Angela Dean
Daniel Voss
Danel Draguljić

Design and Analysis of Experiments

Second Edition

Springer

Springer Texts in Statistics

Series editors

R. DeVeaux
S.E. Fienberg
I. Olkin

Angela Dean · Daniel Voss
Danel Draguljić

Design and Analysis of Experiments

Second Edition

Springer

Angela Dean
The Ohio State University
Columbus, OH
USA

Daniel Voss
Wright State University
Dayton, OH
USA

Danel Draguljić
Franklin & Marshall College
Lancaster, PA
USA

Preface to the Second Edition

Since writing the first edition of *Design and Analysis of Experiments*, there have been a number of additions to the research investigator's toolbox. In this second edition, we have incorporated a few of these modern topics.

Small screening designs are now becoming prevalent in industry for aiding the search for a few influential factors from amongst a large pool of factors of potential interest. In Chap. 15, we have expanded the material on saturated designs and introduced the topic of *supersaturated designs* which have fewer observations than the number of factors being investigated. We have illustrated that useful information can be gleaned about influential factors through the use of supersaturated designs even though their contrast estimators are correlated. When curvature is of interest, we have described *definitive screening designs* which have only recently been introduced in the literature, and which allow second order effects to be measured while retaining independence of linear main effects and requiring barely more than twice as many observations as factors.

Another modern set of tools, now used widely in areas such as biomedical and materials engineering, the physical sciences, and the life sciences, is that of *computer experiments*. To give a flavor of this topic, a new Chap. 20 has been added. Computer experiments are typically used when a mathematical description of a physical process is available, but a physical experiment cannot be run for ethical or cost reasons. We have discussed the major issues in both the design and analysis of computer experiments. While the complete treatment of the theoretical background for the analysis is beyond the scope of this book, we have provided enough technical details of the statistical model, as well as an intuitive explanation, to make the analysis accessible to the intended reader. We have also provided computer code needed for both design and analysis.

Chapter 19 has been expanded to include two new experiments involving split-plot designs from the discipline of human factors engineering. In one case, imbalance due to lost data, coupled with a mixed model, motivates introduction of restricted-maximum-likelihood-based methods implemented in the computer software sections, including a comparison of these methods to those based on least squares estimation.

It is now the case that analysis of variance and computation of confidence intervals is almost exclusively done by computer and rarely by hand. However, we have retained the basic material on these topics since it is

fundamental to the understanding of computer output. We have removed some of the more specialized details of least squares estimates from Chaps. 10–12 and canonical analysis details in Chap. 16, relying on the computer software sections to illustrate these.

SAS® software is still used widely in industry, but many university departments now teach the analysis of data using R (R Development Core Team, 2017). This is a command line software for statistical computing and graphics that is freely available on the web. Consequently, we have made a major addition to the book by including sections illustrating the use of R software for each chapter. These sections run parallel to the "Using SAS Software" sections, retained from the first edition.

A few additions have been made to the "Using SAS Software" sections. For example, in Chap. 11, PROC OPTEX has been included for generation of efficient block designs. PROC MIXED is utilized in Chap. 5 to implement Satterthwaite's method, and also in Chaps. 17–19 to estimate standard errors involving composite variance estimates, and in Chap. 19 to implement *restricted maximum likelihood estimation* given imbalanced data and mixed models.

We have updated the SAS output[1], showing this as reproductions of PC output windows generated by each program. The SAS programs presented can be run on a PC or in a command line environment such as unix, although the latter would use PROC PLOT rather than the graphics PROC SGPLOT.

Some minor modifications have been made to a few other chapters from the first edition. For example, for assessing which contrasts are non-negligible in single replicate or fractional factorial experiments, we have replaced normal probability plots by *half-normal probability plots* (Chaps. 7, 13 and 15). The reason for this change is that contrast signs are dependent upon which level of the factor is labeled as the high level and which is labeled as the low level. Half-normal plots remove this potential arbitrariness by plotting the absolute values of the contrast estimates against "half-normal scores".

Section 7.6 in the first edition on the control of noise variability and Taguchi experiments has been removed, while the corresponding material in Chap. 15 has been expanded. On teaching the material, we found it preferable to have information on mixed arrays, product arrays, and their analysis, in one location. The selection of multiple comparison methods in Chap. 4 has been shortened to include only those methods that were used constantly throughout the book. Thus, we removed the method of multiple comparisons with the best, which was not illustrated often; however, this method remains appropriate and valid for many situations in practice.

Some of the worked examples in Chap. 10 have been replaced with newer experiments, and new worked examples added to Chaps. 15 and 19. Some new exercises have been added to many chapters. These either replace

---

[1]The output in our "Using SAS Software" sections was generated using SAS software Version 9.3 of the SAS System for PC. Copyright © SAS 2012 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

exercises from the first edition or have been added at the end of the exercise list. All other first edition exercises retain their same numbers in this second edition.

A new website http://www.wright.edu/~dan.voss/DeanVossDraguljic.html has been set up for the second edition. This contains material similar to that on the website for the first edition, including datasets for examples and exercises, SAS and R programs, and any corrections.

We continue to owe a debt of gratitude to many. We extend our thanks to all the many students at The Ohio State University and Wright State University who provided imaginative and interesting experiments and gave us permission to include their projects. We thank all the readers who notified us of errors in the first edition and we hope that we have remembered to include all the corrections. We will be equally grateful to readers of the second edition for notifying us of any newly introduced errors. We are indebted to Russell Lenth for updating the R package lsmeans to encompass all the multiple comparisons procedures used in this book. We are grateful to the editorial staff at Springer, especially Rebekah McClure and Hannah Bracken, who were always available to give advice and answer our questions quickly and in detail.

Finally, we extend our love and gratitude to Jeff, Nancy, Tom, Jimmy, Linda, Luka, Nikola, Marija and Anika.

Columbus, USA   Angela Dean
Dayton, USA   Daniel Voss
Lancaster, USA   Danel Draguljić

Preface to the First Edition

The initial motivation for writing this book was the observation from various students that the subject of design and analysis of experiments can seem like "a bunch of miscellaneous topics." We believe that the identification of the objectives of the experiment and the practical considerations governing the design form the heart of the subject matter and serve as the link between the various analytical techniques. We also believe that learning about design and analysis of experiments is best achieved by the planning, running, and analyzing of a simple experiment.

With these considerations in mind, we have included throughout the book the details of the planning stage of several experiments that were run in the course of teaching our classes. The experiments were run by students in statistics and the applied sciences and are sufficiently simple that it is possible to discuss the planning of the entire experiment in a few pages, and the procedures can be reproduced by readers of the book. In each of these experiments, we had access to the investigators' actual report, including the difficulties they came across and how they decided on the treatment factors, the needed number of observations, and the layout of the design. In the later chapters, we have included details of a number of published experiments. The outlines of many other student and published experiments appear as exercises at the ends of the chapters.

Complementing the practical aspects of the design are the statistical aspects of the analysis. We have developed the theory of estimable functions and analysis of variance with some care, but at a low mathematical level. Formulae are provided for almost all analyses so that the statistical methods can be well understood, related design issues can be discussed, and computations can be done by hand in order to check computer output.

We recommend the use of a sophisticated statistical package in conjunction with the book. Use of software helps to focus attention on the statistical issues rather than the calculation. Our particular preference is for the SAS software, and we have included the elementary use of this package at the end of most chapters. Many of the SAS program files and data sets used in the book can be found at www.springer–ny.com. However, the book can equally well be used with any other statistical package. Availability of statistical software has also helped shape the book in that we can discuss more complicated analyses—the analysis of unbalanced designs, for example.

The level of presentation of material is intended to make the book accessible to a wide audience. Standard linear models under normality are used for all analyses. We have avoided using calculus, except in a few optional sections where least squares estimators are obtained. We have also avoided using linear algebra, except in an optional section on the canonical analysis of second-order response surface designs. Contrast coefficients are listed in the form of a vector, but these are interpreted merely as a list of coefficients.

This book reflects a number of personal preferences. First and foremost, we have not put side conditions on the parameters in our models. The reason for this is threefold. Firstly, when side conditions are added to the model, all the parameters appear to be estimable. Consequently, one loses the perspective that in factorial experiments, main effects can be interpreted only as averages over any interactions that happen to be present. Secondly, the side conditions that are the most useful for hand calculation do not coincide with those used by the SAS software. Thirdly, if one feeds a nonestimable parametric function into a computer program such as `PROC GLM` in SAS, the program will declare the function to be "nonestimable," and the user needs to be able to interpret this statement. A consequence is that the traditional solutions to the normal equations do not arise naturally. Since the traditional solutions are for nonestimable parameters, we have tried to avoid giving these, and instead have focused on the estimation of functions of $E[Y]$, all of which are estimable.

We have concentrated on the use of prespecified models and preplanned analyses rather than exploratory data analysis. We have emphasized the experimentwise control of error rates and confidence levels rather than individual error rates and confidence levels.

We rely upon residual plots rather than formal tests to assess model assumptions. This is because of the additional information provided by residual plots when model assumption violations are indicated. For example, plots to check homogeneity of variance also indicate when a variance-stabilizing transformation should be effective. Likewise, nonlinear patterns in a normal probability plot may indicate whether inferences under normality are likely to be liberal or conservative. Except for some tests for lack of fit, we have, in fact, omitted all details of formal testing for model assumptions, even though they are readily available in many computer packages.

The book starts with basic principles and techniques of experimental design and analysis of experiments. It provides a checklist for the planning of experiments, and covers analysis of variance, inferences for treatment contrasts, regression, and analysis of covariance. These basics are then applied in a wide variety of settings. Designs covered include completely randomized designs, complete and incomplete block designs, row-column designs, single replicate designs with confounding, fractional factorial designs, response surface designs, and designs involving nested factors and factors with random effects, including split-plot designs.

In the last few years, "Taguchi methods" have become very popular for industrial experimentation, and we have incorporated some of these ideas. Rather than separating Taguchi methods as special topics, we have interspersed

them throughout the chapters via the notion of including "noise factors" in an experiment and analyzing the variability of the response as the noise factors vary.

We have introduced factorial experiments as early as Chapter 3, but analyzed them as one-way layouts (i.e., using a cell means model). The purpose is to avoid introducing factorial experiments halfway through the book as a totally new topic, and to emphasize that many factorial experiments are run as completely randomized designs. We have analyzed contrasts in a two-factor experiment both via the usual two-way analysis of variance model (where the contrasts are in terms of the main effect and interaction parameters) and also via a cell-means model (where the contrasts are in terms of the treatment combination parameters). The purpose of this is to lay the groundwork for Chapters 13–15, where these contrasts are used in confounding and fractions. It is also the traditional notation used in conjunction with Taguchi methods.

The book is not all-inclusive. For example, we do not cover recovery of interblock information for incomplete block designs with random block effects. We do not provide extensive tables of incomplete block designs. Also, careful coverage of unbalanced models involving random effects is beyond our scope. Finally, inclusion of SAS graphics is limited to low-resolution plots.

The book has been classroom tested successfully over the past five years at The Ohio State University, Wright State University, and Kenyon College, for junior and senior undergraduate students majoring in a variety of fields, first-year graduate students in statistics, and senior graduate students in the applied sciences. These three institutions are somewhat different. The Ohio State University is a large land-grant university offering degrees through the Ph.D., Wright State University is a mid-sized university with few Ph.D. programs, and Kenyon College is a liberal arts undergraduate college. Below we describe typical syllabi that have been used.

At OSU, classes meet for five hours per week for ten weeks. A typical class is composed of 35 students, about a third of whom are graduate students in the applied statistics master's program. The remaining students are undergraduates in the mathematical sciences or graduate students in industrial engineering, biomedical engineering, and various applied sciences. The somewhat ambitious syllabus covers Chapters 1–7 and 10, Sections 11.1–11.4, and Chapters 13, 15, and 17. Students taking these classes plan, run, and analyze their own experiments, usually in a team of four or five students from several different departments. This project serves the function of giving statisticians the opportunity of working with scientists and of seeing the experimental procedure firsthand, and gives the scientists access to colleagues with a broader statistical training. The experience is usually highly rated by the student participants.

Classes at WSU meet four hours per week for ten weeks. A typical class involves about 10 students who are either in the applied statistics master's degree program or who are undergraduates majoring in mathematics with a statistics concentration. Originally, two quarters (20 weeks) of probability and statistics formed the prerequisite, and the course covered much of Chapters 1–4, 6, 7, 10, 11, and 13, with Chapters 3 and 4 being primarily

review material. Currently, students enter with two additional quarters in applied linear models, including regression, analysis of variance, and methods of multiple comparisons, and the course covers Chapters 1 and 2, Sections 3.2, 6.7, and 7.5, Chapters 10, 11, and 13, Sections 15.1–15.2, and perhaps Chapter 16. As at OSU, both of these syllabi are ambitious. During the second half of the course, the students plan, run, and analyze their own experiments, working in groups of one to three. The students provide written and oral reports on the projects, and the discussions during the oral reports are of mutual enjoyment and benefit. A leisurely topics course has also been offered as a sequel, covering the rest of Chapters 14–17.

At Kenyon College, classes meet for three hours a week for 15 weeks. A typical class is composed of about 10 junior and senior undergraduates majoring in various fields. The syllabus covers Chapters 1–7, 10, and 17.

For some areas of application, random effects, nested models, and split-plot designs, which are covered in Chapters 17–19, are important topics. It is possible to design a syllabus that reaches these chapters fairly rapidly by covering Chapters 1–4, 6, 7, 17, 18, 10, 19.

This book has been ten years in the making. In the view of the authors, it is "a work in progress temporarily cast in stone"—or in print, as it were. We are wholly responsible for any errors and omissions, and we would be most grateful for comments, corrections, and suggestions from readers so that we can improve any future editions.

Finally, we extend our love and gratitude to Jeff, Nancy, Tommy, and Jimmy, often neglected during this endeavor, for their enduring patience, love, and support.

Columbus, Ohio   Angela Dean
Dayton, Ohio   Daniel Voss

Contents

Principles and Techniques

<div style="text-align:right">**1**</div>

1.1 Design: Basic Principles and Techniques

1.1.1 The Art of Experimentation

One of the first questions facing an experimenter is, "How many observations do I need to take?" or alternatively, "Given my limited budget, how can I gain as much information as possible?" These are not questions that can be answered in a couple of sentences. They are, however, questions that are central to the material in this book. As a first step towards obtaining an answer, the experimenter must ask further questions, such as, "What is the main purpose of running this experiment?" and "What do I hope to be able to show?"

Typically, an experiment may be run for one or more of the following reasons:

  (i) to determine the principal causes of variation in a measured response,
 (ii) to find the conditions that give rise to a maximum or minimum response,
(iii) to compare the responses achieved at different settings of controllable variables,
(iv) to obtain a mathematical model in order to predict future responses.

Observations can be collected from *observational studies* as well as from *experiments*, but only an experiment allows conclusions to be drawn about cause and effect. For example, consider the following situation:

The output from each machine on a factory floor is constantly monitored by any successful manufacturing company. Suppose that in a particular factory, the output from a particular machine is consistently of low quality. What should the managers do? They could conclude that the machine needs replacing and pay out a large sum of money for a new one. They could decide that the machine operator is at fault and dismiss him or her. They could conclude that the humidity in that part of the factory is too high and install a new air conditioning system. In other words, the machine output has been observed under the current operating conditions (an observational study), and although it has been very effective in showing the management that a problem exists, it has given them very little idea about the *cause* of the poor quality.

It would actually be a simple matter to determine or rule out some of the potential causes. For example, the question about the operator could be answered by moving all the operators from machine to machine over several days. If the poor output follows the operator, then it is safe to conclude that the operator is the cause. If the poor output remains with the original machine, then the operator is

blameless, and the machine itself or the factory humidity is the most likely cause of the poor quality. This is an "experiment." The experimenter has control over a possible cause in the difference in output quality between machines. If this particular cause is ruled out, then the experimenter can begin to vary other factors such as humidity or machine settings.

It is more efficient to examine all possible causes of variation simultaneously rather than one at a time. Fewer observations are usually needed, and one gains more information about the system being investigated. This simultaneous study is known as a "factorial experiment." In the early stages of a project, a list of all factors that conceivably could have an important effect on the response of interest is drawn up. This may yield a large number of factors to be studied, in which case special techniques are needed to gain as much information as possible from examining only a subset of possible factor settings.

The art of designing an experiment and the art of analyzing an experiment are closely intertwined and need to be studied side by side. In designing an experiment, one must take into account the analysis that will be performed. In turn, the efficiency of the analysis will depend upon the particular experimental design that is used to collect the data. Without these considerations, it is possible to invest much time, effort, and expense in the collection of data which seem relevant to the purpose at hand but which, in fact, contribute little to the research questions being asked. A guiding principle of experimental design is to "keep it simple." Interpretation and presentation of the results of experiments are generally clearer for simpler experiments.

Three basic techniques fundamental to experimental design are replication, blocking, and randomization. The first two help to increase precision in the experiment; the last is used to decrease bias. These techniques are discussed briefly below and in more detail throughout the book.

1.1.2   Replication

Replication is the repetition of experimental conditions so that the effects of interest can be estimated with greater precision and the associated variability can be estimated.

There is a difference between "replication" and "repeated measurements." For example, suppose four subjects are each assigned to a drug and a measurement is taken on each subject. The result is four independent observations on the drug. This is "replication." On the other hand, if one subject is assigned to a drug and then measured four times, the measurements are not independent. We call them "repeated measurements." The variation recorded in repeated measurements taken at the same time reflects the variation in the measurement process, while the variation recorded in repeated measurements taken over a time interval reflects the variation in the single subject's response to the drug over time. Neither reflects the variation in independent subjects' responses to the drug. We need to know about the latter variation in order to generalize any conclusion about the drug so that it is relevant to all similar subjects.

1.1.3   Blocking

A designed experiment involves the application of treatments to experimental units to assess the effects of the treatments on some response. The "experimental units," which may be subjects, materials, conditions, points in time, or some combination of these, will be variable and induce variation in the response. Such variation in experimental units may be intentional, as the experimental conditions under which an experiment is run should be representative of those to which the conclusions of the experiment are to be applied. For inferences to be broad in scope, the experimental conditions should

be appropriately varied. Blocking is a technique that can often be used to control and adjust for some of the variation in experimental units.

To block an experiment is to divide, or partition, the experimental units into groups called blocks in such a way that the experimental units in each block are intended to be relatively similar, so that treatments assigned to experimental units in the same block can be compared under relatively similar experimental conditions. If blocking is done well, then comparisons of two or more treatments are made more precisely in the experiment than similar comparisons from an unblocked design. For example, in an experiment to compare the effects of two skin ointments for rash, the two treatments can be compared more precisely on two arms of the same person than on the arms of two different people. Either circumstance can be replicated, ideally using subjects randomly sampled from or at least representative of the population of interest.

1.1.4   Randomization

The purpose of randomization is to prevent systematic and personal biases from being introduced into the experiment by the experimenter. A random assignment of subjects or experimental material to treatments prior to the start of the experiment ensures that observations that are favored or adversely affected by unknown sources of variation are observations "selected in the luck of the draw" and not systematically selected.

Lack of a random assignment of experimental material or subjects leaves the experimental procedure open to *experimenter bias*. For example, a horticulturist may assign his or her favorite variety of experimental crop to the parts of the field that look the most fertile, or a medical practitioner may assign his or her preferred drug to the patients most likely to respond well. The preferred variety or drug may then appear to give better results no matter how good or bad it actually is.

Lack of random assignment can also leave the procedure open to *systematic bias*. Consider, for example, an experiment involving drying time of three paints applied to sections of a wooden board, where each paint is to be observed four times. If no random assignment of order of observation is made, many experimenters would take the four observations on paint 1, followed by those on paint 2, followed by those on paint 3. This order might be perfectly satisfactory, but it could equally well prove to be disastrous. Observations taken over time could be affected by differences in atmospheric conditions, fatigue of the experimenter, systematic differences in the wooden board sections, etc. These could all conspire to ensure that any measurements taken during the last part of the experiment are, say, underrecorded, with the result that paint 3 appears to dry faster than the other paints when, in fact, it may be less good. The order 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3 helps to solve the problem, but it does not remove it completely (especially if the experimenter takes a break after every three observations).

There are also analytical reasons to support the use of a random assignment. It will be seen in Chaps. 3 and 4 that common forms of analysis of the data depend on the F and t distributions. It can be shown that a random assignment ensures that these distributions are the correct ones to use. The interested reader is referred to Kempthorne (1977).

To understand the meaning of randomization, consider an experiment to compare the effects on blood pressure of three exercise programs, where each program is observed four times, giving a total of 12 observations. Now, given 12 subjects, imagine making a list of all possible assignments of the 12 subjects to the three exercise programs so that 4 subjects are assigned to each program. (There are 12!/(4!4!4!), or 34,650 ways to do this.) If the assignment of subjects to programs is done in such a way that every possible assignment has the same chance of occurring, then the assignment is said to be a *completely random assignment*. Completely randomized designs, discussed in Chaps. 3–7 of this book, are randomized in this way. It is, of course, possible that a random assignment itself could

lead to the order 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3. If the experimenter expressly wishes to avoid certain assignments, then a different type of design should be used. An experimenter should not look at the resulting assignment, decide that it does not look very random, and change it.

Without the aid of an objective randomizing device, it is not possible for an experimenter to make a random assignment. In fact, it is not even possible to select a single number at random. This is borne out by a study run at the University of Delaware and reported by Professor Hoerl in the *Royal Statistical Society News and Notes* (January 1988). The study, which was run over several years, asked students to pick a number at random between 0 and 9. The numbers 3 and 7 were selected by about 40% of the students. This is twice as many as would be expected if the numbers were truly selected at random.

The most frequently used objective mechanism for achieving a random assignment in experimental design is a random number generator. A random number generator is a computer program that gives as output a very long string of digits that are integers between 0 and 9 inclusive and that have the following properties. All integers between 0 and 9 occur approximately the same number of times, as do all pairs of integers, all triples, and so on. Furthermore, there is no discernible pattern in the string of digits, and hence the name "random" numbers.

The random numbers in Appendix Table A.1 are part of a string of digits produced by a random number generator (in SAS® version 6.09 on a DEC Model 4000 MODEL 610 computer at Wright State University). Many experimenters and statistical consultants will have direct access to their own random number generator on a computer or calculator and will not need to use the table. The table is divided into six sections (pages), each section containing six groups of six rows and six groups of six columns. The grouping is merely a device to aid in reading the table. To use the table, a random starting place must be found. An experimenter who always starts reading the table at the same place always has the same set of digits, and these could not be regarded as random. The grouping of the digits by six rows and columns allows a random starting place to be obtained using five rolls of a fair die. For example, the five rolls giving 3, 1, 3, 5, 2 tells the experimenter to find the digit that is in Sect. 3 of the table, row group 1, column group 3, row 5, column 2. Then the digits can be read singly, or in pairs, or triples, etc. from the starting point across the rows.

