied Statistics with S-PLUS*. Springer, New York.

Wallenstein, S. and Fisher, A. C. (1977) The analysis of the two-period repeated measurements crossover design with application to clinical trials. *Biometrics* **33**, 261–9.

Walters, S. and Hall, R. C. (1991) Bioequivalence: an Australian regulatory perspective, in P. G. Welling, F. L. S. Tse and S. V. Dighe (eds), *Pharmaceutical Bioequivalence*. Marcel Dekker, New York.

Wang, S. J. and Hung, H. M. (1997) Use of two-stage test statistic in the two-period crossover trials. *Biometrics* **53**, 1081–91.

Welling, P. G., Tse, F. L. S. and Dighe, S. V. (eds) (1991) *Pharmaceutical Bioequivalence*. Marcel Dekker, New York.

Westlake, W. J. (1976) Symmetrical confidence intervals for bioequivalence trials. *Biometrics* **32**, 741–4.

Westlake, W. J. (1981) Response to Kirkwood. *Biometrics* **37**, 591–3.

Wilding, P., Clark, M., Coon, J. T., Lewis, S., Rushton, L., Bennett, J., Oborne, J., Cooper, S. and Tattersfield, A. E. (1997) Effect of long-term treatment with salmeterol on asthma control: a double blind, randomised crossover study. *British Medical Journal* **314**, 1441–6.

Willan, A. R. (1988) Using the maximum test statistic in the two-period crossover clinical trial. *Biometrics* **44**, 211–18.

Willan, A. R. and Pater, J. L. (1986a) Carryover and the two-period crossover clinical trial. *Biometrics* **42**, 593–9.

Willan, A. R. and Pater, J. L. (1986b) Using baseline measurements in the two-period crossover clinical trial. *Control Clinical Trials* **7**, 282–9.

Williams, E. J. (1949) Experimental designs balanced for the estimation of residual effects. *Australian Journal of Scientific Research* **2**, 149–68.

Yates, F. (1984) Tests of significance for 2×2 contigency tables. *Journal of the Royal Statistical Society A* **147**, 426–63.

Youden, J. (1972) Enduring values. *Technometrics* **14**, 1–11.

Zimmerman, H. and Ralfs, V. W. (1980) Model building and testing for the change-over design. *Biometrical Journal* **22**, 197–210.

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

Author Index

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

Subject Index

Cross-over Trials in Clinical Research, Second Edition. Stephen Senn
Copyright © 2002 John Wiley & Sons, Ltd.
Print ISBN: 0-471-49653-7

Statistics in Practice

Human and Biological Sciences

Brown and Prescott – Applied Mixed Models in Medicine
Ellenberg, Fleming and DeMets – Data Monitoring in Clinical Trials: A Practical Perspective
Marubini and Valsecchi – Analysing Survival Data from Clinical Trials and Observation Studies
Parmigiani – Modeling in Medical Decision Making: A Bayesian Approach
Senn – Cross-over Trials in Clinical Research
Senn – Statistical Issues in Drug Development
A. Whitehead – Meta-analysis of Controlled Clinical Trials
J. Whitehead – The Design and Analysis of Sequential Clinical Trials, Revised Second Edition

Earth and Environmental Sciences

Buck, Cavanagh and Litton – Bayesian Approach to Interpreting Archaeological Data
Webster and Oliver – Geostatistics for Environmental Scientists

Industry, Commerce and Finance

Aitken – Statistics and the Evaluation of Evidence for Forensic Scientists
Lehtonen and Pahkinen – Practical Methods for Design and Analysis of Complex Surveys
Ohser and Mücklich – Statistical Analysis of Microstructures in Materials Science