The most common random number generators on computers or calculators generate n-digit real numbers between zero and one. Single digit random numbers can be obtained from an n-digit real number by reading the first digit after the decimal point. Pairs of digits can be obtained by reading the first two digits after the decimal point, and so on. The use of random numbers for randomization is shown in Sects. 3.2, 3.8.1 and 3.9.1.

1.2   Analysis: Basic Principles and Techniques

In the analysis of data, it is desirable to provide both graphical and statistical analyses. Plots that illustrate the relative responses of the factor settings under study allow the experimenter to gain a feel for the practical implications of the statistical results and to communicate effectively the results of the experiment to others. In addition, data plots allow the proposed model to be checked and aid in the identification of unusual observations, as discussed in Chap. 5. Statistical analysis quantifies the relative responses of the factors, thus clarifying conclusions that might be misleading or not at all apparent in plots of the data.

The purpose of an experiment can range from exploratory (discovering new important sources of variability) to confirmatory (confirming that previously discovered sources of variability are sufficiently major to warrant further study), and the philosophy of the analysis depends on the purpose of the experiment. In the early stages of experimentation the analysis may be exploratory, and one would plot and analyze the data in any way that assists in the identification of important sources of variation. In

later stages of experimentation, analysis is usually confirmatory in nature. A mathematical model of the response is postulated and hypotheses are tested and confidence intervals are calculated.

In this book, we use *linear models* to model our response and the *method of least squares* for obtaining estimates of the parameters in the model. These are described in Chap. 3. We also use restricted maximum likelihood estimation of parameters in Chap. 19. Our models include random "error variables" that encompass all the sources of variability not explicitly present in the model. We operate under the assumption that the error terms are normally distributed. However, most of the procedures in this book are generally fairly robust to nonnormality, provided that there are no extreme observations among the data.

It is rare nowadays for experimental data to be analyzed by hand. Most experimenters and statisticians have access to a computer package that is capable of producing, at the very least, a basic analysis of data for the simplest experiments. To the extent possible, for each design discussed, we shall present useful plots and methods of analysis that can be obtained from most statistical software packages. We will also develop many of the mathematical formulas that lie behind the computer analysis. This will enable the reader more easily to appreciate and interpret statistical computer package output and the associated manuals. Computer packages vary in sophistication, flexibility, and the statistical knowledge required of the user. The SAS software (see SAS Institute Inc., 2004) is one of the better commercial statistical packages for analyzing experimental data. The R software (see R Core Team, 2017) is a command line software for statistical computing and graphics which is freely available on the web. Both packages can handle every model discussed in this book, and although they require some knowledge of experimental design on the part of the user, neither is difficult to learn. We provide some basic SAS and R statements and resulting output at the end of most chapters to illustrate data analysis. A reader who wishes to use a different computer package can run the equivalent analyses on his or her own package and compare the output with those shown. It is important that every user know exactly the capabilities of his or her own package and also the likely size of rounding errors.

It is not our intent to teach the best use of SAS and R software, and readers may find better ways of achieving the same analyses. SAS software, being a commercial package, requires purchase of a license, but many universities and companies already have site licenses. R is a free software environment for statistical computing and graphics. Links for downloading and installing R are provided in Sect. 3.9.

Planning Experiments

<div align="right">**2**</div>

2.1 Introduction

Although planning an experiment is an exciting process, it is extremely time-consuming. This creates a temptation to begin collecting data without giving the experimental design sufficient thought. Rarely will this approach yield data that have been collected in exactly the right way and in sufficient quantity to allow a good analysis with the required precision. This chapter gives a step by step guide to the experimental planning process. The steps are discussed in Sect. 2.2 and illustrated via real experiments in Sects. 2.3 and 2.5. Some standard experimental designs are described briefly in Sect. 2.4.

2.2 A Checklist for Planning Experiments

The steps in the following checklist summarize a very large number of decisions that need to be made at each stage of the experimental planning process. The steps are not independent, and at any stage, it may be necessary to go back and revise some of the decisions made at an earlier stage.

Checklist

(a) Define the objectives of the experiment.
(b) Identify all sources of variation, including:

    (i) treatment factors and their levels,
    (ii) experimental units,
    (iii) blocking factors, noise factors, and covariates.

(c) Choose a rule for assigning the experimental units to the treatments.
(d) Specify the measurements to be made, the experimental procedure, and the anticipated difficulties.
(e) Run a pilot experiment.
(f) Specify the model.
(g) Outline the analysis.
(h) Calculate the number of observations that need to be taken.
(i) Review the above decisions. Revise, if necessary.

A short description of the decisions that need to be made at each stage of the checklist is given below. Only after all of these decisions have been made should the data be collected.

(a) **Define the objectives of the experiment**.

A list should be made of the precise questions that are to be addressed by the experiment. It is this list that helps to determine the decisions required at the subsequent stages of the checklist. It is advisable to list only the essential questions, since side issues will unnecessarily complicate the experiment, increasing both the cost and the likelihood of mistakes. On the other hand, questions that are inadvertently omitted may be unanswerable from the data. In compiling the list of objectives, it can often be helpful to outline the conclusions expected from the analysis of the data. The objectives may need to be refined as the remaining steps of the checklist are completed.

(b) **Identify all sources of variation**.

A source of variation is *anything* that could cause an observation to have a different numerical value from another observation. Some sources of variation are minor, producing only small differences in the data. Others are major and need to be planned for in the experiment. It is good practice to make a list of every conceivable source of variation and then label each as either major or minor. Major sources of variation can be divided into two types: those that are of particular interest to the experimenter, called "treatment factors," and those that are not of interest, called "nuisance factors."

(i) Treatment factors and their levels.

Although the term *treatment factor* might suggest a drug in a medical experiment, it is used to mean any substance or item whose effect on the data is to be studied. At this stage in the checklist, the treatment factors and their *levels* should be selected. The levels are the specific types or amounts of the treatment factor that will actually be used in the experiment. For example, a treatment factor might be a drug or a chemical additive or temperature or teaching method, etc. The levels of such treatment factors might be the different amounts of the drug to be studied, different types of chemical additives to be considered, selected temperature settings in the range of interest, different teaching methods to be compared, etc. Few experiments involve more than four levels per treatment factor.

If the levels of a treatment factor are quantitative (i.e., can be measured), then they are usually chosen to be equally spaced. Two levels are needed to model a linear trend, three levels for a quadratic trend, and so forth. If the response or log(response) should be well modeled by a rather simple function of the log of the factor level, then one may choose the factor levels to be equally spaced on a log scale. For convenience, treatment factor levels can be coded. For example, temperature levels 60, 70, 80°, . . . might be coded as 1, 2, 3, . . . in the plan of the experiment, or as 0, 1, 2,. . . . With the latter coding, level 0 does not necessarily signify the absence of the treatment factor. It is merely a label. Provided that the experimenter keeps a clear record of the original choice of levels, no information is lost by working with the codes.

When an experiment involves more than one treatment factor, every observation is a measurement on some combination of levels of the various treatment factors. For example, if there are two treatment factors, temperature and pressure, whenever an observation is taken at a certain pressure, it must necessarily be taken at some temperature, and vice versa. Suppose there are four levels of temperature coded 1, 2, 3, 4 and three levels of pressure coded 1, 2, 3. Then there are twelve combinations of levels coded 11, 12, . . ., 43, where the first digit of each pair refers to the level

of temperature and the second digit to the level of pressure. Treatment factors are often labeled $F_1, F_2, F_3, \ldots$ or $A, B, C, \ldots$. The combinations of their levels are called *treatment combinations* and an experiment involving two or more treatment factors is called a *factorial experiment*. We will use the term *treatment* to mean a level of a treatment factor in a single factor experiment, or to mean a treatment combination in a factorial experiment.

(ii) Experimental units.

Experimental units are the "material" to which the levels of the treatment factor(s) are applied. For example, in agriculture these would be individual plots of land, in medicine they would be human or animal subjects, in industry they might be batches of raw material, factory workers, etc. If an experiment has to be run over a period of time, with the observations being collected sequentially, then the times of day can also be regarded as experimental units.

Experimental units should be representative of the material and conditions to which the conclusions of the experiment will be applied. For example, the conclusions of an experiment that uses university students as experimental units may not apply to all adults in the country. The results of a chemical experiment run in an 80° laboratory may not apply in a 60° factory. Thus it is important to consider carefully the scope of the experiment in listing the objectives in step (a).

It is important to distinguish experimental units from *observational units*—namely, what is measured to obtain observations. For example, in an experiment involving the feeding of animals in a pen to assess the effects of diet on weight gain, it may be that pens of animals fed together are the experimental units while the individual animals are the observational units. In most experiments, the experimental units and observational units are one and the same. However, when there is a distinction, it is important that the data analysis reflect it. Otherwise, mistakenly treating the observational units as experimental units would give the appearance that the experiment provides more data or replication than is indeed present.

(iii) Blocking factors, noise factors, and covariates.

An important part of designing an experiment is to enable the effects of the nuisance factors to be distinguished from those of the treatment factors. There are several ways of dealing with nuisance factors, depending on their nature.

It may be desirable to limit the scope of the experiment and to fix the level of the nuisance factor. This action may necessitate a revision of the objectives listed in step (a) since the conclusions of the experiment will not be so widely applicable. Alternatively, it may be possible to hold the level of a nuisance factor constant for one group of experimental units, change it to a different fixed value for a second group, change it again for a third, and so on. Such a nuisance factor is called a *blocking factor*, and experimental units measured under the same level of the blocking factor are said to be in the same *block* (see Chap. 10). For example, suppose that temperature was expected to have an effect on the observations in an experiment, but it was not itself a factor of interest. The entire experiment could be run at a single temperature, thus limiting the conclusions to that particular temperature. Alternatively, the experimental units could be divided into blocks with each block of units being measured at a different fixed temperature.

Even when the nuisance variation is not measured, it is still often possible to divide the experimental units into blocks of like units. For example, plots of land or times of day that are close together are more likely to be similar than those far apart. Subjects with similar characteristics are more likely to respond in similar ways to a drug than subjects with different characteristics. Observations made in the same factory are more likely to be similar than observations made in different factories.

Sometimes nuisance variation is a property of the experimental units and can be measured before the experiment takes place, (e.g., the blood pressure of a patient in a medical experiment, the I.Q. of a pupil in an educational experiment, the acidity of a plot of land in an agricultural experiment). Such a measurement is called a *covariate* and can play a major role in the analysis (see Chap. 9). Alternatively, the experimental units can be grouped into blocks, each block having a similar value of the covariate. The covariate would then be regarded as a blocking factor.

If the experimenter is interested in the variability of the response as the experimental conditions are varied, then nuisance factors are deliberately included in the experiment and not removed via blocking. Such nuisance factors are called *noise factors*, and experiments involving noise factors form the subject of *robust design*, discussed in Chap. 15.

(c) **Choose a rule by which to assign the experimental units to the levels of the treatment factors**.

The assignment rule, or the *experimental design*, specifies which experimental units are to be observed under which treatments. The choice of design, which may or may not involve blocking factors, depends upon all the decisions made so far in the checklist. There are several standard designs that are used often in practice, and these are introduced in Sect. 2.4. Further details and more complicated designs are discussed later in the book.

The actual assignment of experimental units to treatments should be done at random, subject to restrictions imposed by the chosen design. The importance of a random assignment was discussed in Sect. 1.1.4. Methods of randomization are given in Sect. 3.2.

There are some studies in which it appears to be impossible to assign the experimental units to the treatments either at random or indeed by any method. For example, if the study is to investigate the effects of smoking on cancer with human subjects as the experimental units, it is neither ethical nor possible to assign a person to smoke a given number of cigarettes per day. Such a study would therefore need to be done by observing people who have themselves chosen to be light, heavy, or nonsmokers throughout their lives. This type of study is an *observational study* and not an experiment. Although many of the analysis techniques discussed in this book could be used for observational studies, cause and effect conclusions are not valid, and such studies will not be discussed further.

(d) **Specify the measurements to be made, the experimental procedure, and the anticipated difficulties**.

The data (or observations) collected from an experiment are measurements of a response variable (e.g., the yield of a crop, the time taken for the occurrence of a chemical reaction, the output of a machine). The units in which the measurements are to be made should be specified, and these should reflect the objectives of the experiment. For example, if the experimenter is interested in detecting a difference of 0.5 gram in the response variable arising from two different treatments, it would not be sensible to take measurements to the nearest gram. On the other hand, it would be unnecessary to take measurements to the nearest 0.01 gram. Measurements to the nearest 0.1 gram would be sufficiently sensitive to detect the required difference, if it exists.

There are usually unforeseen difficulties in collecting data, but these can often be identified by taking a few practice measurements or by running a pilot experiment (see step (e)). Listing the anticipated difficulties helps to identify sources of variation required by step (b) of the checklist, and also gives the opportunity of simplifying the experimental procedure before the experiment begins.

Precise directions should be listed as to how the measurements are to be made. This might include details of the measuring instruments to be used, the time at which the measurements are to be made, the way in which the measurements are to be recorded. It is important that everyone involved in running the experiment follow these directions exactly. It is advisable to draw up a data collection sheet that shows the order in which the observations are to be made and also the units of measurement.

(e) **Run a pilot experiment**.

A pilot experiment is a mini experiment involving only a few observations. No conclusions are necessarily expected from such an experiment. It is run to aid in the completion of the checklist. It provides an opportunity to practice the experimental technique and to identify unsuspected problems in the data collection. If the pilot experiment is large enough, it can also help in the selection of a suitable model for the main experiment. The observed experimental error in the pilot experiment can help in the calculation of the number of observations required by the main experiment (step (h)).

At this stage, steps (a)–(d) of the checklist should be reevaluated and changes made as necessary.

(f) **Specify the model**.

The model must indicate explicitly the relationship that is believed to exist between the response variable and the major sources of variation that were identified at step (b). The techniques used in the analysis of the experimental data will depend upon the form of the model. It is important, therefore, that the model represent the true relationship reasonably accurately.

The most common type of model is the linear model, which shows the response variable set equal to a linear combination of terms representing the major sources of variation plus an error term representing all the minor sources of variation taken together. A pilot experiment (step (e)) can help to show whether or not the data are reasonably well described by the model.

There are two different types of treatment or block factors that need to be distinguished, since they lead to somewhat different analyses. The effect of a factor is said to be a *fixed effect* if the factor levels have been specifically selected by the experimenter and if the experimenter is interested in comparing the effects on the response variable of these specific levels. This is the most common type of factor and is the type considered in the early chapters. A model containing only fixed-effect factors (apart from the response and error random variables) is called a *fixed-effects model*.

Occasionally, however, a factor has an extremely large number of possible levels, and the levels included in the experiment are a random sample from the population of all possible levels. The effect of such a factor is said to be a *random effect*. Since the levels are not specifically chosen, the experimenter has little interest in comparing the effects on the response variable of the particular levels used in the experiment. Instead, it is the variability of the response due to the entire population of levels that is of interest. Models for which all factors are random effects are called *random-effects models*. Models for which some factors are random effects and others are fixed effects are called *mixed models*. Experiments involving random effects will be considered in Chaps. 17 and 18.

(g) **Outline the analysis**.

The type of analysis that will be performed on the experimental data depends on the objectives determined in step (a), the design selected in step (c), and its associated model specified in step (f).

The entire analysis should be outlined (including hypotheses to be tested and confidence intervals to be calculated). The analysis not only determines the calculations at step (h), but also verifies that the design is suitable for achieving the objectives of the experiment.

(h) **Calculate the number of observations needed**.

At this stage in the checklist, a calculation should be done for the number of observations that are needed in order to achieve the objectives of the experiment. If too few observations are taken, then the experiment may be inconclusive. If too many are taken, then time, energy, and money are needlessly expended.

Formulae for calculating the number of observations are discussed in Sects. 3.6 and 4.5 for the completely randomized design, and in later chapters for more complex designs. The formulae require a knowledge of the size of the experimental variability. This is the amount of variability in the data caused by the sources of variation designated as minor in step (b) (plus those sources that were forgotten!). Estimating the size of the experimental error prior to the experiment is not easy, and it is advisable to err on the large side. Methods of estimation include the calculation of the experimental error in a pilot experiment (step (e)) and previous experience of working with similar experiments.

(i) **Review the above decisions**. **Revise if necessary**.

Revision is necessary when the number of observations calculated at step (h) exceeds the number that can reasonably be taken within the time or budget available. Revision must begin at step (a), since the scope of the experiment usually has to be narrowed. If revisions are not necessary, then the data collection may commence.

It should now be obvious that a considerable amount of thought needs to precede the running of an experiment. The data collection is usually the most costly and the most time-consuming part of the experimental procedure. Spending a little extra time in planning helps to ensure that the data can be used to maximum advantage. No method of analysis can save a badly designed experiment.

Although an experimental scientist well trained in the principles of design and analysis of experiments may not need to consult a statistician, it usually helps to talk over the checklist with someone not connected with the experiment. Step (a) in the checklist is often the most difficult to complete. A consulting statistician's first question to a client is usually, "Tell me *exactly* why you are running the experiment. *Exactly* what do you want to show?" If these questions cannot be answered, it is not sensible for the experimenter to go away, collect some data, and worry about it later. Similarly, it is essential that a consulting statistician understand reasonably well not only the purpose of the experiment but also the experimental technique. It is not helpful to tell an experimenter to run a pilot experiment that eats up most of the budget.

The experimenter needs to give clear directions concerning the experimental procedure to all persons involved in running the experiment and in collecting the data. It is also necessary to check that these directions are being followed exactly as prescribed. An amusing anecdote told by Salvadori (1980) in his book *Why Buildings Stand Up* illustrates this point. The story concerns a quality control study of concrete. Concrete consists of cement, sand, pebbles, and water and is mixed in strictly controlled proportions in a concrete plant. It is then carried to a building site in a revolving drum on a large truck. A sample of concrete is taken from each truckload and, after seven days, is tested for compressive strength. Its strength depends partly upon the ratio of water to cement, and decreases as the proportion of water increases. The anecdote concerns a problem that occurred during construction of an airport

terminal in New York. Although the concrete reaching the site before noon showed good strength, some of the concrete arriving shortly after noon did not. The supervisor investigated the most plausible causes until he decided to follow the trucks as they went from the plant to the site. He spotted a truck driver regularly stopping for beer and a sandwich at noon, and to prevent the concrete hardening, he added extra water into the drums. Thus, Salvadori concludes "the prudent engineer must not only be cautious about material properties, but be aware, most of all, of human behavior."

This applies to prudent experimenters, too! In the chapters that follow, most of the emphasis falls on the statistical analysis of well-designed experiments. It is crucial to keep in mind the ideas in these first sections while reading the rest of the book. Unfortunately, there are no nice formulae to summarize everything. Both the experimenter and the statistical consultant should use the checklist and lots of common sense!

2.3   A Real Experiment—Cotton-Spinning Experiment

The experiment to be described was reported in the November 1953 issue of the journal *Applied Statistics* by Robert Peake, of the British Cotton Industry Research Association. Although the experiment was run many years ago, the types of decisions involved in planning experiments have changed very little. The original report was not written in checklist form, but all of the relevant details were provided by the author in the article.

Checklist

(a) **Define the objectives of the experiment**.

At an intermediate stage of the cotton-spinning process, a strand of cotton (known as "roving") thicker than the final thread is produced. Roving is twisted just before it is wound onto a bobbin. As the degree of twist increases, so does the strength of the cotton, but unfortunately, so does the production time and hence, the cost. The twist is introduced by means of a rotary guide called a "flyer." The purpose of the experiment was twofold; first, to investigate the way in which different degrees of twist (measured in turns per inch) affected the breakage rate of the roving, and secondly, to compare the ordinary flyer with the newly devised special flyer.

(b) **Identify all sources of variation**.

(i) Treatment factors and their levels.

There are two treatment factors, namely "type of flyer" and "degree of twist." The first treatment factor, flyer, has two levels, "ordinary" and "special." We code these as 1 and 2, respectively. The levels of the second treatment factor, twist, had to be chosen within a feasible range. A pilot experiment was run to determine this range, and four non equally spaced levels were selected, 1.63, 1.69, 1.78, and 1.90 turns per inch. Coding these levels as 1, 2, 3, and 4, there are eight possible treatment combinations, as shown in Table 2.1.

The two treatment combinations 11 and 24 were omitted from the experiment, since the pilot experiment showed that these did not produce satisfactory roving. The experiment was run with the six treatment combinations 12, 13, 14, 21, 22, 23.

Table 2.1 Treatment combinations for the cotton-spinning experiment

		Twist		
Flyer	1.63	1.69	1.78	1.90
Ordinary	(11)	12	13	14
Special	21	22	23	(24)

(ii) Experimental units.

An experimental unit consisted of the thread on the set of full bobbins in a machine on a given day. It was not possible to assign different bobbins in a machine to different treatment combinations. The bobbins needed to be fully wound, since the tension, and therefore the breakage rate, changed as the bobbin filled. It took nearly one day to wind each set of bobbins completely.

(iii) Blocking factors, noise factors, and covariates.

Apart from the treatment factors, the following sources of variation were identified: the different machines, the different operators, the experimental material (cotton), and the atmospheric conditions.
There was some discussion among the experimenters over the designation of the blocking factors. Although similar material was fed to the machines and the humidity in the factory was controlled as far as possible, it was still thought that the experimental conditions might change over time. A blocking factor representing the day of the experiment was contemplated. However, the experimenters finally decided to ignore the day-to-day variability and to include just one blocking factor, each of whose levels represented a machine with a single operator. The number of experimental units per block was limited to six to keep the experimental conditions fairly similar within a block.

(c) **Choose a rule by which to assign the experimental units to the treatments**.

A randomized complete block design, which is discussed in detail in Chap. 10, was selected. The six experimental units in each block were randomly assigned to the six treatment combinations. The design of the final experiment was similar to that shown in Table 2.2.

(d) **Specify the measurements to be made, the experimental procedure, and the anticipated difficulties**.

It was decided that a suitable measurement for comparing the effects of the treatment combinations was the number of breaks per hundred pounds of material. Since the job of machine operator included mending every break in the roving, it was easy for the operator to keep a record of every break that occurred.
The experiment was to take place in the factory during the normal routine. The major difficulties were the length of time involved for each observation, the loss of production time caused by changing the flyers, and the fact that it was not known in advance how many machines would be available for the experiment.

Table 2.2 Part of the design for the cotton-spinning experiment

| | Time order | | | | | |
Block	1	2	3	4	5	6
I	22	12	14	21	13	23
II	21	14	12	13	22	23
III	23	21	14	12	13	22
IV	23	21	12	$\cdots$	$\cdots$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

(e) **Run a pilot experiment**.

The experimental procedure was already well known. However, a pilot experiment was run in order to identify suitable levels of the treatment factor "degree of twist" for each of the flyers; see step (b).

(f) **Specify the model**.

The model was of the form

$$\text{Breakage rate} = \text{constant} + \text{effect of treatment combination}$$
$$+ \text{effect of block} + \text{error} \ .$$

Models of this form and the associated analyses are discussed in Chap. 10.

(g) **Outline the analysis**.

The analysis was planned to compare differences in the breakage rates caused by the six flyer/twist combinations. Further, the trend in breakage rates as the degree of twist was increased was of interest for each flyer separately.

(h) **Calculate the number of observations that need to be taken**.

The experimental variability was estimated from a previous experiment of a somewhat different nature. This allowed a calculation of the required number of blocks to be done (see Sect. 10.5.2). The calculation was based on the fact that the experimenters wished to detect a true difference in breakage rates of at least 2 breaks per 100 pounds with high probability. The calculation suggested that 56 blocks should be observed (a total of 336 observations!).

(i) **Review the above decisions. Revise, if necessary**.

Since each block would take about a week to observe, it was decided that 56 blocks would not be possible. The experimenters decided to analyze the data after the first 13 blocks had been run. The effect of decreasing the number of observations from the number calculated is that the requirements stated in step (h) would not be met. The probability of detecting differences of 2 breaks per 100 lbs was substantially reduced.

Table 2.3   Data from the cotton-spinning experiment

| Treatment combination | Block number | | | | | |
	1	2	3	4	5	6
12	6.0	9.7	7.4	11.5	17.9	11.9
13	6.4	8.3	7.9	8.8	10.1	11.5
14	2.3	3.3	7.3	10.6	7.9	5.5
21	3.3	6.4	4.1	6.9	6.0	7.4
22	3.7	6.4	8.3	3.3	7.8	5.9
23	4.2	4.6	5.0	4.1	5.5	3.2

| Treatment combination | Block number | | | | | | |
	7	8	9	10	11	12	13
12	10.2	7.8	10.6	17.5	10.6	10.6	8.7
13	8.7	9.7	8.3	9.2	9.2	10.1	12.4
14	7.8	5.0	7.8	6.4	8.3	9.2	12.0
21	6.0	7.3	7.8	7.4	7.3	10.1	7.8
22	8.3	5.1	6.0	3.7	11.5	13.8	8.3
23	10.1	4.2	5.1	4.6	11.5	5.0	6.4

Source Peake (1953). Copyright © 1953 Royal Statistical Society. Reprinted with permission

Fig. 2.1   A subset of the data for the cotton-spinning experiment
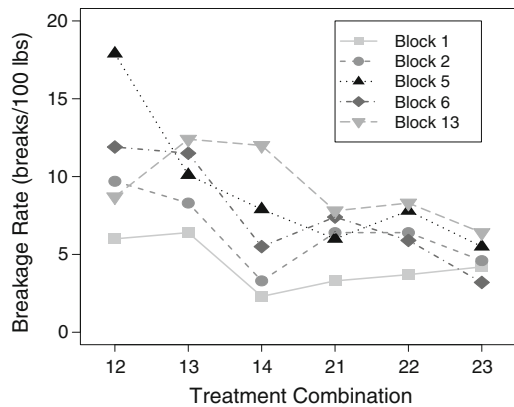
The results from the 13 blocks are shown in Table 2.3, and the data from five of these are plotted in Fig. 2.1. The data show that there are certainly differences in blocks. For example, results in block 5 are consistently above those for block 1. The breakage rate appears to be somewhat higher for treatment combinations 12 and 13 than for 23. However, the observed differences may not be any larger than the inherent variability in the data. Therefore, it is important to subject these data to a careful statistical analysis. This will be done in Sect. 10.5.

2.4   Some Standard Experimental Designs

An experimental design is a rule that determines the assignment of the experimental units to the treatments. Although experiments differ from each other greatly in most respects, there are some standard designs that are used frequently. These are described briefly in this section.

2.4.1   Completely Randomized Designs

A *completely randomized design* is the name given to a design in which the experimenter assigns the experimental units to the treatments completely at random, subject only to the number of observations to be taken on each treatment. Completely randomized designs are used for experiments that involve no blocking factors. They are discussed in depth in Chaps. 3–9 and again in some of the later chapters. The mechanics of the randomization procedure are illustrated in Sect. 3.2. The statistical properties of the design are completely determined by specification of $r_1, r_2, \ldots, r_v$, where r_i denotes the number of observations on the ith treatment, $i = 1, \ldots, v$.

The model is of the form

$$\text{Response} \ = \ \text{constant} \ + \ \text{effect of treatment} \ + \ \text{error}\,.$$

Factorial experiments often have a large number of treatments. This number can even exceed the number of available experimental units, so that only a subset of the treatment combinations can be observed. Special methods of design and analysis are needed for such experiments, and these are discussed in Chap. 15.

2.4.2   Block Designs

A *block design* is a design in which the experimenter partitions the experimental units into blocks, determines the allocation of treatments to blocks, and assigns the experimental units within each block to the treatments completely at random. Block designs are discussed in depth in Chaps. 10–14.

In the analysis of a block design, the blocks are treated as the levels of a single blocking factor even though they may be defined by a combination of levels of more than one nuisance factor. For example, the cotton-spinning experiment of Sect. 2.3 is a block design with each block corresponding to a combination of a machine and an operator. The model is of the form

$$
\begin{aligned}
\text{Response} = \ &\text{constant} \ + \ \text{effect of block} \\
&+ \ \text{effect of treatment} \ + \ \text{error}\ .
\end{aligned}
$$

The simplest block design is the *complete block design*, in which each treatment is observed the same number of times in each block. Complete block designs are easy to analyze. A complete block design whose blocks contain a single observation on each treatment is called a *randomized complete block design* or, simply, a *randomized block design*.

When the block size is smaller than the number of treatments, so that it is not possible to observe every treatment in every block, a block design is called an *incomplete block design*. The precision with which treatment effects can be compared and the methods of analysis that are applicable depend on the choice of the design. Some standard design choices, and appropriate methods of randomization, are covered in Chap. 11. Incomplete block designs for factorial experiments are discussed in Chap. 13.

2.4.3   Designs with Two or More Blocking Factors

When an experiment involves two major sources of variation that have each been designated as blocking factors, these blocking factors are said to be either *crossed* or *nested*. The difference between these is illustrated in Table 2.4. Each experimental unit occurs at some combination of levels of the two blocking

Table 2.4 Schematic plans of experiments with two blocking factors

(i) Crossed blocking factors				(ii) Nested blocking factors			
		Block				Block	
		Factor 1				Factor 1	
		1	2	3		1	2	3
Block	1	*	*	*	1	*		
Factor	2	*	*	*	2	*		
2	3	*	*	*	3	*		
					Block　4		*	
					Factor　5		*	
					2　6		*	
					7			*
					8			*
					9			*

factors, and an asterisk denotes experimental units that are to be assigned to treatment factors. It can be seen that when the block factors are crossed, experimental units are used from all possible combinations of levels of the blocking factors. When the block factors are nested, a particular level of one of the blocking factors occurs at only one level of the other blocking factor.

Crossed Blocking Factors

A design involving two crossed blocking factors is sometimes called a "row–column" design. This is due to the pictorial representation of the design, in which the levels of one blocking factor are represented by rows and the levels of the second are represented by columns as in Table 2.4(i). An intersection of a row and a column is called a "cell." Experimental units in the same cell should be similar. The model is of the form

$$\text{Response} = \text{constant} + \text{effect of row block} + \text{effect of column block}$$
$$+ \text{effect of treatment} + \text{error} \,.$$

Some standard choices of row–column designs with one experimental unit per cell are discussed in Chap. 12, and an example is given in Sect. 2.5.3 (p. 26) of a row–column design with six experimental units per cell.

The example shown in Table 2.5 is a basic design (prior to randomization) that was considered for the cotton-spinning experiment. The two blocking factors were "machine with operator" and "day." Notice that if the column headings are ignored, the design looks like a randomized complete block design. Similarly, if the row headings are ignored, the design with columns as blocks looks like a randomized complete block design. Such designs are called Latin squares and are discussed in Chap. 12. For the cotton-spinning experiment, which was run in the factory itself, the experimenters could not guarantee that the same six machines would be available for the same six days, and this led them to select a randomized complete block design. Had the experiment been run in a laboratory, so that every machine was available on every day, the Latin square design would have been used, and the day-to-day variability could have been removed from the analysis of treatments.

Nested (or Hierarchical) Blocking Factors.

Two blocking factors are said to be nested when observations taken at two different levels of one blocking factor are automatically at two different levels of the second blocking factor as in Table 2.4(ii).

Table 2.5   A Latin square for the cotton-spinning experiment

Machine with operator	Days					
	1	2	3	4	5	6
1	12	13	14	21	22	23
2	13	14	21	22	23	12
3	14	21	22	23	12	13
4	22	23	12	13	14	21
5	23	12	13	14	21	22
6	21	22	23	12	13	14

As an example, consider an experiment to compare the effects of a number of diets (the treatments) on the weight (the response variable) of piglets (the experimental units). Piglets vary in their metabolism, as do human beings. Therefore, the experimental units are extremely variable. However, some of this variability can be controlled by noting that piglets from the same litter are more likely to be similar than piglets from different litters. Also, litters from the same sow are more likely to be similar than litters from different sows. The different sows can be regarded as blocks, the litters regarded as subblocks, and the piglets as the experimental units within the subblocks. A piglet belongs only to one litter (piglets are nested within litters), and a litter belongs only to one sow (litters are nested within sows). The random assignment of piglets to diets would be done separately litter by litter in exactly the same way as for any block design.

In the industrial setting, the experimental units may be samples of some experimental material (e.g., cotton) taken from several different batches that have been obtained from several different suppliers. The samples, which are to be assigned to the treatments, are "nested within batches," and the batches are "nested within suppliers." The random assignment of samples to treatment factor levels is done separately batch by batch.

In an ordinary block design, the experimental units can be thought of as being nested within blocks. In the above two examples, an extra "layer" of nesting is apparent. Experimental units are nested within subblocks, subblocks are nested within blocks. The subblocks can be assigned at random to the levels of a further treatment factor. When this is done, the design is often known as a *split-plot design* (see Sect. 2.4.4).

2.4.4   Split-Plot Designs

A *split-plot design* is a design with at least one blocking factor where the experimental units within each block are assigned to the treatment factor levels as usual, and *in addition,* the blocks are assigned at random to the levels of a further treatment factor. This type of design is used when the levels of one (or more) treatment factors are easy to change, while the alteration of levels of other treatment factors are costly, or time-consuming. For example, this type of situation occurred in the cotton-spinning experiment of Sect. 2.3. Setting the degree of twist involved little more than a turn of a dial, but changing the flyers involved stripping down the machines. The experiment was, in fact, run as a randomized complete block design, as shown in Table 2.2. However, it could have been run as a split-plot design, as shown in Table 2.6. The time slots have been grouped into blocks, which have been assigned at random to the two flyers. The three experimental units within each cell have been assigned at random to degrees of twist.

Table 2.6 A split-plot design for the cotton-spinning experiment

	Time order					
	1	2	3	4	5	6
	Block I			Block II		
		Flyer 2			Flyer 1	
Machine I	Twist 2	Twist 1	Twist 3	Twist 2	Twist 4	Twist 3
		Flyer 2			Flyer 1	
Machine II	Twist 1	Twist 2	Twist 3	Twist 4	Twist 2	Twist 3
		Flyer 1			Flyer 2	
Machine III	Twist 4	Twist 2	Twist 3	Twist 3	Twist 1	Twist 2
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Split-plot designs also occur in medical and psychological experiments. For example, suppose that several subjects are assigned at random to the levels of a drug. In each time-slot each subject is asked to perform one of a number of tasks, and some response variable is measured. The subjects can be regarded as blocks, and the time-slots for each subject can be regarded as experimental units within the blocks. The blocks and the experimental units are each assigned to the levels of the treatment factors—the subject to drugs and the time-slots to tasks. Split-plot designs are discussed in detail in Chap. 19.

In a split-plot design, the effect of a treatment factor whose levels are assigned to the experimental units is generally estimated more precisely than a treatment factor whose levels are assigned to the blocks. It was this reason that led the experimenters of the cotton-spinning experiment to select the randomized complete block design in Table 2.2 rather than the split-plot design of Table 2.6. They preferred to take the extra time in running the experiment rather than risk losing precision in the comparison of the flyers.

2.5 More Real Experiments

Three experiments are described in this section. The first, called the "soap experiment," was run as a class project by Suyapa Silvia in 1985. The second, called the "battery experiment," was run by one of the authors. Both of these experiments are designed as completely randomized designs. The first has one treatment factor at three levels while the second has two treatment factors, each at two levels. The soap and battery experiments are included here to illustrate the large number of decisions that need to be made in running even the simplest investigations. Their data are used in Chaps. 3–5 to illustrate methods of analysis. The third experiment, called the "cake-baking experiment," includes some of the more complicated features of the designs discussed in Sect. 2.4.

2.5.1 Soap Experiment

The checklist for this experiment has been obtained from the experimenter's report. Our comments are in parentheses. The reader is invited to critically appraise the decisions made by this experimenter and to devise alternative ways of running her experiment.

Checklist (Suyapa Silvia, 1985)

(a) **Define the objectives of the experiment**.

The purpose of this experiment is to compare the extent to which three particular types of soap dissolve in water. It is expected that the experiment will answer the following questions: Are there any differences in weight loss due to dissolution among the three soaps when allowed to soak in water for the same length of time? What are these differences?

Generalizations to other soaps advertised to be of the same type as the three used for this experiment cannot be made, as each soap differs in terms of composition, i.e., has different mixtures of ingredients. Also, because of limited laboratory equipment, the experimental conditions imposed upon these soaps cannot be expected to mimic the usual treatment of soaps, i.e., use of friction, running water, etc. Conclusions drawn can only be discussed in terms of the conditions posed in this experiment, although they could give indications of what the results might be under more normal conditions.

(We have deleted the details of the actual soaps used).

(b) **Identify all sources of variation**.

(i) Treatment factors and their levels

The treatment factor, soap, has been chosen to have three levels: regular, deodorant, and moisturizing brands, all from the same manufacturer. The particular brands used in the experiment are of special interest to this experimenter.

The soap will be purchased at local stores and cut into cubes of similar weight and size—about $1''$ cubes. The cubes will be cut out of each bar of soap using a sharp hacksaw so that all sides of the cube will be smooth. They will then be weighed on a digital laboratory scale showing a precision of 10 mg. The weight of each cube will be made approximately equal to the weight of the smallest cube by carefully shaving thin slices from it. A record of the preexperimental weight of each cube will be made.

(Note that the experimenter has no control over the age of the soap used in the experiment. She is assuming that the bars of soap purchased will be typical of the population of soap bars available in the stores. If this assumption is not true, then the results of the experiment will not be applicable in general. Each cube should be cut from a different bar of soap purchased from a random sample of stores in order for the experiment to be as representative as possible of the populations of soap bars.)

(ii) Experimental units

The experiment will be carried out using identical metal muffin pans. Water will be heated to 100°F (approximate hot bath temperature), and each section will be quickly filled with 1/4 cup of water. A pilot study indicated that this amount of water is enough to cover the tops of the soaps. The water-filled sections of the muffin pans are the experimental units, and these will be assigned to the different soaps as described in step (c).

(iii) Blocking factors, noise factors, and covariates

(Apart from the differences in the composition of the soaps themselves, the initial sizes of the cubes were not identical, and the sections of the muffin pan were not necessarily all exposed to the

same amount of heat. The initial sizes of the cubes were measured by weight. These could have been used as covariates, but the experimenter chose instead to measure the weight changes, that is, "final weight minus initial weight." The sections of the muffin pan could have been grouped into blocks with levels such as "outside sections," "inside sections," or such as "center of heating vent" and "off-center of heating vent." However, the experimenter did not feel that the experimental units would be sufficiently variable to warrant blocking. Other sources of variation include inaccuracies of measuring initial weights, final weights, amounts and temperature of water. All of these were designated as minor. No noise factors were incorporated into the experiment.)

(c) **Choose a rule by which to assign the experimental units to the levels of the treatment factors**.

An equal number of observations will be made on each of the three treatment factor levels. Therefore, r cubes of each type of soap will be prepared. These cubes will be randomly matched to the experimental units (muffin pan sections) using a random-number table.
(This assignment rule defines a completely randomized design with r observations on each treatment factor level, see Chap. 3).

(d) **Specify the measurements to be made, the experimental procedure, and the anticipated difficulties**.

The cubes will be carefully placed in the water according to the assignment rule described in paragraph (c). The pans will be immediately sealed with aluminum foil in order to prevent excessive moisture loss. The pans will be positioned over a heating vent to keep the water at room temperature. Since the sections will be assigned randomly to the cubes, it is hoped that if water temperature differences do exist, these will be randomly distributed among the three treatment factor levels. After 24 hours, the contents of the pans will be inverted onto a screen and left to drain and dry for a period of 4 days in order to ensure that the water that was absorbed by each cube has been removed thoroughly. The screen will be labeled with the appropriate soap numbers to keep track of the individual soap cubes.
After the cubes have dried, each will be carefully weighed. These weights will be recorded next to the corresponding preexperimental weights to study the changes, if any, that may have occurred. The analysis will be carried out on the differences between the post- and preexperimental weights.

Expected Difficulties

 (i) The length of time required for a cube of soap to dissolve noticeably may be longer than is practical or assumed. Therefore, the data may not show any differences in weights.
 (ii) Measuring the partially dissolved cubes may be difficult with the softer soaps (e.g., moisturizing soap), since they are likely to lose their shape.
(iii) The drying time required may be longer than assumed and may vary with the soaps, making it difficult to know when they are completely dry.
(iv) The heating vent may cause the pan sections to dry out prematurely.

(After the experiment was run, Suyapa made a list of the actual difficulties encountered. They are reproduced below. Although she had run a pilot experiment, it failed to alert her to these difficulties ahead of time, since not all levels of the treatment factor had been observed.)

Difficulties Encountered

 (i) When the cubes were placed in the warm water, it became apparent that some soaps absorbed water very quickly compared to others, causing the tops of these cubes to become exposed eventually. Since this had not been anticipated, no additional water was added to these chambers in order to keep the experiment as designed. This created a problem, since the cubes of soap were not all completely covered with water for the 24-hour period.

 (ii) The drying time required was also different for the regular soap compared with the other two. The regular soap was still moist, and even looked bigger, when the other two were beginning to crack and separate. This posed a real dilemma, since the loss of weight due to dissolution could not be judged unless all the water was removed from the cubes. The soaps were observed for two more days after the data was collected and the regular soap did lose part of the water it had retained.

 (iii) When the contents of the pans were deposited on the screen, it became apparent that the dissolved portion of the soap had become a semisolid gel, and a decision had to be made to regard this as "nonusable" and not allow it to solidify along with the cubes (which did not lose their shape).

(The remainder of the checklist together with the analysis is given in Sect. 3.7. The calculations at step (h) showed that four observations should be taken on each soap type. The data were collected and are shown in Table 2.7. A plot of the data is shown in Fig. 2.2.)

The weightloss for each cube of soap measured in grams to the nearest 0.01 gm is the difference between the initial weight of the cube (pre-weight) and the weight of the same cube at the end of the experiment (post-weight). Negative values indicate a weight gain, while positive values indicate a weight loss (a large value being a greater loss). As can be seen, the regular soap cubes experienced the smallest changes in weight, and in fact, appear to have retained some of the water. Possible reasons for this will be examined in the discussion section (see Sect. 3.7.3). The data show a clear difference in the weight loss of the different soap types. This will be verified by a statistical hypothesis test (Sect. 3.7.2).

Table 2.7   Weight loss for soaps in the soap experiment

Soap (Level)	Cube	Pre-weight (grams)	Post-weight (grams)	Weightloss (grams)
Regular (1)	1	13.14	13.44	−0.30
	2	13.17	13.27	−0.10
	3	13.17	13.31	−0.14
	4	13.17	12.77	0.40
Deodorant (2)	5	13.03	10.40	2.63
	6	13.18	10.57	2.61
	7	13.12	10.71	2.41
	8	13.19	10.04	3.15
Moisturizing (3)	9	13.14	11.28	1.86
	10	13.19	11.16	2.03
	11	13.06	10.80	2.26
	12	13.00	11.18	1.82

Fig. 2.2 Weight loss for the soap experiment
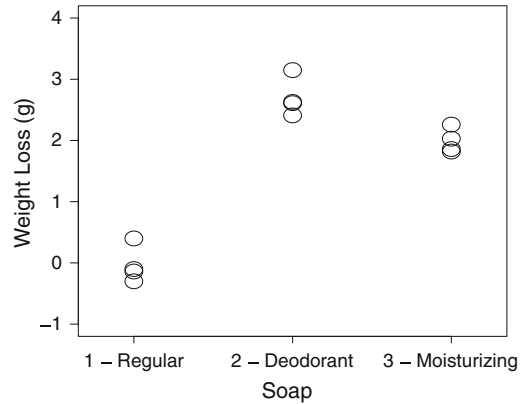
2.5.2 Battery Experiment

Checklist

(a) **Define the objectives of the experiment**.

Due to the frequency with which his family needed to purchase flashlight batteries, one of the authors (Dan Voss) was interested in finding out which type of nonrechargeable battery was the most economical. In particular, Dan was interested in comparing the lifetime per unit cost of the particular name brand that he most often purchased with the store brand where he usually shopped. He also wanted to know whether it was worthwhile paying the extra money for alkaline batteries over heavy duty batteries.

A further objective was to compare the lifetimes of the different types of battery regardless of cost. This was due to the fact that whenever there was a power cut, all the available flashlights appeared to have dead batteries! (Only the first objective will be discussed in Chaps. 3 and 4. The second objective will be addressed in Chap. 5.)

(b) **Identify all sources of variation**.

There are several sources of variation that are easy to identify in this experiment. Clearly, different duty batteries such as alkaline and heavy duty could well be an important factor in the lifetime per unit cost, as could the brand of the battery. These two sources of variation are the ones of most interest in the experiment and form the levels of the two treatment factors "duty" and "brand." Dan decided not to include regular duty batteries in the experiment.

Other possible sources of variation include the date of manufacture of the purchased battery, and whether the lifetime was monitored under continuous running conditions or under the more usual setting with the flashlight being turned on and off, the temperature of the environment, the age and variability of the flashlight bulbs.

The first of these could not be controlled in the experiment. The batteries used in the experiment were purchased at different times and in different locations in order to give a wide representation of dates of manufacture. The variability caused by this factor would be measured as part of the natural variability (error variability) in the experiment along with measurement error. Had the dates been marked on the packets, they could have been included in the analysis of the experiment as covariates. However, the dates were not available.

The second of these possible sources of variation (running conditions) was fixed. All the measurements were to be made under constant running conditions. Although this did not mimic the usual operating conditions of flashlight batteries, Dan thought that the relative ordering of the different battery types in terms of life per unit cost would be the same. The continuous running setting was much easier to handle in an experiment since each observation was expected to take several hours and no sophisticated equipment was available.

The third source of variation (temperature) was also fixed. Since the family living quarters are kept at a temperature of about $68°$ in the winter, Dan decided to run his experiment at this usual temperature. Small fluctuations in temperature were not expected to be important.

The variability due to the age of the flashlight bulb was more difficult to handle. A decision had to be made whether to use a new bulb for each observation and risk muddling the effect of the battery with that of the bulb, or whether to use the same bulb throughout the experiment and risk an effect of the bulb age from biasing the data. A third possibility was to divide the observations into blocks and to use a single bulb throughout a block, but to change bulbs between blocks. Since the lifetime of a bulb is considerably longer than that of a battery, Dan decided to use the same bulb throughout the experiment.

(i) Treatment factors and their levels

There are two treatment factors each having two levels. These are battery "duty" (level 1 = alkaline, level 2 = heavy duty) and "brand" (level 1 = name brand, level 2 = store brand). This gives four treatment combinations coded 11, 12, 21, 22. In Chaps. 3–5, we will recode these treatment combinations as 1, 2, 3, 4, and we will often refer to them as the four different treatments or the four different levels of the factor "battery type." Thus, the levels of battery type are:

Level	Treatment Combination
1	alkaline, name brand (11)
2	alkaline, store brand (12)
3	heavy duty, name brand (21)
4	heavy duty, store brand (22)

(ii) Experimental units

The experimental units in this experiment are the time slots. These were assigned at random to the battery types so as to determine the order in which the batteries were to be observed. Any fluctuations in temperature during the experiment form part of the variability between the time slots and are included in the error variability.

(iii) Blocking factors, noise factors, and covariates

As mentioned above, it was decided not to include a blocking factor representing different flashlight bulbs. Also, the date of manufacture of each battery was not available, and small fluctuations in room temperature were not thought to be important. Consequently, there were no covariates in the experiment, and no noise factors were incorporated.

(c) **Choose a rule by which to assign the experimental units to the levels of the treatment factor**.

Since there were to be no blocking factors, a completely randomized design was selected, and the time slots were assigned at random to the four different battery types.

(d) **Specify the measurements to be made, the experimental procedure, and the anticipated difficulties**.

The first difficulty was in deciding exactly how to measure lifetime of a flashlight battery. First, a flashlight requires two batteries. In order to keep the cost of the experiment low, Dan decided to wire a circuit linking just one battery to a flashlight bulb. Although this does not mimic the actual use of a flashlight, Dan thought that as with the constant running conditions, the relative lifetimes per unit cost of the four battery types would be preserved. Secondly, there was the difficulty in determining when the battery had run down. Each observation took several hours, and it was not possible to monitor the experiment constantly. Also, a bulb dims slowly as the battery runs down, and it is a judgment call as to when the battery is flat. Dan decided to deal with both of these problems by including a small clock in the circuit. The clock stopped before the bulb had completely dimmed, and the elapsed time on the clock was taken as a measurement of the battery life. The cost of a battery was computed as half of the cost of a two-pack, and the lifetime per unit cost was measured in minutes per dollar (min/$).

(e) **Run a pilot experiment**.

A few observations were run as a pilot experiment. This ensured that the circuit did indeed work properly. It was discovered that the clock and the bulb had to be wired in parallel and not in series, as Dan had first thought! The pilot experiment also gave a rough idea of the length of time each observation would take (at least four hours), and provided a very rough estimate of the error variability that was used at step (h) to calculate that four observations were needed on each treatment combination.

Difficulties Encountered

The only difficulty encountered in running the main experiment was that during the fourth observation, it was discovered that the clock was running but the bulb was out. This was due to a loose connection. The connection was repaired, a new battery inserted into the circuit, and the clock reset.

Data

The data collected in the main experiment are shown in Table 2.8 and plotted in Fig. 2.3. The experiment was run in 1993.

Table 2.8 Data for the battery experiment

Battery type	Life (min)	Unit cost ($)	Life per unit cost	Time order
1	602	0.985	611	1
2	863	0.935	923	2
1	529	0.985	537	3
4	235	0.495	476	4
1	534	0.985	542	5
1	585	0.985	593	6
2	743	0.935	794	7
3	232	0.520	445	8
4	282	0.495	569	9
2	773	0.935	827	10
2	840	0.935	898	11
3	255	0.520	490	12
4	238	0.495	480	13
3	200	0.520	384	14
4	228	0.495	460	15
3	215	0.520	413	16

Fig. 2.3 Battery life per unit cost versus battery
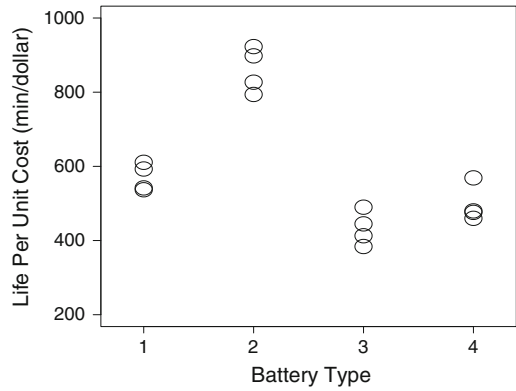
2.5.3 Cake-Baking Experiment

The following factorial experiment was run in 1979 by the baking company Spillers Ltd. (in the U.K.) and was reported in the *Bulletin in Applied Statistics* in 1980 by S.M. Lewis and A.M. Dean.

Checklist

(a) **Define the objectives of the experiment**.

The experimenters at Spillers, Ltd. wanted to know how "cake quality" was affected by adding different amounts of glycerine and tartaric acid to the cake mix.

(b) **Identify all sources of variation**.

(i) Treatment factors and their levels

The two treatment factors of interest were glycerine and tartaric acid. Glycerine was called the "first treatment factor" and labeled F_1, while tartaric acid was called the "second treatment factor" and labeled F_2. The experimenters were very familiar with the problems of cake baking and determinations of cake quality. They knew exactly which amounts of the two treatment factors they wanted to compare. They selected four equally spaced amounts of glycerine and three equally spaced amounts of tartaric acid. These were coded as 1, 2, 3, 4 for glycerine and 1, 2, 3 for tartaric acid. Therefore, the twelve coded treatment combinations were 11, 12, 13, 21, 22, 23, 31, 32, 33, 41, 42, 43.

(ii) Identify the experimental units

Before the experimental units can be identified, it is necessary to think about the experimental procedure. One batch of cake-mix was divided into portions. One of the twelve treatment combinations (i.e., a certain amount of glycerine and a certain amount of tartaric acid) was added to each portion. Each portion was then thoroughly mixed and put into a container for baking. The containers were placed on a tray in an oven at a given temperature for the required length of time. The experimenters required an entire tray of cakes to make one measurement of cake quality. Only one tray would fit on any one shelf of an oven. An experimental unit was, therefore, "an oven shelf with a tray of containers of cake-mix," and these were assigned at random to the twelve treatment combinations.

(iii) Blocking factors, noise factors, and covariates

There were two crossed blocking factors. The first was time of day with two levels (morning and afternoon). The second was oven, which had three levels, one level for each of the three ovens that were available on the day of the experiment. Each cell (defined by oven and time of day) contained six experimental units, since an oven contained six shelves (see Table 2.9). Each set of six experimental units was assigned at random to six of the twelve treatment combinations, and it

Table 2.9 Basic design for the baking experiment

Oven codes	Time of day codes											
	1						2					
1	11	13	22	24	32	34	12	14	21	23	31	33
2	12	14	21	23	32	34	11	13	22	24	31	33
3	12	14	22	24	31	33	11	13	21	23	32	34

was decided in advance which six treatment combinations should be observed together in a cell (see step (c) of the checklist).

Although the experimenters expected differences in the ovens and in different runs of the same oven, their experience showed that differences between the shelves of their industrial ovens were very minor. Otherwise, a third blocking factor representing oven shelf would have been needed.

It was possible to control carefully the amount of cake mix put into each container, and the experimenters did not think it was necessary to monitor the precooked weight of each cake. Small differences in these weights would not affect the measurement of the quality. Therefore, no covariates were used in the analysis.

(c) **Choose a rule by which to assign the experimental units to the levels of the treatment factors**.

Since there were two crossed blocking factors, a row–column design with six experimental units per cell was required. It was not possible to observe every treatment combination in every cell. However, it was thought advisable to observe all twelve treatment combinations in each oven, either in the morning or the afternoon. This precaution was taken so that if one of the ovens failed on the day of the experiment, the treatment combinations could still all be observed twice each. The basic design (before randomization) that was used by Spillers is shown in Table 2.9. The experimental units (the trays of containers on the six oven shelves) need to be assigned at random to the 6 treatment combinations cell by cell. The oven codes need to be assigned to the actual ovens at random, and the time of day codes 1 and 2 to morning and afternoon.

Exercises

Exercises 1–7 refer to the list of experiments in Table 2.10.

1. Table 2.10 gives a list of experiments that can be run as class projects. Select a simple experiment of interest to you, but preferably not on the list. Complete steps (a)–(c) of the checklist with the intention of actually running the experiment when the checklist is complete.
2. For experiments 1 and 7 in Table 2.10, complete steps (a) and (b) of the checklist. There may be more than one treatment factor. Give precise definitions of their levels.
3. For experiment 2, complete steps (a)–(c) of the checklist.
4. For experiment 3, complete steps (a)–(c) of the checklist.
5. For experiment 4, list sources of variation. Decide which sources can be controlled by limiting the scope of the experiment or by specifying the exact experimental procedure to be followed. Of the

Table 2.10 Some simple experiments

1.	Compare the growth rate of bean seeds under different watering and lighting schedules.
2.	Does the boiling point of water differ with different concentrations of salt?
3.	Compare the strengths of different brands of paper towel.
4.	Do different makes of popcorn give different proportions of unpopped kernels? What about cooking methods?
5.	Compare the effects of different locations of an observer on the speed at which subjects locate the occurrences of the letter "e" in a written passage.
6.	Do different colored candles burn at different speeds?
7.	Compare the proportions of words remembered from lists of related or unrelated words, and under various conditions such as silence and distraction.
8.	Compare the effects of different colors of exam paper on students' performance in an examination.

remaining sources of variation, decide which are minor and which are major. Are there any blocking factors in this experiment?

6. For experiment 6, specify what measurements should be made, how they should be made, and list any difficulties that might be expected.

7. For experiment 8, write down all the possible sources of variation. In your opinion, should this experiment be run as a completely randomized design, a block design, or a design with more than one blocking factor? Justify your answer.

8. Read critically through the checklists in Sect. 2.5. Would you suggest any changes? Would you have done anything differently? If you had to criticize these experiments, which points would you address?

9. The following description was given by Clifford Pugh in the 1953 volume of *Applied Statistics*.

"The widespread use of detergents for domestic dish washing makes it desirable for manufacturers to carry out tests to evaluate the performance of their products. ... Since foaming is regarded as the main criterion of performance, the measure adopted is the number of plates washed before the foam is reduced to a thin surface layer. The five main factors which may affect the number of plates washed by a given product are (i) the concentration of detergent, (ii) the temperature of the water, (iii) the hardness of the water, (iv) the type of "soil" on the plates, and (v) the method of washing used by the operator. ... The difficulty of standardizing the soil is overcome by using the plates from a works canteen (cafeteria) for the test and adopting a randomized complete block technique in which plates from any one course form a block .... One practical limitation is the number of plates available in any one block. This permits only four ... tests to be completed (in a block)."

Draw up steps (a)–(d) of a checklist for an experiment of the above type and give an example of a design that fits the requirements of your checklist.

3

3.1 Introduction

In working through the checklist in Chap. 2, the experimenter must choose an experimental design at step (c). A design is the rule that determines the assignment of the experimental units to treatments. The simplest possible design is the *completely randomized design*, where the experimental units are assigned to the treatments completely at random, subject to the number of observations to be taken on each treatment. Completely randomized designs involve no blocking factors.

Two ways of calculating the required number of observations (sample sizes) on each treatment are presented in Sects. 3.6 and 4.5. The first method chooses sample sizes to obtain desired powers of hypothesis tests, and the second chooses sample sizes to achieve desired lengths of confidence intervals. We sometimes refer to the list of treatments and the corresponding sample sizes as the design, with the understanding that the assignment of experimental units to treatments is to be done completely at random.

In this chapter, we discuss the random assignment procedure for the completely randomized design, we introduce the method of least squares for estimating model parameters, and we develop a procedure for testing equality of the treatment parameters. Analyses by the SAS and R software are described at the end of the chapter.

3.2 Randomization

In this section we provide a procedure for randomization that is very easily applied using a computer, but can equally well be done by hand. On a computer, the procedure requires the availability of software that stores data in rows and columns (like spreadsheet software, a SAS data set, or a data.frame or matrix in R), that includes a function that randomly generates real numbers between zero and one, and that includes the capacity to sort rows by the values in one column.

We use r_i to denote the number of observations to be taken on the ith treatment, and $n = \Sigma r_i$ to denote the total number of observations (and hence the required number of experimental units). We code the treatments from 1 to v and label the experimental units 1 to n.

Step 1: Enter into one column r_1 1's, then r_2 2's, …, and finally r_v v's, giving a total of $n = \Sigma r_i$ entries. These represent the treatment labels.

Step 2: Enter into another column $n = \Sigma r_i$ random numbers, including enough digits to avoid ties. (The random numbers can be generated by a computer program or read from Table A.1).

Table 3.1 Randomization

Unsorted treatments	Unsorted random numbers	Sorted treatments	Sorted random numbers	Experimental unit
1	0.533	3	0.139	1
1	0.683	2	0.379	2
2	0.702	3	0.411	3
2	0.379	1	0.533	4
3	0.411	1	0.683	5
3	0.962	2	0.702	6
3	0.139	3	0.962	7

Step 3:   Reorder both columns so that the random numbers are put in ascending order. This arranges the treatment labels into a random order.

Step 4:   Assign experimental unit t to the treatment whose label is in row t.

If the number n of experimental units is a k-digit integer, then the list in step 2 should be a list of k-digit random numbers. To obtain k-digit random numbers from Table A.1, a random starting place is found as described in Sect. 1.1.4, p. 3. The digits are then read across the rows in groups of k (ignoring spaces).

We illustrate the randomization procedure using the SAS software in Sect. 3.8.1, p. 52, and using the R software in Sect. 3.9.1, p. 59. The procedure can equally well be done using the random digits in Table A.1 and sorting by hand.

Example 3.2.1 Randomization

Consider a completely randomized design for three treatments and sample sizes $r_1 = r_2 = 2, r_3 = 3$. The unrandomized design (step 1 of the randomization procedure) is 1 1 2 2 3 3 3, and is listed in column 1 of Table 3.1. Suppose step 2 generates the random numbers in column 2 of Table 3.1. In step 3, columns 1 and 2 are sorted so that the entries in column 2 are in ascending order. This gives columns 3 and 4. In step 4, the entries in column 3 are matched with experimental units 1–7 in order, so that column 3 contains the design after randomization. Treatment 1 is in rows 4 and 5, so experimental units 4 and 5 are assigned to treatment 1. Likewise, units 2 and 6 are assigned to treatment 2, and units 1, 3 and 7 are assigned to treatment 3. The randomly ordered treatments are then 3 2 3 1 1 2 3, and the experimental units 1–7 are assigned to the treatments in this order.   □

3.3   Model for a Completely Randomized Design

A model is an equation that shows the dependence of the response variable upon the levels of the treatment factors. (Models involving block effects or covariates are considered in later chapters.)

Let Y_{it} be a random variable that represents the response obtained on the tth observation of the ith treatment. Let the parameter μ_i denote the "true response" of the ith treatment, that is, the response that would always be obtained from the ith treatment if it could be observed under *identical* experimental conditions and measured without error. Of course, this ideal situation can never happen—there is always some variability in the experimental procedure even if only caused by inaccuracies in reading measuring instruments. Sources of variation that are deemed to be minor and ignored during the planning of the experiment also contribute to variation in the response variable. These sources of nuisance variation are usually represented by a single variable ϵ_{it}, called an *error variable*, which is a random variable with zero mean. The model is then

$$Y_{it} = \mu_i + \epsilon_{it}, \quad t = 1, \ldots, r_i, \quad i = 1, \ldots, v,$$

where v is the number of treatments and r_i is the number of observations to be taken on the ith treatment. An alternative way of writing this model is to replace the parameter μ_i by $\mu + \tau_i$, so that the model becomes

$$Y_{it} = \mu + \tau_i + \epsilon_{it}, \quad t = 1, \ldots, r_i, \quad i = 1, \ldots, v.$$

In this model, $\mu + \tau_i$ denotes the true mean response for the ith treatment, and examination of differences between the parameters μ_i in the first model is equivalent to examination of differences between the parameters τ_i in the second model.

It will be seen in Sect. 3.4 that unique estimates of the parameters in the second formulation of the model cannot be obtained. Nevertheless, many experimenters prefer this model. The parameter μ is a constant, and the parameter τ_i represents the positive or negative deviation of the response from this constant when the ith treatment is observed. This deviation is called the "effect" on the response of the ith treatment.

The above models are *linear models*, that is, the response variable is written as a linear function of the parameters. Any model that is not, or cannot, be transformed into a linear model cannot be treated by the methods in this book. Linear models often provide reasonably good approximations to more complicated models, and they are used extensively in practice.

The specific forms of the distributions of the random variables in a model need to be identified before any statistical analyses can be done. The error variables represent all the minor sources of variation taken together, including all the measurement errors. In many experiments, it is reasonable to assume that the error variables are independent and that they have a normal distribution with zero mean and unknown variance σ^2, which must be estimated. We call these assumptions the *error assumptions*. It will be shown in Chap. 5 that plots of the experimental data give good indications of whether or not the error assumptions are likely to be true. Proceeding with the analysis when the constant variance, normality, or independence assumptions are violated can result in a totally incorrect analysis.

A complete statement of the model for any experiment should include the list of error assumptions. Thus, for a completely randomized design with v specifically selected treatments (fixed effects), the model is

$$
\begin{aligned}
Y_{it} &= \mu + \tau_i + \epsilon_{it}\,, \\
\epsilon_{it} &\sim N(0, \sigma^2)\,, \\
&\epsilon_{it}\text{'s are mutually independent,} \\
t &= 1, \ldots, r_i, \quad i = 1, \ldots, v,
\end{aligned}
\tag{3.3.1}
$$

where "$\sim N(0, \sigma^2)$" denotes "has a normal distribution with mean 0 and variance σ^2." This is sometimes called a *one-way analysis of variance model*, since the model includes only one major source of variation, namely the treatment effect, and because the standard analysis of data using this model involves a comparison of measures of variation.

Notice that it is unnecessary to specify the distribution of Y_{it} in the model, as it is possible to deduce this from the stated information. Since Y_{it} is modeled as the sum of a treatment mean $\mu + \tau_i$ and a normally distributed random variable ϵ_{it}, it follows that

$$Y_{it} \sim N(\mu + \tau_i, \sigma^2).$$

Also, since the ϵ_{it}'s are mutually independent, the Y_{it}'s must also be mutually independent. Therefore, if the model is a true representation of the behavior of the response variable, then the data values y_{it} for the ith treatment form a random sample from a $N(\mu + \tau_i, \sigma^2)$ distribution.

3.4 Estimation of Parameters

3.4.1 Estimable Functions of Parameters

A function of the parameters of any linear model is said to be *estimable* if and only if it can be written as the expected value of a linear combination of the response variables. Only estimable functions of the parameters have unique linear unbiased estimates. Since it makes no sense to work with functions that have an infinite possible number of values, it is important that the analysis of the experiment involve only the estimable functions. For the one-way analysis of variance model (3.3.1), every estimable function is of the form

$$
E\left[\sum_i \sum_t a_{it} Y_{it}\right] = \sum_i \sum_t a_{it} E[Y_{it}]
$$

$$
= \sum_i \sum_t a_{it}(\mu + \tau_i) \;\; = \;\; \sum_i b_i(\mu + \tau_i),
$$

where $b_i = \Sigma_t a_{it}$ and the a_{it}'s are real numbers. Any function not of this form is *nonestimable*.

Clearly, $\mu + \tau_1$ is estimable, since it can be obtained by setting $b_1 = 1$ and $b_2 = b_3 = \cdots = b_v = 0$. Similarly, each $\mu + \tau_i$ is estimable. If we choose $b_i = c_i$ where $\sum c_i = 0$, we see that $\sum c_i \tau_i$ is estimable. Any such function $\sum c_i \tau_i$ for which $\sum_i c_i = 0$ is called a *contrast*, so all contrasts are estimable in the one-way analysis of variance model. For example, setting $b_1 = 1$, $b_2 = -1$, $b_3 = \cdots = b_v = 0$ shows that $\tau_1 - \tau_2$ is estimable. Similarly, each $\tau_i - \tau_s$, $i \neq s$, is estimable. Notice that there are no values of b_i that give μ, τ_1, τ_2, ..., or τ_v separately as the expected value. Therefore, these parameters are not individually estimable.

3.4.2 Notation

We write the ith treatment sample mean as

$$
\overline{Y}_{i.} = \frac{1}{r_i}\left(\sum_{t=1}^{r_i} Y_{it}\right)
$$

and the corresponding observed sample mean as $\overline{y}_{i.}$. The "dot" notation means "add over all values of the subscript replaced with a dot," and the "bar" means "divide by the number of terms that have been added up." This notation will be extremely useful throughout this book. For example, in the next subsection we write

$$
\frac{1}{n}\sum_{i=1}^{v}\sum_{t=1}^{r_i} y_{it} = \frac{1}{n}\sum_{i=1}^{v} y_{i.} = \frac{1}{n}y_{..} = \overline{y}_{..}, \;\; \text{where } n = \sum_{i=1}^{v} r_i = r_{.},
$$

so that $\overline{y}_{..}$ is the average of all of the observations. Note that if the summation applies to a subscript on two variables, the dot notation cannot be used. For example, $\sum r_i \hat{\tau}_i$ cannot be written as $r_{.}\hat{\tau}_{.}$, since $r_{.}\hat{\tau}_{.}$ denotes $(\sum r_i)(\sum \hat{\tau}_i)$. Also note that when notation involves both a sum and a square, such as $y_{..}^2$ or $\overline{y}_{i.}^2$, the sum is taken first and then the sum is squared.

3.4.3   Obtaining Least Squares Estimates

The *method of least squares* is used to obtain estimates and estimators for estimable functions of parameters in linear models. We shall show that the ith treatment sample mean $\overline{Y}_{i.}$ and its observed value $\overline{y}_{i.}$ are the "least squares estimator" and "least squares estimate," respectively, of $\mu + \tau_i$. Least squares solutions for the parameters $\mu, \tau_1, \ldots, \tau_v$ are any set of corresponding values $\hat{\mu}, \hat{\tau}_1, \ldots, \hat{\tau}_v$ that minimize the sum of squared errors

$$\sum_{i=1}^{v} \sum_{t=1}^{r_i} e_{it}^2 = \sum_{i=1}^{v} \sum_{t=1}^{r_i} (y_{it} - \mu - \tau_i)^2. \tag{3.4.2}$$

The estimated model $\hat{y}_{it} = \hat{\mu} + \hat{\tau}_i$ is the model that best fits the data in the sense of minimizing (3.4.2).

Finding least squares solutions is a standard problem in calculus.[1] The sum of squared errors (3.4.2) is differentiated with respect to each of the parameters $\mu, \tau_1, \ldots, \tau_v$ in turn. Then each of the $v + 1$ resulting derivatives is set equal to zero, yielding a set of $v + 1$ equations. These $v + 1$ equations are called the *normal equations*. Any solution to the normal equations gives a minimum value of the sum of squared errors (3.4.2) and provides a set of least squares solutions for the parameters.

The reader is asked to verify in Exercise 6 that the normal equations for the one-way analysis of variance model (3.3.1) are those shown in (3.4.3). The first equation in (3.4.3) is obtained by setting the derivative of the sum of squared errors of (3.4.2) with respect to μ equal to zero, and the other v equations are obtained by setting the derivatives with respect to each τ_i in turn equal to zero. We put "hats" on the parameters at this stage to denote solutions. The $v + 1$ normal equations are

$$y_{..} - n\hat{\mu} - \sum_i r_i \hat{\tau}_i = 0, \tag{3.4.3}$$

$$y_{i.} - r_i \hat{\mu} - r_i \hat{\tau}_i = 0, \quad i = 1, \ldots, v,$$

and include $v + 1$ unknown parameters. From the last v equations, we obtain

$$\hat{\mu} + \hat{\tau}_i = \overline{y}_{i.}, \quad i = 1, \ldots, v,$$

so the least squares solution for the ith treatment mean $\mu + \tau_i$ is the corresponding sample mean $\overline{y}_{i.}$.

There is a problem in solving the normal equations to obtain least squares solutions for each parameter $\mu, \tau_1, \ldots, \tau_v$ individually. If the last v normal equations (3.4.3) are added together, the first equation results. This means that the $v+1$ equations are not distinct (not linearly independent). The last v normal equations *are* distinct, since they each contain a different τ_i. Thus, there are exactly v distinct normal equations in $v + 1$ unknown parameters, and there is no unique solution for the parameters. This is not surprising, in view of the fact that we have already seen in Sect. 3.4.1 that these parameters are not individually estimable. For practical purposes, any one of the infinite number of solutions will be satisfactory, since they lead to identical solutions for the estimable parameters. To obtain any one of these solutions, it is necessary to add a further equation to the set of normal equations. *Any* extra equation can be added, provided that it is not a linear combination of the equations already present. The trick is to add whichever equation will aid most in solving the entire set of equations.

---

[1]Readers without a background in calculus may note that the least squares solutions for the parameters, individually, are not unique and then may skip forward to Sect. 3.4.4.

One obvious possibility is to add the equation $\hat{\mu} = 0$, in which case the normal equations become

$$\hat{\mu} = 0\,,$$
$$y_{..} - \sum_i r_i \hat{\tau}_i = 0\,,$$
$$y_{i.} - r_i \hat{\tau}_i = 0\,, \quad i = 1, \ldots, v.$$

It is then a simple matter to solve the last v equations for the $\hat{\tau}_i$'s, yielding $\hat{\tau}_i = y_{i.}/r_i = \overline{y}_{i.}$. Thus, one solution to the normal equations is

$$\hat{\mu} = 0\,,$$
$$\hat{\tau}_i = \overline{y}_{i.}\,, \quad i = 1, \ldots, v.$$

A more common solution is obtained by adding the extra equation $\sum_i r_i \hat{\tau}_i = 0$ to (3.4.3). In this case, the normal equations become

$$\sum_i r_i \hat{\tau}_i = 0\,,$$
$$y_{..} - n\hat{\mu} = 0\,,$$
$$y_{i.} - r_i \hat{\mu} - r_i \hat{\tau}_i = 0\,, \quad i = 1, \ldots, v,$$

from which we obtain the least squares solutions

$$\hat{\mu} = \overline{y}_{..}\,,$$
$$\hat{\tau}_i = \overline{y}_{i.} - \overline{y}_{..}\,, \quad i = 1, \ldots, v.$$

Still another solution, used, for example, by the SAS software, is obtained by adding the equation $\hat{\tau}_v = 0$. Then the solutions to the normal equations are

$$\hat{\mu} = \overline{y}_{v.}\,,$$
$$\hat{\tau}_i = \overline{y}_{i.} - \overline{y}_{v.}\,, \quad i = 1, \ldots, v.$$

The default solution for the R software is similar and obtained by adding the equation $\hat{\tau}_1 = 0$. In each of the sets of solutions just obtained, it is always true that

$$\hat{\mu} + \hat{\tau}_i = \overline{y}_{i.}.$$

No matter which extra equation is added to the normal equations, $\overline{y}_{i.}$ will *always* be the least squares solution for $\mu + \tau_i$. Thus, although it is not possible to obtain unique least squares solutions for μ and τ_i separately, the least squares solution for the estimable true treatment mean $\mu + \tau_i$ *is* unique. We call $\overline{y}_{i.}$ the *least squares estimate* and $\overline{Y}_{i.}$ the *least squares estimator* of $\mu + \tau_i$. The notation $\hat{\mu} + \hat{\tau}_i$ is used somewhat ambiguously to mean both the least squares estimator and estimate. It should be clear from the context which of these is meant.

3.4.4   Properties of Least Squares Estimators

An important property of a least squares estimator is that

the least squares estimator of any estimable function of the parameters is the unique best linear unbiased estimator.

This statement, called the *Gauss–Markov Theorem*, is true for all linear models whose error variables are independent and have common variance σ^2. The theorem tells us that for the one-way analysis of variance model (3.3.1), the least squares estimator $\sum b_i \overline{Y}_{i.}$ of the estimable function $\sum b_i(\mu + \tau_i)$ is unique, is unbiased and has smallest variance. The theorem also tells us that τ_i cannot be estimable, since we have three different solutions for τ_i and none of the corresponding estimators has expected value equal to τ_i.

For the one-way analysis of variance model, Y_{it} has a normal distribution with mean $\mu + \tau_i$ and variance σ^2 (see Sect. 3.3), so $E[\overline{Y}_{i.}] = \mu + \tau_i$ and $\text{Var}(\overline{Y}_{i.}) = \sigma^2/r_i$. Therefore, the distribution of the least squares estimator $\overline{Y}_{i.}$ of $\mu + \tau_i$ is

$$\overline{Y}_{i.} \sim N(\mu + \tau_i \, , \ \sigma^2/r_i) \, .$$

The $\overline{Y}_{i.}$'s are independent, since they are based on different Y_{it}'s. Consequently, the distribution of the least squares estimator $\sum c_i \overline{Y}_{i.}$ of the contrast $\sum c_i \tau_i$, with $\sum c_i = 0$, is

$$\sum c_i \overline{Y}_{i.} \sim N(\Sigma c_i \tau_i, \ \Sigma \frac{c_i^2}{r_i} \sigma^2).$$

Example 3.4.1 Heart-lung pump experiment
The following experiment was run by Richard Davis at The Ohio State University in 1987 to determine the effect of the number of revolutions per minute (rpm) of the rotary pump head of an Olson heart–lung pump on the fluid flow rate. The rpm was set directly on the tachometer of the pump console and PVC tubing of size 3/8" by 3/32" was used. The flow rate was measured in liters per minute. Five equally spaced levels of the treatment factor "rpm" were selected, namely, 50, 75, 100, 125, and 150 rpm, and these were coded as 1, 2, 3, 4, 5, respectively. The experimental design was a completely randomized design with $r_1 = r_3 = r_5 = 5$, $r_2 = 3$, and $r_4 = 2$. The data, in the order collected, are given in Table 3.2, and the summary information is

$$\begin{aligned}
y_{1.} &= \ \ 5.676, \ r_1 = 5, \ \overline{y}_{1.} = 1.1352, \\
y_{2.} &= \ \ 5.166, \ r_2 = 3, \ \overline{y}_{2.} = 1.7220, \\
y_{3.} &= 11.634, \ r_3 = 5, \ \overline{y}_{3.} = 2.3268, \\
y_{4.} &= \ \ 5.850, \ r_4 = 2, \ \overline{y}_{4.} = 2.9250, \\
y_{5.} &= 17.646, \ r_5 = 5, \ \overline{y}_{5.} = 3.5292.
\end{aligned}$$

The least squares estimate of the mean fluid flow rate when the pump is operating at 150 rpm is

$$(\hat{\mu} + \hat{\tau}_5) = \overline{y}_{5.} = 3.5292$$

liters per minute. The other mean fluid flow rates are estimated in a similar way. The experimenter expected the flow rate to increase as the rpm of the pump head was increased. Figure 3.1 supports this expectation.

Table 3.2 Fluid flow obtained from the rotary pump head of an Olson heart–lung pump

Observation	rpm	Level	Liters/minute
1	150	5	3.540
2	50	1	1.158
3	50	1	1.128
4	75	2	1.686
5	150	5	3.480
6	150	5	3.510
7	100	3	2.328
8	100	3	2.340
9	100	3	2.298
10	125	4	2.982
11	100	3	2.328
12	50	1	1.140
13	125	4	2.868
14	150	5	3.504
15	100	3	2.340
16	75	2	1.740
17	50	1	1.122
18	50	1	1.128
19	150	5	3.612
20	75	2	1.740

Fig. 3.1 Plot of data for the heart–lung pump experiment
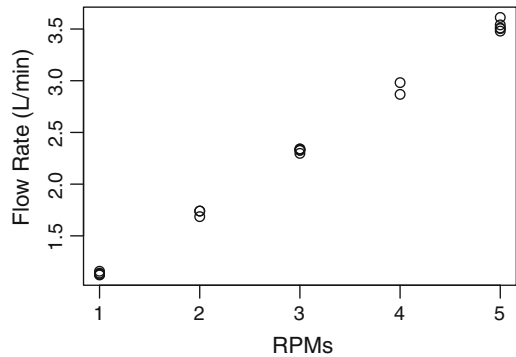
Since the variance of the least squares estimator $\overline{Y}_{i.}$ of $\mu + \tau_i$ is σ^2/r_i, the first, third, and fifth treatment means are more precisely measured than the second and fourth.

The least squares estimate of the difference in fluid flow rate between 50 and 150 rpm is

$$(\hat{\tau}_5 - \hat{\tau}_1) = (\hat{\mu} + \hat{\tau}_5) - (\hat{\mu} + \hat{\tau}_1) = \overline{y}_{5.} - \overline{y}_{1.} = 2.394$$

liters per minute. The associated variance is

$$\sum_{i=1}^{5} \frac{c_i^2}{r_i}\sigma^2 = \left(\frac{1}{5} + \frac{1}{5}\right)\sigma^2 = 0.4\,\sigma^2.$$

3.4.5 Estimation of σ^2

The least squares estimates $\hat{\mu} + \hat{\tau}_i = \overline{y}_{i.}$ of $\mu + \tau_i$ $(i = 1, \ldots, v)$ minimize the sum of squared errors. Therefore, for the one-way analysis of variance model (3.3.1), the minimum possible value of the sum of squared errors (3.4.2), which we write as *ssE*, is equal to

$$ssE = \sum_i \sum_t \hat{e}_{it}^2 = \sum_i \sum_t (y_{it} - \hat{\mu} - \hat{\tau}_i)^2.$$

Here, $\hat{e}_{it} = (y_{it} - \hat{\mu} - \hat{\tau}_i)$ is the deviation of the tth observation on the ith treatment from the estimated ith treatment mean. This is called the (it)th *residual*. Substituting the least squares estimates $\hat{\mu} + \hat{\tau}_i = \overline{y}_{i.}$ into the formula for *ssE*, we have

$$ssE = \sum_i \sum_t (y_{it} - \overline{y}_{i.})^2. \tag{3.4.4}$$

The minimum sum of squared errors, *ssE*, is called the *sum of squares for error* or the *error sum of squares*, and is used below to find an unbiased estimate of the error variance σ^2. A useful computational formula for *ssE* is obtained by multiplying out the quantity in parentheses in (3.4.4); that is,

$$ssE = \sum_i \sum_t y_{it}^2 - \sum_i r_i \overline{y}_{i.}^2 \tag{3.4.5}$$

Now, the random variable *SSE* corresponding to the minimum sum of squared errors *ssE* in (3.4.4) is

$$SSE = \sum_i \sum_t (Y_{it} - \overline{Y}_{i.})^2 = \sum_i (r_i - 1) S_i^2, \tag{3.4.6}$$

where $S_i^2 = \sum_{t=1}^{r_i} (Y_{it} - \overline{Y}_{i.})^2 / (r_i - 1))$ is the sample variance for the ith treatment. In Exercise 3.11, the reader is asked to verify that S_i^2 is an unbiased estimator of the error variance σ^2. Then, the expected value of *SSE* is

$$E(SSE) = \sum_i (r_i - 1) E(S_i^2) = (n - v)\sigma^2,$$

giving an unbiased estimator of σ^2 as

$$\hat{\sigma}^2 = SSE/(n - v) = MSE. \tag{3.4.7}$$

The corresponding unbiased estimate of σ^2 is the observed value of *MSE*, namely $msE = ssE/(n-v)$. Both *MSE* and *msE* are called the *mean square for error* or *error mean square*. The estimate *msE* is sometimes called the "within groups (or within treatments) variation."

3.4.6 Confidence Bound for σ^2

If an experiment were to be repeated in the future, the estimated value of σ^2 obtained from the current experiment could be used at step (h) of the checklist to help calculate the number of observations that should be taken in the new experiment (see Sects. 3.6.2 and 4.5). However, the error variance in the

new experiment is unlikely to be exactly the same as that in the current experiment, and in order not to underestimate the number of observations needed, it is advisable to use a larger value of σ^2 in the sample size calculation. One possibility is to use the upper limit of a one-sided confidence interval for σ^2.

It can be shown that the distribution of SSE/σ^2 is chi-squared with $n - v$ degrees of freedom, denoted by χ^2_{n-v}. Consequently,

$$P \left(\frac{SSE}{\sigma^2} \geq \chi^2_{n-v,1-\alpha} \right) = 1 - \alpha \,, \tag{3.4.8}$$

where $\chi^2_{n-v,1-\alpha}$ is the percentile of the chi-squared distribution with $n - v$ degrees of freedom and with probability of $1 - \alpha$ in the right-hand tail.

Manipulating the inequalities in (3.4.8), and replacing SSE by its observed value ssE, gives a one-sided $100(1 - \alpha)\%$ confidence bound for σ^2 as

$$\sigma^2 \leq \frac{ssE}{\chi^2_{n-v,1-\alpha}} \,. \tag{3.4.9}$$

This upper bound is called a $100(1 - \alpha)\%$ *upper confidence limit* for σ^2.

Example 3.4.2 Battery experiment, continued
The data of the battery experiment (Sect. 2.5.2, p. 24) are summarized in Table 3.3. The sum of squares for error is obtained from (3.4.5); that is,

$$\begin{aligned}
ssE &= \sum_i \sum_t y_{it}^2 - \sum_i r_i \bar{y}_{i.}^2 \\
&= 6,028,288 - 4(570.75^2 + 860.50^2 + 433.00^2 + 496.25^2) \\
&= 28,412.5.
\end{aligned}$$

An unbiased estimate of the error variance is then obtained as

$$msE = ssE/(n - v) = 28,412.5/(16 - 4) = 2367.71 \,.$$

A 95% upper confidence limit for σ^2 is given by

$$\sigma^2 \ \leq \ \frac{ssE}{\chi^2_{12,0.95}} \ = \ \frac{28,412.5}{5.23} \ = \ 5432.60 \,,$$

and taking the square root of the confidence limit, a 95% upper confidence limit for σ is 73.71 minutes per dollar. If the experiment were to be repeated in the future, the calculation for the number of observations at step (h) of the checklist might take the largest likely value for σ to be around 70–75 minutes per dollar. $\qquad\square$

Table 3.3   Data for the battery experiment

Battery type	Life per unit cost (minutes per dollar)				$\overline{y}_{i.}$
1	611	537	542	593	570.75
2	923	794	827	898	860.50
3	445	490	384	413	433.00
4	476	569	480	460	496.25

3.5   One-Way Analysis of Variance

3.5.1   Testing Equality of Treatment Effects

In an experiment involving v treatments, an obvious question is whether or not the treatments differ at all in terms of their effects on the response variable. Thus one may wish to test the null hypothesis

$$H_0 : \{\tau_1 = \tau_2 = \cdots = \tau_v\}$$

that the treatment effects are all equal against the alternative hypothesis

$$H_A : \{\text{at least two of the } \tau_i\text{'s differ}\}.$$

At first glance, the null hypothesis appears to involve nonestimable parameters. However, we can easily rewrite it in terms of $v - 1$ estimable contrasts, as follows:

$$H_0 : \{\tau_1 - \tau_2 = 0 \ \text{ and } \ \tau_1 - \tau_3 = 0 \ \text{ and } \ \cdots \ \text{ and } \ \tau_1 - \tau_v = 0\}.$$

This is not the only way to rewrite H_0 in terms of estimable contrasts. For example, we could use the contrasts $\tau_i - \overline{\tau}_.$ (where $\overline{\tau}_. = \sum \tau_i / v$) and write the null hypothesis as follows:

$$H_0 : \{\tau_1 - \overline{\tau}_. = 0 \text{ and } \tau_2 - \overline{\tau}_. = 0 \text{ and } \ \cdots \ \text{ and } \ \tau_v - \overline{\tau}_. = 0\} \,.$$

Now $\overline{\tau}_.$ is the average of the τ_i's, so the $\tau_i - \overline{\tau}_.$'s add to zero. Consequently, if $\tau_i - \overline{\tau}_. = 0$ for $i = 1, \ldots, v-1$, then $\tau_v - \overline{\tau}_.$ must also be zero. Thus, this form of the null hypothesis could be written in terms of just the first $v - 1$ estimable functions $\tau_1 - \overline{\tau}_., \ldots, \tau_{v-1} - \overline{\tau}_..$

Any way that we rewrite H_0 in terms of estimable functions of the parameters, it will always depend on $v - 1$ distinct contrasts. The number $v - 1$ is called the *treatment degrees of freedom*.

The basic idea behind an analysis of variance test is that the sum of squares for error measures how well the model fits the data. Consequently, a way of testing H_0 is to compare the sum of squares for error under the original one-way analysis of variance model (3.3.1), known as the *full model*, with that obtained from the modified model, which assumes that the null hypothesis is true. This modified model is called the *reduced model*.

Under H_0, the τ_i's are equal, and we can write the common value of $\tau_1, \ldots, \tau_v$ as τ. If we incorporate this into the one-way analysis of variance model, we obtain the reduced model

Fig. 3.2 Residuals under the full and reduced models when H_0 is false

Residuals; full model   Residuals; reduced model
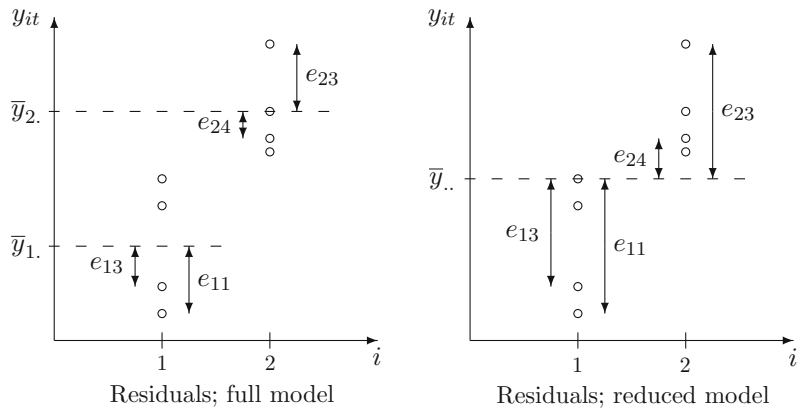
$$Y_{it} = \mu + \tau + \epsilon_{it}^0 \,,$$
$$\epsilon_{it}^0 \sim N(0, \sigma^2) \,,$$
$$\epsilon_{it}^0{}'s \text{ are mutually independent} \,,$$
$$t = 1, \ldots, r_i, \quad i = 1, \ldots, v,$$

where we write ϵ_{it}^0 for the (it)th error variable in the reduced model. To calculate the sum of squares for error, ssE_0, we need to determine the value of $\mu + \tau$ that minimizes the sum of squared errors

$$\sum_i \sum_t (y_{it} - \mu - \tau)^2 \,.$$

Using calculus, the reader is asked to show in Exercise 7 that the unique least squares estimate of $\mu + \tau$ is the sample mean of all the observations; that is, $\hat{\mu} + \hat{\tau} = \overline{y}_{..}$. Therefore, the error sum of squares for the reduced model is

$$\begin{aligned} ssE_0 &= \sum_i \sum_t (y_{it} - \overline{y}_{..})^2 \\ &= \sum_i \sum_t y_{it}^2 - n\overline{y}_{..}^2 \,. \end{aligned} \tag{3.5.10}$$

If the null hypothesis $H_0 : \{\tau_1 = \tau_i = \cdots = \tau_v\}$ is false, and the treatment effects differ, the sum of squares for error ssE under the full model (3.3.1) is considerably smaller than the sum of squares for error ssE_0 for the reduced model. This is depicted in Fig. 3.2. On the other hand, if the null hypothesis is true, then ssE_0 and ssE will be very similar. The analysis of variance test is based on the difference $ssE_0 - ssE$, relative to the size of ssE; that is, the test is based on $(ssE_0 - ssE)/ssE$. We would want to reject H_0 if this quantity is large.

We call $ssT = ssE_0 - ssE$ the *sum of squares for treatments* or the *treatment sum of squares*, since its value depends on the differences between the treatment effects. Using formulas (3.5.10) and (3.4.5) for ssE_0 and ssE, the treatment sum of squares is

$$ssT = ssE_0 - ssE \tag{3.5.11}$$

$$= \left(\sum_i \sum_t y_{it}^2 - n\overline{y}_{..}^2 \right) - \left(\sum_i \sum_t y_{it}^2 - \sum_i r_i \overline{y}_{i.}^2 \right)$$

$$= \sum_i r_i \overline{y}_{i.}^2 - n\overline{y}_{..}^2 . \tag{3.5.12}$$

An equivalent formulation is

$$ssT = \sum_i r_i (\overline{y}_{i.} - \overline{y}_{..})^2 . \tag{3.5.13}$$

The reader is invited to multiply out the parentheses in (3.5.13) and verify that (3.5.12) is obtained. There is a shortcut method of expanding (3.5.13) to obtain (3.5.12). First write down each term in y and square it. Then associate with each squared term the signs in (3.5.13). Finally, precede each term with the summations and constant outside the parentheses in (3.5.13). This quick expansion will work for all terms like (3.5.13) in this book. Formula (3.5.13) is probably the easier form of ssT to remember, while (3.5.12) is easier to manipulate for theoretical work and use for computations.

Since we will reject H_0 if ssT/ssE is large, we need to know what "large" means. This in turn means that we need to know the distribution of the corresponding random variable SST/SSE when H_0 is true, where

$$SST = \sum_i r_i (\overline{Y}_{i.} - \overline{Y}_{..})^2 \quad \text{and} \quad SSE = \sum_i \sum_t (Y_{it} - \overline{Y}_{i.})^2 . \tag{3.5.14}$$

Now, as mentioned in Sect. 3.4.6, it can be shown that SSE/σ^2 has a chi-squared distribution with $n - v$ degrees of freedom, denoted by χ^2_{n-v}. Similarly, it can be shown that when H_0 is true, SST/σ^2 has a χ^2_{v-1} distribution, and that SST and SSE are independent. The ratio of two independent chi-squared random variables, each divided by their degrees of freedom, has an F distribution. Therefore, if H_0 is true, we have

$$\frac{SST/\sigma^2(v - 1)}{SSE/\sigma^2(n - v)} \sim F_{v-1,n-v} .$$

We now know the distribution of SST/SSE multiplied by the constant $(n - v)/(v - 1)$, and we want to reject the null hypothesis $H_0 : \{\tau_1 = \cdots = \tau_v\}$ in favor of the alternative hypothesis $H_A : \{$at least two of the treatment effects differ$\}$ if this ratio is large. Thus, if we write $msT = ssT/(v - 1)$, $msE = ssE/(n - v)$, where ssT and ssE are the observed values of the treatment sum of squares and error sum of squares, respectively, our decision rule is to

$$\text{reject } H_0 \quad \text{if} \quad \frac{msT}{msE} > F_{v-1,n-v,\alpha} , \tag{3.5.15}$$

where $F_{v-1,n-v,\alpha}$ is the critical value from the F distribution with $v - 1$ and $n - v$ degrees of freedom with α in the right-hand tail. The probability α is often called the *significance level* of the test and is the probability of rejecting H_0 when in fact it is true (a Type I error). Thus, α should be selected to be small if it is important not to make a Type I error ($\alpha = 0.01$ and 0.001 are typical choices); otherwise, α can be chosen to be a little larger ($\alpha = 0.10$ and 0.05 are typical choices). Critical values $F_{v-1,n-v,\alpha}$ for the F distribution are given in Table A.6. Due to lack of space, only a few typical values of α have been tabulated.

Table 3.4 One-way analysis of variance table

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	Expected mean square
Treatments	v-1	ssT	$\frac{ssT}{v-1}$	$\frac{msT}{msE}$	$\sigma^2 + Q(\tau_i)$
Error	n-v	ssE	$\frac{ssE}{n-v}$		σ^2
Total	n-1	$sstot$			

<div align="center">Computational formulae</div>

$ssT = \sum_i r_i \overline{y}_{i.}^2 - n\overline{y}_{..}^2$ $\qquad\qquad\qquad\qquad ssE = \sum_i \sum_t y_{it}^2 - \sum_i r_i \overline{y}_{i.}^2$

$sstot = \sum_i \sum_t y_{it}^2 - n\overline{y}_{..}^2$

$Q(\tau_i) = \sum_i r_i (\tau_i - \sum_h r_h \tau_h / n)^2 / (v-1)$

The calculations involved in the test of the hypothesis H_0 against H_A are usually written as an *analysis of variance table* as shown in Table 3.4. The last line shows the *total sum of squares* and *total degrees of freedom*. The total sum of squares, *sstot*, is $(n-1)$ times the sample variance of all of the data values. Thus,

$$sstot = \sum_i \sum_t (y_{it} - \overline{y}_{..})^2 = \sum_i \sum_t y_{it}^2 - n\overline{y}_{..}^2. \qquad (3.5.16)$$

From (3.5.10), we see that *sstot* happens to be equal to ssE_0 for the one-way analysis of variance model, and from (3.5.11) we see that

$$sstot = ssT + ssE.$$

Thus, the total sum of squares consists of a part *ssT* that is explained by differences between the treatment effects and a part *ssE* that is not explained by any of the parameters in the model.

Example 3.5.1 `Battery experiment, continued`
Consider the battery experiment introduced in Sect. 2.5.2, p. 24. The sum of squares for error was calculated in Example 3.4.2, p. 40, to be $ssE = 28,412.5$. The life per unit cost responses and treatment averages are given in Table 3.3, p. 41. From these, we have $\Sigma\Sigma y_{it}^2 = 6,028,288$, $\overline{y}_{..} = 590.125$, and $r_i = 4$. Hence, the sums of squares ssT (3.5.12) and $sstot$ (3.5.16) are

$$\begin{aligned}
ssT &= \sum r_i \overline{y}_{i.}^2 - n\overline{y}_{..}^2 \\
&= 4(570.75^2 + 860.50^2 + 433.00^2 + 496.25^2) - 16(590.125)^2 \\
&= 427,915.25, \\
sstot = ssE_0 &= \sum\sum y_{it}^2 - n\overline{y}_{..}^2 \\
&= 6,028,288 - 16(590.125)^2 = 456,327.75,
\end{aligned}$$

and we can verify that $sstot = ssT + ssE$.

The decision rule for testing the null hypothesis $H_0 : \{\tau_1 = \tau_2 = \tau_3 = \tau_4\}$ that the four battery types have the same average life per unit cost against the alternative hypothesis that at least two of the battery types differ, at significance level α, is

$$\text{reject } H_0 \text{ if } msT/msE = 60.24 > F_{3,12,\alpha}.$$

Table 3.5   One-way analysis of variance table for the battery experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Type	3	427,915.25	142,638.42	60.24	0.0001
Error	12	28,412.50	2,367.71		
Total	15	456,327.75			

From Table A.6, it can be seen that $60.24 > F_{3,12,\alpha}$ for any of the tabulated values of α. For example, if α is chosen to be 0.01, then $F_{3,12,0.01} = 5.95$. Thus, for any tabulated choice of α, the null hypothesis is rejected, and it is concluded that at least two of the battery types differ in mean life per unit cost. In order to investigate which particular pairs of battery types differ, we would need to calculate confidence intervals. This will be done in Chap. 4.   □

3.5.2   Use of p-Values

The *p-value* of a test is the smallest choice of α that would allow the null hypothesis to be rejected. For convenience, computer packages usually print the p-value as well as the ratio *msT/msE*. Having information about the p-value saves looking up $F_{v-1,n-v,\alpha}$ in Table A.6. All we need to do is to compare the p-value with our selected value of α. Therefore, the decision rule for testing $H_0 : \{\tau_1 = \cdots \tau_v\}$ against $H_A : \{\text{not all of } \tau_i\text{'s are equal}\}$ can be written as

$$\text{reject } H_0 \text{ if } p < \alpha.$$

Example 3.5.2 Battery experiment, continued
In the battery experiment of Example 3.5.1, the null hypothesis $H_0 : \{\tau_1 = \tau_2 = \tau_3 = \tau_4\}$ that the four battery types have the same average life per unit cost was tested against the alternative hypothesis that they do not. The p-value generated by SAS software for the test is shown in Table 3.5 as $p = 0.0001$. A value of 0.0001 in the SAS computer output indicates that the p-value is less than or equal to 0.0001. Smaller values are not printed explicitly. If α were chosen to be 0.01, then the null hypothesis would be rejected, since $p < \alpha$.   □

3.6   Sample Sizes

Before an experiment can be run, it is necessary to determine the number of observations that should be taken on each treatment. This forms step (h) of the checklist in Sect. 2.2. In order to make this determination, the experimenter must first ascertain the approximate cost, in both time and money, of taking each observation and whether the cost differs for different levels of the treatment factor(s). There will probably be a fixed budget for the entire experiment. Therefore, remembering to set aside sufficient resources for the analysis of the experimental data, a rough calculation can be made of the maximum number, N, of observations that can be afforded. After having worked through steps (a)–(g) of the checklist, the experimenter will have identified the objectives of the experiment and the type of analysis required. It must now be ascertained whether or not the objectives of the experiment can be achieved within the budget. The calculations at step (h) may show that it is unnecessary to take as many as N observations, in which case valuable resources can be saved. Alternatively, and unfortunately

the more likely, it may be found that more than N observations are needed in order to fulfill all the experimenter's requirements of the experiment. In this case, the experimenter needs to go back and review the decisions made so far in order to try to relax some of the requirements. Otherwise, an increase in budget needs to be obtained. There is little point in running the experiment with smaller sample sizes than those required without finding out what effect this will have on the analysis. The following quotation from J. N. R. Jeffers in his article "Acid rain and tree roots: an analysis" in *The Statistical Consultant in Action* (1987) is worth careful consideration:

> There is a quite strongly held view among experimenters that statisticians always ask for more replication than can be provided, and hence jeopardize the research by suggesting that it is not worth doing unless sufficient replication can be provided. There is, of course, some truth in this allegation, and equally, some truth in the view that, unless an experiment can be done with adequate replication, and with due regard to the size of the difference which it is important to be able to detect, the research may indeed not be worth doing.

We will consider two methods of determining the number of observations on each treatment (the sample sizes). One method, which involves specifying the desired length of confidence intervals, will be presented in Sect. 4.5. The other method, which involves specifying the power required of the analysis of variance, is the topic of this section. Since the method uses the expected value of the mean square for treatments, we calculate this first.

3.6.1   Expected Mean Squares for Treatments

The formula for *SST*, the treatment sum of squares, was given in (3.5.14) on p. 43. Its expected value is

$$E[SST] = E[\sum r_i (\overline{Y}_{i.} - \overline{Y}_{..})^2]$$
$$= E[\Sigma r_i \overline{Y}_{i.}^2 - n\overline{Y}_{..}^2]$$
$$= \sum r_i E[\overline{Y}_{i.}^2] - nE[\overline{Y}_{..}^2] .$$

From the definition of the variance of a random variable, we know that $\text{Var}(X) = E[X^2] - (E[X])^2$, so we can write $E[SST]$ as

$$E[SST] = \Sigma r_i [\text{Var}(\overline{Y}_{i.}) + (E[\overline{Y}_{i.}])^2] - n[\text{Var}(\overline{Y}_{..}) + (E[\overline{Y}_{..}])^2] .$$

For the one-way analysis of variance model (3.3.1), the response variables Y_{it} are independent, and each has a normal distribution with mean $\mu + \tau_i$ and variance σ^2. So,

$$E[SST] = \sum r_i \left(\sigma^2/r_i + (\mu + \tau_i)^2 \right)$$
$$-n \left(\sigma^2/n + \left(\mu + \sum r_i \tau_i/n \right)^2 \right)$$
$$= v\sigma^2 + n\mu^2 + 2\mu \sum r_i \tau_i + \sum r_i \tau_i^2$$
$$-\sigma^2 - n\mu^2 - 2\mu \sum r_i \tau_i - (\sum r_i \tau_i)^2/n$$
$$= (v - 1)[\sigma^2 + Q(\tau_i)] ,$$

where

$$Q(\tau_i) = \Sigma_i r_i \left(\tau_i - \Sigma_h r_h \tau_h / n\right)^2 / (v - 1), \tag{3.6.17}$$

which reduces to $Q(\tau_i) = r \sum_i (\tau_i - \overline{\tau}.)^2 / (v - 1)$ when $r_1 = r_2 = \ldots r_v = r$. The expected value of the mean square for treatments $MST = SST / (v - 1)$ is

$$E[MST] = \sigma^2 + Q(\tau_i),$$

which is the quantity we listed in the analysis of variance table, Table 3.4. We note that when the treatment effects are all equal, $Q(\tau_i) = 0$, and $E[MST] = \sigma^2$.

3.6.2   Sample Sizes Using Power of a Test

Suppose that one of the major objectives of an experiment is to examine whether or not the treatments all have a similar effect on the response. The null hypothesis is actually somewhat unrealistic. The effects of the treatments are almost certainly not *exactly* equal, and even if they were, the nuisance variability in the experimental data would mask this fact. In any case, if the different levels produce only a very small difference in the response variable, the experimenter may not be interested in discovering this fact. For example, a difference of 5 minutes in life per dollar in two different batteries would probably not be noticed by most users. However, a larger difference such as 60 minutes may well be noticed. Thus the experimenter might require H_0 to be rejected with high probability if $\tau_i - \tau_s > 60$ minutes per dollar for some $i \neq s$ but may not be concerned about rejecting the null hypothesis if $\tau_i - \tau_s \leq 5$ minutes per dollar for all $i \neq s$. In most experiments, there is some value Δ such that if the difference in the effects of any two of the treatments exceeds Δ, the experimenter would like to reject the null hypothesis in favor of the alternative hypothesis with high probability.

The *power* of the test at Δ, denoted by $\pi(\Delta)$, is the probability of rejecting H_0 when the effects of at least two of the treatments differ by Δ. The power of the test $\pi(\Delta)$ is a function of Δ and also of the sample sizes, the number of treatments, the significance level α, and the error variance σ^2. Consequently, the sample sizes can be determined if $\pi(\Delta)$, v, α, and σ^2 are known. The values of Δ, $\pi(\Delta)$, v, and α are chosen by the experimenter, but the error variance has to be guessed using data from a pilot study or another similar experiment. In general, the largest likely value of σ^2 should be used. If the guess for σ^2 is too small, then the power of the test will be lower than the specified $\pi(\Delta)$. If the guess for σ^2 is too high, then the power will be higher than needed, and differences in the τ_i's smaller than Δ will cause H_0 to be rejected with high probability.

The rule for testing the null hypothesis $H_0 : \{\tau_1 = \cdots = \tau_v\}$ against H_A: {at least two of the τ_i's differ}, given in (3.5.15), on p. 43, is

$$\text{reject } H_0 \text{ if } \frac{msT}{msE} > F_{v-1, n-v, \alpha} .$$

As stated in Sect. 3.5.1, the test statistic MST/MSE has an F distribution if the null hypothesis is correct. But if the null hypothesis is false, then MST/MSE has a related distribution called a noncentral F distribution. The noncentral F distribution is denoted by $F_{v-1, n-v, \delta^2}$, where δ^2 is called the *noncentrality parameter* and is defined to be

$$\delta^2 = (v - 1)Q(\tau_i) / \sigma^2, \tag{3.6.18}$$

where $Q(\tau_i)$ was calculated in (3.6.17). When $Q(\tau_i) = 0$, then $\delta^2 = 0$, and the distribution of MST/MSE becomes the usual F-distribution. Otherwise, δ^2 is greater than zero, and the mean and spread of the distribution of MST/MSE are larger than those of the usual F-distribution. For equal sample sizes $r_1 = r_2 = \cdots = r_v = r$, we see that δ^2 is

$$\delta^2 = r \sum_i (\tau_i - \overline{\tau}.)^2 / \sigma^2.$$

The calculation of the sample size r required to achieve a power $\pi(\Delta)$ at Δ for given v, α, and σ^2 rests on the fact that the hardest situation to detect is that in which the effects of two of the factor levels (say, the first and last) differ by Δ, and the others are all equal and midway between; that is,

$$\mu + \tau_2 = \mu + \tau_3 = \cdots = \mu + \tau_{v-1} = c,$$

$$\mu + \tau_1 = c + \Delta/2, \quad \text{and} \quad \mu + \tau_v = c - \Delta/2,$$

for some constant c. In this case,

$$\delta^2 = r \sum_i \frac{(\tau_i - \overline{\tau}.)^2}{\sigma^2} = \frac{r\Delta^2}{2\sigma^2}. \tag{3.6.19}$$

The power of the test depends on the sample size r through the distribution of MST/MSE, which depends on δ^2. Since the power of the test is the probability of rejecting H_0, we have

$$\pi(\Delta) = P\left(\frac{MST}{MSE} > F_{v-1, n-v, \alpha}\right).$$

The noncentral F distribution is tabulated in Table A.7, with power π given as a function of $\phi = \delta/\sqrt{v}$ for various values of $\nu_1 = v - 1$, $\nu_2 = n - v$, and α. Using (3.6.19),

$$\phi^2 = \frac{\delta^2}{v} = \frac{r\Delta^2}{2v\sigma^2},$$

so

$$r = \frac{2v\sigma^2\phi^2}{\Delta^2}. \tag{3.6.20}$$

Hence, given α, Δ, v, and σ^2, the value of r can be determined from Table A.7 to achieve a specified power $\pi(\Delta)$. The determination has to be done iteratively, since the denominator degrees of freedom, $\nu_2 = n - v = v(r - 1)$, depend on the unknown r. The procedure is as follows:

(a) Find the section of Table A.7 for the numerator degrees of freedom $\nu_1 = v - 1$ and the specified α (only $\alpha = 0.05$ is shown).
(b) Calculate the denominator degrees of freedom using $\nu_2 = 1000$ in the first iteration and $\nu_2 = n - v = v(r - 1)$ in the following iterations, and locate the appropriate row of the table, taking the smaller listed value of ν_2 if necessary.
(c) For the required power $\pi(\Delta)$, use interpolation to determine the corresponding value of ϕ, or take the larger listed value if necessary.

(d) Calculate $r = 2v\sigma^2\phi^2/\Delta^2$, rounding up to the nearest integer. (The first iteration gives a lower bound for r.)

(e) Repeat steps (b)–(d) until the value of r is unchanged or alternates between two values. Select the larger of alternating values.

Example 3.6.1 `Soap experiment, continued`
The first part of the checklist for the soap experiment is given in Sect. 2.5.1, p. 20, and is continued in Sect. 3.7, below. At step (h), the experimenter calculated the number of observations needed on each type of soap as follows.

The error variance was estimated to be about 0.007 grams squared from the pilot experiment. In testing the hypothesis $H_0 : \{\tau_1 = \tau_2 = \tau_3\}$, the experimenter deemed it important to be able to detect a difference in weight loss of at least $\Delta = 0.25$ g between any two soap types, with a probability 0.90 of correctly doing so, and a probability 0.05 of a Type I error. This difference was considered to be the smallest discrepancy in the weight loss of soaps that would be noticeable.

Using a one-way analysis of variance model, for $v = 3$ treatments, with $\Delta = 0.25$, $r = 2v\sigma^2\phi^2/\Delta^2 = 0.672\phi^2$, and $\nu_2 = v(r - 1) = 3(r - 1)$, r was calculated as follows. Using Table A.7 for $\nu_1 = v - 1 = 2$, $\alpha = 0.05$, and $\pi(\Delta) = 0.90$:

r	$\nu_2 = 3(r-1)$	ϕ	$r = 0.672\phi^2$	Action
	1000	2.25	3.40	Round up to $r = 4$
4	9	2.50	4.20	Round up to $r = 5$
5	12	2.50	4.20	Stop, and use $r = 4$ or 5.

The experimenter decided to take $r = 4$ observations on each soap type. Sections 3.8.3 and 3.9.4 show how to make these calculations using the SAS and R software, respectively.   □

3.7   A Real Experiment—Soap Experiment, Continued

The objective of the soap experiment described in Sect. 2.5.1, p. xx, was to compare the extent to which three different types of soap dissolve in water. The three soaps selected for the experiment were a regular soap, a deodorant soap, and a moisturizing soap from a single manufacturer, and the weight-loss after 24 h of soaking and 4 days drying is reproduced in Table 3.6. Steps (a)–(d) of the checklist were given in Sect. 2.5.1. The remaining steps and part of the analysis of the experimental data are described below. The first part of the description is based on the written report of the experimenter, Suyapa Silvia.

Table 3.6   Data for the soap experiment

Soap	Weight-loss (grams)				$\overline{y}_{i.}$
1	−0.30	−0.10	−0.14	0.40	−0.0350
2	2.63	2.61	2.41	3.15	2.7000
3	1.86	2.03	2.26	1.82	1.9925

3.7.1   Checklist, Continued

(e) **Run a pilot experiment**.

A pilot experiment was run and used for two purposes. First, it helped to identify the difficulties listed at step (d) of the checklist. Secondly, it provided an estimate of σ^2 for step (h). The error variance was estimated to be about $0.007\,\text{g}^2$. The value $0.007\ \text{gm}^2$ was the value of *msE* in the pilot experiment. In fact, this is an underestimate, and it would have been better to have used the one-sided confidence bound (3.4.9) for σ^2.

(f) **Specify the model**.

Since care will be taken to control all extraneous sources of variation, it is assumed that the following model will be a reasonable approximation to the true model.

$$Y_{it} = \mu + \tau_i + \epsilon_{it}\,,$$
$$\epsilon_{it} \sim N(0, \sigma^2)\,,$$
$$\epsilon_{it}'\text{s are mutually independent}$$
$$i = 1, 2, 3; \quad t = 1, \dots r_i;$$

where τ_i is the (fixed) effect on the response of the ith soap, μ is a constant, Y_{it} is the weight loss of the tth cube of the ith soap, and ϵ_{it} is a random error.

Before analyzing the experimental data, the assumptions concerning the distribution of the error variables will be checked using graphical methods. (Assumption checking will be discussed in Chap. 5).

(g) **Outline the analysis**.

In order to address the question of differences in weights, a one-way analysis of variance will be computed at $\alpha = 0.05$ to test

versus
$$H_0 : \{\tau_1 = \tau_2 = \tau_3\}$$
$$H_A : \{ \text{ the effects of at least two pairs of soap types differ}\}.$$

To find out more about the differences among pairs of treatments, 95% confidence intervals for the pairwise differences of the τ_i will be calculated using Tukey's method (Tukey's method will be discussed in Sect. 4.4.4).

(h) **Calculate the number of observations that need to be taken**.

Four observations will be taken on each soap type. (See Example 3.6.1, p. 49, for the calculation.)

(i) **Review the above decisions. Revise if necessary**.

It is not difficult to obtain 4 observations on each of 3 soaps, and therefore the checklist does not need revising. Small adjustments to the experimental procedure that were found necessary during the pilot experiment have already been incorporated into the checklist.

3.7.2   Data Collection and Analysis

The data collected by the experimenter are plotted in Fig. 2.2, p. 24, and reproduced in Table 3.6. The assumptions that the error variables are independent and have a normal distribution with constant

Table 3.7 One-way analysis of variance table for the soap experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Soap	2	16.1220	8.0610	104.45	0.0001
Error	9	0.6946	0.0772		
Total	11	16.8166			

variance were checked (using methods to be described in Chap. 5) and appear to be satisfied. The least squares estimates, $\hat{\mu} + \hat{\tau}_i = \bar{y}_{i.}$, of the average weight loss values (in grams) are

$$\bar{y}_{1.} = -0.0350 , \qquad \bar{y}_{2.} = 2.7000 , \qquad \bar{y}_{3.} = 1.9925.$$

The hypothesis of no differences in weight loss due to the different soap types is tested below using an analysis of variance test.

Using the values $\bar{y}_{i.}$ given above, together with $\sum\sum y_{it}^2 = 45.7397$ and $r_1 = r_2 = r_3 = 4$, the sums of squares for Soap and Total are calculated using (3.5.12) and (3.5.16), pp. 43 and 44, as

$$
\begin{aligned}
ssT &= \sum r_i \bar{y}_{i.}^2 - n\bar{y}_{..}^2 \\
&= \left[4(-0.0350)^2 + 4(2.7000)^2 + 4(1.9925)^2\right] - \left[12(1.5525)^2\right] = 16.1220 , \\
sstot = ssE_0 &= \sum\sum y_{it}^2 - n\bar{y}_{..}^2 \\
&= 45.7397 - 12(1.5525)^2 = 16.8166.
\end{aligned}
$$

The sum of squares for error can be calculated by subtraction, giving $ssE = sstot - ssT = 0.6946$, or directly from (3.4.5), p. 39, as

$$
\begin{aligned}
ssE &= \sum\sum y_{it}^2 - \sum r_i \bar{y}_{i.}^2 \\
&= 45.7397 - \left[4(-0.0350)^2 + 4(2.7000)^2 + 4(1.9925)^2\right] = 0.6946 .
\end{aligned}
$$

The estimate of error variability is then

$$\hat{\sigma}^2 = msE = ssE/(n - v) = 0.6945/(12 - 3) = 0.0772 .$$

The sums of squares and mean squares are shown in the analysis of variance table, Table 3.7. Notice that the estimate of σ^2 is ten times larger than the estimate of $0.007\,\mathrm{g}^2$ provided by the pilot experiment. This suggests that the pilot experiment was not sufficiently representative of the main experiment. As a consequence, the actual power of detecting a difference of $\Delta = 0.25\,\mathrm{g}$ between the weight losses of the soaps is, in fact, somewhat below the desired probability of 0.90.

The decision rule for testing $H_0 : \{\tau_1 = \tau_2 = \tau_3\}$ against the alternative hypothesis, that at least two of the soap types differ in weight loss, using a significance level of $\alpha = 0.05$, is to reject H_0 if $msT/msE = 104.45 > F_{2,9,0.05}$. From Table A.6, $F_{2,9,0.05} = 4.26$. Consequently, the null hypothesis is rejected, and it is concluded that at least two of the soap types do differ in their weight loss after 24 h in water (and 4 days drying time). This null hypothesis would have been rejected for most practical choices of α. If α had been chosen to be as small as 0.005, $F_{2,9,\alpha}$ is still only 10.1. Alternatively, if the

analysis is done by computer, the p-value would be printed in the computer output. Here the p-value is less than 0.0001, and H_0 would be rejected for any choice of α above this value.

The experimenter was interested in estimating the contrasts $\tau_i - \tau_u$ for all $i \neq u$, that is, she was interested in comparing the effects on weight loss of the different types of soaps. For the one-way analysis of variance model (3.3.1) and a completely randomized design, all contrasts are estimable, and the least squares estimate of $\tau_i - \tau_u$ is

$$\hat{\tau}_i - \hat{\tau}_u = (\hat{\mu} + \hat{\tau}_i) - (\hat{\mu} + \hat{\tau}_u) = \overline{y}_{i.} - \overline{y}_{u.} \, .$$

Hence, the least square estimates of the differences in the treatment effects are

$$\hat{\tau}_2 - \hat{\tau}_3 = 0.7075 \, , \qquad \hat{\tau}_2 - \hat{\tau}_1 = 2.7350 \, , \qquad \hat{\tau}_3 - \hat{\tau}_1 = 2.0275 \, .$$

Confidence intervals for the differences will be evaluated in Example 4.4.5.

3.7.3   Discussion by the Experimenter

The results of this experiment were unexpected in that the soaps reacted with the water in very different ways, each according to its ingredients. An examination of the soap packages showed that for the deodorant soap and the moisturizing soap, water is listed as the third ingredient, whereas the regular soap claims to be 99.44% pure soap. Information on the chemical composition of soaps revealed that soaps are sodium and/or potassium salts of oleic, palmitic, and coconut oils and therefore in their pure form (without water) should float as the regular soap bars do. The other two soaps under discussion contain water and therefore are more dense and do not float.

One possible reason for the regular soap's actual increase in weight is that this "dry" soap absorbed and retained the water and dissolved to a lesser extent during the soaking period. The deodorant soap and the moisturizing soap, on the other hand, already contained water and did not absorb as much as the regular soap. They dissolved more easily during the soaking phase as a consequence. This is somewhat supported by the observation that the dissolved soap gel that formed extensively around the deodorant soap and the moisturizing soap did not form as much around the regular soap. Furthermore, the regular soap appeared to increase in size and remain larger, even at the end of the drying period.

3.7.4   Further Observations by the Experimenter

The soaps were weighed every day for one week after the experimental data had been collected in order to see what changes continued to occur. The regular soap eventually lost most of the water it retained, and the average loss of weight (due to dissolution) was less than that for the other two soaps.

If this study were repeated, with a drying period of at least one week, I believe that the results would indicate that regular soap loses less weight due to dissolution than either of the deodorant soap or the moisturizing soap

3.8   Using SAS Software

3.8.1   Randomization

A simple procedure for randomizing a completely randomized design was given in Sect. 3.2, p. 31. This procedure is easily implemented using the SAS software, as we now illustrate. Consider a completely randomized design for two treatments and $r = 3$ observations on each, giving a total of $n = 6$ observations. The following SAS statements create and print a data set named DESIGN, which includes the lists of values of the two variables TRTMT and RANNO as required by steps 1 and 2 of the randomization

procedure in Sect. 3.2. The statements INPUT and LINES are instructions to SAS that the values of TRTMT are being input on the lines that follow rather than from an external data file. Inclusion of "@@" in the INPUT statement allows the levels of TRTMT to be entered on one line as opposed to one per line. For each treatment label entered for the variable TRTMT, a corresponding value of RANNO is generated using the SAS random number generating function RANUNI which generates uniform random numbers between 0 and 1.

```
DATA DESIGN;
  INPUT TRTMT @@;
  RANNO=RANUNI(0);
  LINES;
  1 1 1 2 2 2
;
PROC PRINT; RUN;
```

The statement PROC PRINT then prints the following output. The column labeled OBS (observation number) is generated by the SAS software for reference.

```
The SAS System
Obs     TRTMT      RANNO
 1        1        0.74865
 2        1        0.62288
 3        1        0.87913
 4        2        0.32869
 5        2        0.47360
 6        2        0.72967
```

The following additional statements sort the data set by the values of RANNO, as required by step 3 of the randomization procedure, and print the randomized design along with the ordered experimental unit labels 1–6 under the heading OBS.

```
PROC SORT; BY RANNO;
PROC PRINT; RUN;
```

The resulting output is as follows.

```
The SAS System
Obs     TRTMT      RANNO
 1        2        0.32869
 2        2        0.47360
 3        1        0.62288
 4        2        0.72967
 5        1        0.74865
 6        1        0.87913
```

Experimental units 3, 5, and 6 are assigned to treatment 1, and experimental units 1, 2, and 4 are assigned to treatment 2.

Table 3.8 SAS program for the soap experiment

```
Line   SAS Program
  1    OPTIONS LINESIZE = 72;

  2    DATA SOAP;
  3      INPUT WTLOSS SOAP;
  4      LINES;
  5      -0.30 1
  6      -0.10 1
  7      -0.14 1
  8        :   :
  9       1.82 3
 10    ;

 11    PROC PRINT;

 12    PROC SGPLOT;
 13      SCATTER X = SOAP Y = WTLOSS;
 14      XAXIS TYPE = DISCRETE LABEL = 'Soap';
 15      YAXIS LABEL = 'Weight Loss (grams)';

 16    PROC GLM;
 17      CLASS SOAP;
 18      MODEL WTLOSS = SOAP;
 19      LSMEANS SOAP;

 20    RUN; QUIT;
```

3.8.2   Analysis of Variance

In this section we illustrate how SAS software can be used to conduct a one-way analysis of variance test for equality of the treatment effects, assuming that model (3.3.1) is appropriate. We use the data in Table 2.7, p. 23, from the soap experiment.

A sample SAS program to analyze the data is given in Table 3.8. Line numbers have been included for reference, but the line numbers are not part of the SAS program and if included would cause SAS software to generate error messages. SAS programs and data files for this edition are available at the following website.

http://www.wright.edu/~dan.voss/DeanVossDraguljic.html

The option LINESIZE = 72 in the OPTIONS statement in line 1 of the program causes all list output generated by the program to be restricted to 72 characters per line, which is convenient for printing list output on 8.5 by 11 inch paper in the portrait orientation. This option has no effect on the standard html output, however, so can be ignored by readers running the SAS software in a windows environment, which is most likely to be the norm. Some of our SAS programs, including those using PROC SGPLOT for example, assume the user is running SAS in a windows environment, whereas PROC PLOT might be used instead if running SAS in a command line mode. All SAS statements are ended by a semicolon.

Lines 2–10 of the program create a SAS data set named SOAP that includes as variables the response variable WTLOSS and the corresponding level of the treatment factor SOAP. The LINES statement indicates that subsequent lines contain data to be read directly from the program, until data entry is

Fig. 3.3 Data plot from
the SAS software for the
soap experiment
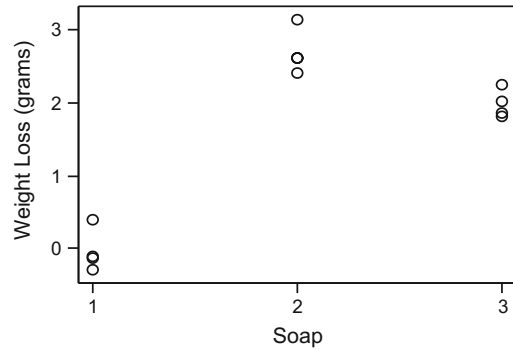
stopped by the next semicolon (line 10). Line 8 must be replaced by the additional data not shown here.

Alternatively, the same data could be read from a file, `soap.txt` say, by replacing lines 2–10 with the following code, including a correct file path.

```
INFILE 'c:\path\soap.txt' FIRSTOBS = 2; INPUT WTLOSS SOAP;
```

Include the option `FIRSTOBS = 2` if the data file contains headers on line one then data starting on line two, but delete it if the data starts on line one with no headers.

The `PRINT` procedure (line 11) prints the data. While this is good practice to verify that the data were read correctly, the `PRINT` procedure will not routinely be shown in subsequent programs.

The `SGPLOT` procedure (lines 12–15) generates a scatterplot of `WTLOSS` versus `SOAP` like that shown in Fig. 3.3. The *x*-axis option `TYPE = DISCRETE` instructs the SAS software to use integer values for *x*-axis tick marks. The `LABEL` option sets the desired label for each axis.
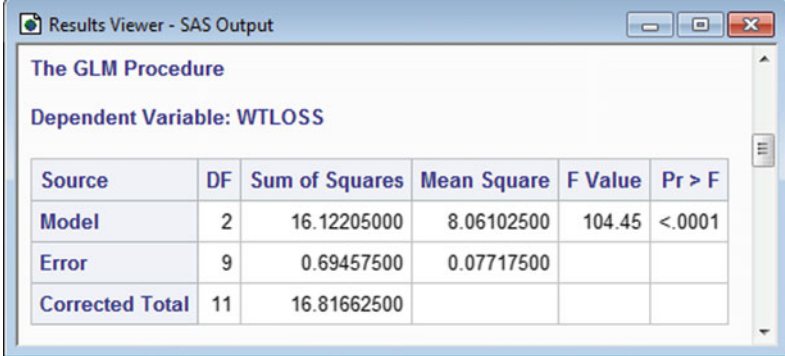
The resulting scatterplot is displayed in a SAS output window, by default (on a PC). Alternatively, one could redirect the scatterplot to be saved in the file `ch3soap.pdf` in pdf format, for example, as illustrated by the following code.

```
ODS GRAPHICS / RESET IMAGENAME ='ch3soap' IMAGEFMT = PDF
               HEIGHT = 1.5in WIDTH = 2in;
ODS LISTING GPATH ='c:\path\figs';
* insert PROC SGPLOT and its subcommands here;
RUN; * Run PROC SGPLOT before closing output to pdf file;
ODS GRAPHICS / RESET;
```

The statements beginning "*" and ending ";" are comments which are not executed by the SAS software. The first `ODS GRAPHICS` command redirects graphics output to the file `ch3soap.pdf` using pdf as the image format, and specifies the dimension of the graphic image to be saved. The `ODS LISTING` command then specifies the directory where the SAS software is to store the file, so the user must replace "`c:\path\figs`" with an existing path and directory on the user's computer. Following `PROC SGPLOT` and its statements, the second `ODS GRAPHICS` command resets the graphics defaults, so graphics output reverts again to the SAS output window. However, before doing so, the `RUN` command causes `SGPLOT` to execute while output is still directed to the file `ch3soap.pdf`.

The General Linear Models procedure `PROC GLM` (lines 16–19) generates an analysis of variance table. The `CLASS` statement identifies `SOAP` as a major source of variation to be modeled as a *classification variable*, so a parameter is associated with each of its levels. The `MODEL` statement defines the response variable as `WTLOSS`, and the only source of variation included in the model is `SOAP`. The
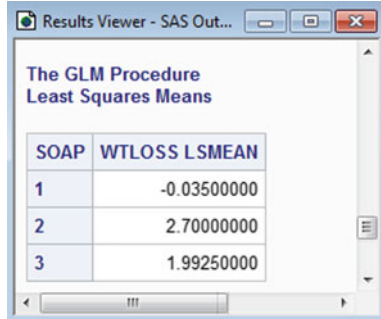
Fig. 3.4 Sample SAS
output from PROC GLM
for the soap experiment

Fig. 3.5 Output from the
LSMEANS statement for
the soap experiment

parameter μ and the error variables are automatically included in the model. The MODEL statement causes the analysis of variance table shown in Fig. 3.4 to be calculated. The F Value is the value of the ratio *msT/msE* for testing the null hypothesis that the three treatment effects are all equal. The value Pr > F is the *p*-value of the test to be compared with the chosen significance level. When the *p*-value is listed as <.0001, the null hypothesis would be rejected for any chosen significance level larger than 0.0001.

The LSMEANS statement (line 19 of Table 3.8), which is part of the GLM procedure, causes the least squares means, $\hat{\mu} + \hat{\tau}_i = \bar{y}_{i.}$, to be printed. The output from this statement is shown in Fig. 3.5.

The RUN statement in line 20 is needed to cause the last procedure to be executed when the program is run in an interactive line mode, typical of a PC Windows environment for example, and the QUIT statement ends the procedure. Though necessary for interactive program processing, the RUN and QUIT statements will not be shown from now on in any programs.

3.8.3 Calculating Sample Size Using Power of a Test

In Table 3.9, we show a sample SAS program which calculates the power of the test of the null hypothesis $H_0 : \{\tau_1 = \cdots = \tau_v\}$ against H_A: {at least two of the τ_i's differ}. The program uses a DO statement, which allows the calculation to be done for a selected range of sample sizes r, using the formulae in Sect. 3.6.2. The line DO R = 3 TO 6 BY 1; asks the SAS software to do the calculations for each value of r between 3 and 6, increasing r by 1 each time.

The code shown is for the soap experiment in Example 3.6.1, but is easily modified for other experiments by changing the values of the number of levels of the treatment factor (V), the difference

Table 3.9 Calculating sample sizes using power of the test

```
DATA POWER;
  V = 3;
  DEL = 0.25;
  SIGMA2 = 0.007;
  ALPHA = 0.05;
  NU1 = V - 1;
  LHTPB = 1 - ALPHA;

DO R = 3 TO 6 BY 1;
  NU2 = V*(R - 1);
  PHI = (SQRT(R / (2*V*SIGMA2)))*DEL;
  FVALUE = FINV(LHTPB, NU1, NU2);
  NONCN = V*PHI**2;
  POWER = 1 - PROBF(FVALUE, NU1, NU2, NONCN);
  OUTPUT;
END;

PROC PRINT;
  VAR R POWER;
```

(DEL) to be detected (i.e. Δ), the assumed largest value of the error variance (SIGMA2), the significance level of the test (ALPHA), and the range of values of r to be investigated.

In Table 3.9, the degrees of freedom $\nu_1 = v - 1$ and $\nu_2 = n - v = v(r - 1)$ for the F-distribution are donated by NU1 and NU2. The "left-hand tail probability" $1 - \alpha$ is called LHTPB, and is used in calculating the critical value $F_{v-1,n-v,\alpha}$, called FVALUE. From (3.6.20), the value of ϕ, labelled PHI is calculated as $\sqrt{r \Delta^2 / (2v\sigma^2)}$. The "non-centrality parameter" NONCN is $\delta^2 = v\phi^2$ and this is needed by the non-central F distribution in the calculation of the power for the range of values of r specified. The output, generated by the PROC PRINT statement is

```
Obs     R      POWER
 2      3     0.70934
 3      4     0.89565
 4      5     0.96715
 5      6     0.99058
```

and we can see that, just as in Example 3.6.1, to achieve a power of approximately 0.9, the number of observations needed is $r = 4$ per level of the treatment factor. To see the values of all the variables calculated at each step, remove the line VAR R POWER; which restricts which variables are printed.

3.9 Using R Software

Preliminaries

R is a free software environment for statistical computing and graphics, used extensively in this book. Readers can install the R software after downloading it from http://cran.us.r-project.org, for example, choosing the appropriate version for the computer and operating system. RStudio is free software providing an enhanced environment for running R. After installing R, readers are recommended to also download and install RStudio; it can be downloaded from http://www.rstudio.com, choosing

again the appropriate version for the computer and operating system. Run either R or RStudio, as RStudio invokes R.

Throughout the book, we shall assume that the reader has set up a working directory for R called `RCode` and that R program files are either in the working directory or in specified subdirectories of the working directory. For example, we assume that RCode includes a subdirectory called `data` containing any data set that is to be read by R and a subdirectory called `figs` in which R will store any plots generated. R programs and data files for this edition are available at the following website.

http://www.wright.edu/~dan.voss/DeanVossDraguljic.html

We shall assume that the user will execute the following commands or similar each time that R (or RStudio) is started.

```
rm(list = ls())   # Remove all objects (start with clean slate)
opar = par()      # Save default graphics parameters as opar
setwd("/RCode")   # Set the working directory
getwd()           # Confirm working directory
options(show.signif.stars = FALSE) # Show no stars for significance tests
options(width = 72, digits = 5, scipen = 2) # Control printed output
ooptions = options(width = 72, digits = 5, scipen = 2) # Save print options
```

The first command removes all existing R objects created previously by the user, clearing the slate for a new session. The symbol "#" starts a line comment, used for program documentation. The second command assigns R's current graphics parameters `par()`—initially the default graphics parameter values—to the object `opar`, saving them so they can be restored later via the command `par(opar)` if desired. We will routinely use "=" for assignment, though use of "<−" is more traditional in R. The `setwd` command in the third line sets `/RCode` as the *working directory*, where the software reads and writes files by default. If RCode is not in the root directory, then use `setwd("/`*path*`/RCode")` but specify the correct directory path to RCode. The `getwd()` command in the fourth line displays the working directory, to confirm it is now `/RCode`. For functions that conduct hypothesis tests, the `options` command in the fifth line suppresses printing of stars for various levels of significance. The `options` command in line six controls printed output, restricting it to be at most 72 columns wide with five significant digits, and penalizing use of scientific notation. We have initialized R in this way or similarly when running our programs, though these commands will generally not be shown in our subsequent program code. The last line saves these print options as `ooptions`, so if changed they can be restored by the command `options(ooptions)`.

While the above commands can be typed into the R Console and executed one by one, it is more convenient to save them in a file, `startup.r` say, in the working directory. Then the single command

```
source("/RCode/startup.r")
```

will execute the commands in the `startup.r` file. We routinely executed this code line each time we started R (or RStudio) to produce R program output in this book, though we do not show this code line in our programs.

As will be seen in the following sections, when R is waiting for the next command, a prompt > is displayed, and if the user command is not complete when a line is entered (for example if the final parenthesis is missing), the prompt will change to + on the next line, prodding the user to enter the rest of the command. To end an R session, type `q()`.

It is prudent to use the latest production version of R. On a Windows operating system, the `updateR` command of the `installr` package will detect if there is a new R version available, and if so it will download and install it and update previously installed add-on packages. The following commands, when executed from within R, install and load the `installr` package and execute the `updateR` command.

```
install.packages("installr"); library(installr); updateR()
```

Assuming the user has installed and set up R as noted above, we are ready to use the software.

3.9.1   Randomization

A simple procedure for randomizing a completely randomized design was given in Sect. 3.2, p. 31. This procedure is easily implemented using the R software, as we now illustrate. Consider a completely randomized design for two treatments and $r = 3$ observations on each, giving a total of $n = 6$ observations.

The following R statements create and display a data frame named design, a *data frame* being a data set consisting of equal-length columns of information. Here, the columns of design are the lists of values of the two variables trtmt and ranno as required by steps 1 and 2 of the randomization procedure in Sect. 3.2. In particular, the first statement creates the column trtmt of treatment labels. The second statement creates the column ranno consisting of six uniform random numbers between 0 and 1, six being the length of the column trtmt. The third statement puts the columns trtmt and ranno into a data frame, and assigns this object the name design. Then the fourth statement displays design. The R output is shown immediately following the R statements.

```
> trtmt = c(1, 1, 1, 2, 2, 2)  # Create column trtmt = (1, 1, 1, 2, 2, 2)
> ranno = runif(length(trtmt))  # Create column of 6 unif(0, 1) RVs
> design = data.frame(trtmt, ranno) # Create data.frame "design"
> design  # Display the data.frame design

   trtmt     ranno
1      1 0.447827
2      1 0.494462
3      1 0.174414
4      2 0.894132
5      2 0.473540
6      2 0.010771
```

We digress to provide additional information about R, before finishing the randomization process in the next paragraph. The command trtmt = rep(c(1, 2), each = 3) would yield the same column trtmt, but by replicating 1 and 2 three times each—a more convenient approach for larger designs. Each time the above R commands are run, a different set of random numbers will result. Typing the command design causes the entire data frame design to be displayed. One could display only what is in column 1 named trtmt, for example, by typing design$trtmt, design[, 1], or design[, "trtmt"]. The column trtmt also still exists as a separate object, that can be displayed simply by typing trtmt. One could remove this redundant object by the command rm(trtmt). Note that R is "case-sensitive", so if the column name is trtmt, then R will not be able to locate a column called, say, Trtmt with a capital T. The details about any of the commands used can be found by typing ?commandName, for example typing ?runif will bring up the command help file containing many details about the use of runif.

The following additional statements sort the trtmt column of the data frame design by the values of ranno, as required by step 3 of the randomization procedure. Specifically, the statement order(ranno) yields the order 6 3 1 5 2 4, since the smallest random variate is in row 6, the second smallest is in row 3, etc. So, the first statement below redefines the data frame design to have its rows reordered accordingly, effectively sorting the rows based on the values of the random numbers (RNs). The second statement below defines a new column of design named EU containing the integers from

Table 3.10 R program for the soap experiment: reading and plotting data

```
Line   R Code and Output
   1   > # Read the data into the data.frame "soap.data"
   2   > soap.data = read.table("data/soap.txt", header = TRUE)
   3   > head(soap.data, 5)  # Display first 5 lines of soap.data

   4       Soap Cube PreWt PostWt WtLoss
   5   1    1    1  13.14  13.44  -0.30
   6   2    1    2  13.17  13.27  -0.10
   7   3    1    3  13.17  13.31  -0.14
   8   4    1    4  13.17  12.77   0.40
   9   5    2    5  13.03  10.40   2.63

  10   > # Add factor variable fSoap to soap.data for later ANOVA
  11   > soap.data$fSoap = factor(soap.data$Soap)

  12   > # Plot WtLoss vs Soap, specify axis labels, suppress x-axis.
  13   > plot(WtLoss ~ Soap, data = soap.data, xlab = "Soap",
  14   +        ylab = "Weight Loss (grams)", las = 1, xaxt = "n")
  15   > # Insert x-axis (axis 1) with tick marks from 1 to 3 by 1.
  16   > axis(1, at = seq(1,3,1))
```

1 to 6 as labels for the experimental units. Then the last statement asks for the sorted design to be displayed.

```
> design = design[order(ranno), ]  # Sort rows by RNs, save
> design$EU = c(1:6)  # Add col EU = (1,2,3,4,5,6) to design
> design  # Display the results of the randomization

  trtmt     ranno EU
6     2 0.010771  1
3     1 0.174414  2
1     1 0.447827  3
5     2 0.473540  4
2     1 0.494462  5
4     2 0.894132  6
```

Experimental units 2, 3, and 5 are to be assigned to treatment 1, and experimental units 1, 4, and 6 are to be assigned to treatment 2.

3.9.2   Reading and Plotting Data

A sample R program to input, display, and plot the data is given in Table 3.10. Line numbers have been included for reference, but they are not part of the R program and if included in the R code would yield error messages. The prompt ">" and the continuation prompt "+" are supplied by R and should not be typed by the user.

We use the data in Table 2.7, p. 23, from the soap experiment, and assume that the data are stored in the file soap.txt in the data subdirectory of the working directory; that is, in data/soap.txt.

Line 2 of Table 3.10 reads the data from the file `soap.txt` and puts it into an R data set (data frame) called `soap.data`. Alternatively, one could enter the data via the keyboard, which is the standard input device `stdin()`, by replacing line 2 with the following.

```
> soap.data = read.table(stdin(), header = TRUE)
0: Soap Cube PreWt PostWt WtLoss
1:    1    1 13.14  13.44   -0.30
2:    1    2 13.17  13.27   -0.10
 :    :    :    :       :        :
12:   3   12 13.00  11.18    1.82
13:
```

Keyboard data entry is ended by hitting the return key twice.

The `head(soap.data, 5)` command in line 3 displays the first five lines of the data set shown in lines 4–9. Alternatively, the command `head(soap.data)` would display the first six lines by default, and the command `soap.data` would display the full data set. The first column displayed indicates the data is from rows 1–5 of the data set, and the other five columns show the five variables in the data set, including the response variable `WtLoss` and the corresponding level of the treatment factor `Soap`. In line 2, the statement `header = TRUE` (i.e. `header = T`) tells R that the columns of data in the file `soap.txt` have headings, and these can be seen in line 4. If a data file has no headings, the header statement may be omitted, the default being `header = FALSE` (i.e. `header = F`).

In line 11, the data set `soap.data` is augmented with a new variable, `fSoap`, created by converting the numerical variable `Soap` to a factor variable, needed later for the analysis of variance.

The remaining code generates a plot of the data. The `plot` command in lines 13–14 generates a scatterplot of `WtLoss` versus `Soap` like that shown in Fig. 3.6, with labels specified for each axis by `xlab` and `ylab`. The option `data = soap.data` indicates that the variables being plotted are in the data set `soap.data`. Alternatively, one could use the syntax

```
plot(soap.data$WtLoss ~ soap.data$Soap, xlab ="Soap",
```

in line 13. The dollar sign identifies specific columns of the data set, so, for example, `soap.data$Soap` just means "read the column labeled `Soap` from the data set `soap.data`". The + prompt in line 14 indicates that the prior command is not yet complete. The option `las = 1` sets *labels style* 1, making tick mark labels horizontal, impacting *y*-axis labeling. The option `xaxt = "n"` in line 14 suppresses the automatically generated x-axis (which would have five tick marks), then the `axis` command in line 16 includes instead an x-axis with three tick marks specified to be at 1, 2 and 3—namely, a sequence starting at 1 and ending at 3 in steps of size 1. If the numerical variable `Soap` was replaced by the factor variable `fSoap` in line 13, then a box plot would be obtained instead of a scatterplot.

The resulting scatterplot is displayed in a graphics window, by default (on a PC). Alternatively, one can redirect the scatterplot to be saved in the file `ch3soap.pdf` in pdf format, for example, as illustrated by the following code.

```
pdf("figs/ch3soapplot.pdf", width = 5, height = 3)  # Open a pdf file
# Insert plot command and its subcommands here
dev.off()  # Close the pdf file
```

The `pdf` command opens the pdf file `ch3soapplot.pdf` in the `figs` subdirectory of the working directory (as specified by the user upon startup), and specifies the dimension of the graphic image to be saved. Once this has been done, `plot` function calls will send output to the pdf file.

Fig. 3.6 R data plot for the
soap experiment
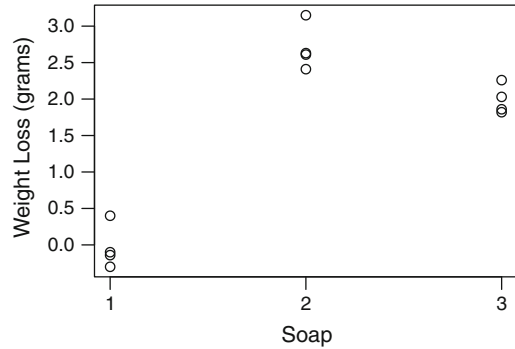
Then the dev.off() command closes the pdf file, so graphical output reverts to the default graphics
window.

3.9.3 Analysis of Variance

In this section we illustrate how the R software can be used to conduct a one-way analysis of variance
test for equality of the treatment effects, assuming model (3.3.1) is appropriate. Table 3.11 contains a
sample program and output, with line numbers again included for reference only. We continue to use the
soap experiment data, which in line 1 is read from the file soap.txt into the data set soap.data.

In line 3, we convert the numerical variable Soap to a factor variable, saving it as a new variable
fSoap of the soap.data data set. The statement in line 4 generates summary statistics for the
variables in columns 1, 5 and 6 of the data set, with the output shown in lines 5–11. If all six variables
are to be summarized, the statement summary(soap.data) without column numbers is sufficient.
One can see that the summary command treats the numeric variable Soap and the factor variable
fSoap differently, providing summary statistics for the Soap values, but levels and frequencies for
fSoap. A *factor variable* is treated as a qualitative variable by R, analogous to a CLASS variable in
SAS software.

The aov function in line 12 fits a linear model to the soap data, specifying WtLoss as the response
variable and fSoap as the primary source of variation, the symbol "∼" separating and distinguishing
these, saving the resulting information as the object model1. Because fSoap is a factor variable, it is
modeled as a classification variable as desired. The parameter μ and the error variables are automatically
included in the model. The anova(model1) command in line 13 displays the one-way analysis of
variance information shown in lines 14–19. In lines 17–18, the F value is the value of the ratio
msT/msE for testing the null hypothesis that the three treatment effects are all equal, and Pr(>F)
is the *p*-value of the test to be compared with the chosen significance level. The listed *p*-value is
5.91×10^{-7}, so the null hypothesis is rejected for any chosen significance level larger than this small
value.

The library(lsmeans) command in line 21 loads the package lsmeans from the user's
library for subsequent use; (this assumes the reader has already installed the package as discussed in
the next paragraph). The lsmeans command in line 22 generates the least squares means for the three
levels of the factor fSoap, plus further statistical information which will be discussed in Chap. 4. The
output is shown in lines 23–28.

Table 3.11 R program for the soap experiment: analysis of variance and least squares means

```
Line  R Code or Output
  1   > soap.data = read.table("data/soap.txt", header = TRUE)
  2   > # Add factor variable fSoap to soap.data for ANOVA
  3   > soap.data$fSoap = factor(soap.data$Soap)
  4   > summary(soap.data[,c(1,5:6)]) # Summarize data in cols 1, 5, 6

  5        Soap         WtLoss         fSoap
  6    Min.   :1   Min.   :-0.300   1:4
  7    1st Qu.:1   1st Qu.: 0.275   2:4
  8    Median :2   Median : 1.945   3:4
  9    Mean   :2   Mean   : 1.552
 10    3rd Qu.:3   3rd Qu.: 2.460
 11    Max.   :3   Max.   : 3.150

 12   > model1 = aov(WtLoss ~ fSoap, data = soap.data)
 13   > anova(model1)

 14   Analysis of Variance Table
 15
 16   Response: WtLoss
 17             Df Sum Sq Mean Sq F value  Pr(>F)
 18   fSoap      2  16.12    8.06     104 5.9e-07
 19   Residuals  9   0.69    0.08

 20   > # install.packages("lsmeans")
 21   > library(lsmeans)
 22   > lsmeans(model1, "fSoap")

 23    fSoap  lsmean      SE df lower.CL upper.CL
 24    1     -0.0350 0.1389  9 -0.34922  0.27922
 25    2      2.7000 0.1389  9  2.38578  3.01422
 26    3      1.9925 0.1389  9  1.67828  2.30672
 27
 28    Confidence level used: 0.95
```

Add-On Packages

Installation of the R software (see Sect. 1.2) installs the base software, including some base packages providing limited functionality. There are thousands of additional user-defined packages that the user may freely download, the lsmeans package introduced above being one example. To use any such function not included in the base software installation, the "add-on" package containing the function must first be installed and loaded. For example, the command install.packages("lsmeans") in line 20 installs the lsmeans package, permanently saving it in a library of packages on the user's computer, so the command library(lsmeans) can load the lsmeans package from the user's library. A package only needs to be installed once. However, any add-on package must be loaded by the user prior to its first use in any new R session. As such, when our programs require an add-on package, we will routinely include the necessary library command to load the package. Furthermore, when we use an add-on package for the first time, the corresponding program will include the corresponding install.packages command, but commented out. Before running such a program the first time,

Table 3.12 Calculating sample sizes using power of the test

```
> #install.packages(pwr)
> library(pwr)
> v = 3; del = 0.25; sig2 = 0.007; alpha = 0.05; pwr = 0.90
> pwr.anova.test(k = v, sig.level = alpha, power = pwr,
+                  f = sqrt(del^2/(2*v*sig2)))

     Balanced one-way analysis of variance power calculation

             k = 3
             n = 4.038656
             f = 1.219875
     sig.level = 0.05
         power = 0.9

NOTE: n is number in each group
```

the reader can simply delete the comment character, "#", so the package gets installed. On most systems, the process is automatic.

For linux users, when the `install.packages` command is invoked for the first time, you may get a warning that says in effect that a library is not writable and it will ask you whether you would like to use a personal library. If you answer `y`, it will ask you if you would like to create one. If you answer `y` again, it will ask you to select a CRAN "mirror" (i.e. a site from which to download the package for installation). Select (give the number of) any site near your location, and then R will create the personal library and download the package. After this, the `install.packages` command will proceed automatically.

3.9.4   Calculating Sample Size Using Power of a Test

In Table 3.12, we show a sample R program which calculates the power of the test of the null hypothesis $H_0 : \{\tau_1 = \cdots = \tau_v\}$ against H_A: {at least two of the τ_i's differ}. The code shown is for the soap experiment in Example 3.6.1, but is easily modified for other experiments by changing the values of the number of levels of the treatment factor (`v`), the difference (`del`) to be detected (i.e. Δ), the assumed largest value of the error variance (`sig2`), the significance level of the test (`alpha`), and the desired power of the test `pwr`.

The R program from Table 3.12 uses function `pwr.anova.test` which can be found in the package `pwr`. The function's inputs `k = v`, `sig.level = alpha`, and `power = pwr` are self-explanatory. The degrees of freedom $\nu_1 = v - 1$, $\nu_2 = n - v = v(r - 1)$, and the critical value $F_{v-1,n-v,\alpha}$ for the F-distribution corresponding to the required significance level are calculated internally by R. From (3.6.20), the value of $\phi/\sqrt{r}$, labeled as input `f`, is calculated as $\sqrt{\Delta^2/(2v\sigma^2)}$ and is needed by the non-central F distribution in the calculation of the power for different values of r. The output is also shown in Table 3.12. and we can see that, just as in Example 3.6.1, to achieve a power of approximately 0.9, the number of observations needed is $r = 4$ per level of the treatment factor. Notice that r is labelled `n` in the R output.

Exercises

1. Suppose that you are planning to run an experiment with one treatment factor having four levels and no blocking factors. Suppose that the calculation of the required number of observations has given $r_1 = r_2 = r_3 = r_4 = 5$. Assign at random 20 experimental units to the $v = 4$ levels of the treatments, so that each treatment is assigned 5 units.

2. Suppose that you are planning to run an experiment with one treatment factor having three levels and no blocking factors. It has been determined that $r_1 = 3$, $r_2 = r_3 = 5$. Assign at random 13 experimental units to the $v = 3$ treatments, so that the first treatment is assigned 3 units and the other two treatments are each assigned 5 units.

3. Suppose that you are planning to run an experiment with three treatment factors, where the first factor has two levels and the other two factors have three levels each. Write out the coded form of the 18 treatment combinations. Assign 36 experimental units at random to the treatment combinations so that each treatment combination is assigned two units.

4. For the one-way analysis of variance model (3.3.1), p. 33, the solution to the normal equations used by the SAS software is $\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{v.}$ $(i = 1, \ldots, v)$ and $\hat{\mu} = \bar{y}_{v.}$.

   (a) Is τ_i estimable? Explain.
   (b) Calculate the expected value of the least squares estimator for $\tau_1 - \tau_2$ corresponding to the above solution. Is $\tau_1 - \tau_2$ estimable? Explain.

5. Consider a completely randomized design with observations on three treatments (coded 1, 2, 3). For the one-way analysis of variance model (3.3.1), p. 33, determine which of the following are estimable. For those that are estimable, state the least squares estimator.

   (a) $\tau_1 + \tau_2 - 2\tau_3$.
   (b) $\mu + \tau_3$.
   (c) $\tau_1 - \tau_2 - \tau_3$.
   (d) $\mu + (\tau_1 + \tau_2 + \tau_3)/3$.

6. (requires calculus) Show that the normal equations for estimating $\mu, \tau_1, \ldots, \tau_v$ are those given in Eq. (3.4.3) on p. 35.

7. (requires calculus) Show that the least squares estimator of $\mu + \tau$ is $\overline{Y}_{..}$ for the linear model $Y_{it} = \mu + \tau + \epsilon_{it}^0$ $(t = 1, \ldots, r_i; i = 1, 2, \ldots, v)$, where the ϵ_{it}^0's are independent random variables with mean zero and variance σ^2. (This is the reduced model for the one-way analysis of variance test, Sect. 3.5.1, p. 41.)

8. For the model in the previous exercise, find an unbiased estimator for σ^2. (Hint: first calculate $E[ssE_0]$ in (3.5.10), p. 42.)

9. (requires calculus) Find the least squares estimates of $\mu_1, \mu_2, \ldots, \mu_v$ for the linear model $Y_{it} = \mu_i + \epsilon_{it}$ $(t = 1, \ldots, r_i; i = 1, 2, \ldots, v)$, where the ϵ_{it}'s are independent random variables with mean zero and variance σ^2. Compare these estimates with the least squares estimates of $\mu + \tau_i$ $(i = 1, 2, \ldots, v)$ in model (3.3.1), p. 33.

10. For the model in the previous exercise, find an unbiased estimator for σ^2. Compare the estimator with that in (3.4.7), p. 39.

11. Verify, for the one-way analysis of variance model (3.3.1), p. 33, that each treatment sample variance S_i^2 is an unbiased estimator of the error variance σ^2, so that

$$E(SSE) = \sum_i (r_i - 1)E(S_i^2) = (n - v)\sigma^2.$$

Table 3.13   Times (in seconds) for the balloon experiment

Time order	1	2	3	4	5	6	7	8
Coded color	1	3	1	4	3	2	2	2
Inflation time	22.0	24.6	20.3	19.8	24.3	22.2	28.5	25.7
Time order	9	10	11	12	13	14	15	16
Coded color	3	1	2	4	4	4	3	1
Inflation time	20.2	19.6	28.8	24.0	17.1	19.3	24.2	15.8
Time order	17	18	19	20	21	22	23	24
Coded color	2	1	4	3	1	4	4	2
Inflation time	18.3	17.5	18.7	22.9	16.3	14.0	16.6	18.1
Time order	25	26	27	28	29	30	31	32
Coded color	2	4	2	3	3	1	1	3
Inflation time	18.9	16.0	20.1	22.5	16.0	19.3	15.9	20.3

12. **Balloon experiment**

Prior to 1985, the experimenter (Meily Lin) had observed that some colors of birthday balloons seem to be harder to inflate than others. She ran this experiment to determine whether balloons of different colors are similar in terms of the time taken for inflation to a diameter of 7 inches. Four colors were selected from a single manufacturer. An assistant blew up the balloons and the experimenter recorded the times (to the nearest 1/10 second) with a stop watch. The data, in the order collected, are given in Table 3.13, where the codes 1, 2, 3, 4 denote the colors pink, yellow, orange, blue, respectively.

(a) Plot inflation time versus color and comment on the results.
(b) Estimate the mean inflation time for each balloon color, and add these estimates to the plot from part (a).
(c) Construct an analysis of variance table and test the hypothesis that color has no effect on inflation time.
(d) Plot the data for each color in the order that it was collected. Are you concerned that the assumptions on the model are not satisfied? If so, why? If not, why not?
(e) Is the analysis conducted in part (c) satisfactory?

13. **Heart–lung pump experiment, continued**

The heart–lung pump experiment was described in Example 3.4.1, p. 37, and the data were shown in Table 3.2, p. 38.

(a) Calculate an analysis of variance table and test the null hypothesis that the different number of revolutions per minute have the same effects on the fluid flow rate.
(b) Are you happy with your conclusion? Why or why not?
(c) Calculate a 90% upper confidence limit for the error variance σ^2.

14. **Meat cooking experiment**

(L. Alvarez, M. Burke, R. Chow, S. Lopez, and C. Shirk, 1998)

Table 3.14 Post-cooking weight data (in grams) for the meat cooking experiment

Frying			Grilling		
Fat content			Fat content		
10%	15%	20%	10%	15%	20%
81	85	71	84	83	78
88	80	77	84	88	75
85	82	72	82	85	78
84	80	80	81	86	79
84	82	80	86	88	82

Table 3.15 Data for the trout experiment

Code	Hemoglobin (grams per 100 ml)									
1	6.7	7.8	5.5	8.4	7.0	7.8	8.6	7.4	5.8	7.0
2	9.9	8.4	10.4	9.3	10.7	11.9	7.1	6.4	8.6	10.6
3	10.4	8.1	10.6	8.7	10.7	9.1	8.8	8.1	7.8	8.0
4	9.3	9.3	7.2	7.8	9.3	10.2	8.7	8.6	9.3	7.2

Source: Gutsell (1951). Copyright © 1951 International Biometric Society. Reprinted with permission

An experiment was run to investigate the amount of weight lost (in grams) by ground beef hamburgers after grilling or frying, and how much the weight loss is affected by the percentage fat in the beef before cooking. The experiment involved two factors: cooking method (factor A, with two levels frying and grilling, coded 1, 2), and fat content (factor B, with three levels 10, 15, and 20%, coded 1, 2, 3). Thus there were six treatment combinations 11, 12, 13, 21, 22, 23, relabeled as treatment levels 1, 2, …, 6, respectively. Hamburger patties weighing 110 g each were prepared from meat with the required fat content. There were 30 "cooking time slots" which were randomly assigned to the treatments in such a way that each treatment was observed five times ($r = 5$). The patty weights after cooking are shown in Table 3.14.

(a) Plot the data and comment on the results.
(b) Write down a suitable model for this experiment.
(c) Calculate the least squares estimate of the mean response for each treatment. Show these estimates on the plot obtained in part (a).
(d) Test the null hypothesis that the treatments have the same effect on patty post-cooking weight.
(e) Estimate the contrast $\tau_1 - (\tau_2 + \tau_3)/2$ which compares the effect on the post-cooked weight of the average of the two higher fat contents versus the leanest meat for the fried hamburger patties.
(f) Calculate the variance associated with the contrast in part (e). How does the value of the variance compare with the variance σ^2 of the random error variables?

15. **Trout experiment** (Gutsell 1951, Biometrics)

The data in Table 3.15 show the measurements of hemoglobin (grams per 100 ml) in the blood of brown trout. The trout were placed at random in four different troughs. The fish food added to the troughs contained, respectively, 0, 5, 10, and 15 g of sulfamerazine per 100 pounds of fish (coded 1, 2, 3, 4). The measurements were made on ten randomly selected fish from each trough after 35 days.

(a) Plot the data and comment on the results.
(b) Write down a suitable model for this experiment, assuming trough effects are negligible.
(c) Calculate the least squares estimate of the mean response for each treatment. Show these estimates on the plot obtained in (a). Can you draw any conclusions from these estimates?
(d) Test the hypothesis that sulfamerazine has no effect on the hemoglobin content of trout blood.
(e) Calculate a 95% upper confidence limit for σ^2.

16. **Trout experiment, continued**

Suppose the trout experiment of Exercise 3.15 is to be repeated with the same $v = 4$ treatments, and suppose that the same hypothesis, that the treatments have no effect on hemoglobin content, is to be tested.

(a) For calculating the number of observations needed on each treatment, what would you use as a guess for σ^2?
(b) Calculate the sample sizes needed for an analysis of variance test with $\alpha = 0.05$ to have power 0.95 if: (i) $\Delta = 1.5$; (ii) $\Delta = 1.0$; (iii) $\Delta = 2.0$.

17. **Meat cooking experiment, continued**

Suppose the meat cooking experiment of Exercise 3.14 is to be repeated with the same $v = 6$ treatments, and suppose the same hypothesis, that the treatments have the same effect on burger patty weight loss, is to be tested.

(a) Calculate an unbiased estimate of σ^2 and a 90% upper confidence limit for it.
(b) Calculate the sample sizes needed for an analysis of variance test with $\alpha = 0.05$ to have power 0.90 if:
(i) $\Delta = 5.0$; (ii) $\Delta = 10.0$.

18. The diameter of a ball bearing is to be measured using three different calipers. How many observations should be taken on each caliper type if the null hypothesis H_0:{effects of the calipers are the same} is to be tested against the alternative hypothesis that the three calipers give different average measurements. It is required to detect a difference of 0.01 mm in the effects of the caliper types with probability 0.98 and a Type I error probability of $\alpha = 0.05$. It is thought that σ is about 0.03 mm.

19. An experiment is to be run to determine whether or not time differences in performing a simple manual task are caused by different types of lighting. Five levels of lighting are selected ranging from dim colored light to bright white light. The one-way analysis of variance model (3.3.1), p. 33 is thought to be a suitable model, and H_0 : {$\tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5$} is to be tested against the alternative hypothesis H_A:{the τ_i's are not all equal} at significance level 0.05. How many observations should be taken at each light level given that the experimenter wishes to reject H_0 with probability 0.90 if the difference in the effects of any two light levels produces a 4.5-second time difference in the task? It is thought that σ is at most 3.0 seconds.

Inferences for Contrasts and Treatment Means

<div align="right">**4**</div>

4.1 Introduction

The objective of an experiment is often much more specific than merely determining whether or not all of the treatments give rise to similar responses. For example, a chemical experiment might be run primarily to determine whether or not the yield of the chemical process increases as the amount of the catalyst is increased. A medical experiment might be concerned with the efficacy of each of several new drugs as compared with a standard drug. A nutrition experiment may be run to compare high fiber diets with low fiber diets. Such treatment comparisons are formalized in Sect. 4.2. The purpose of this chapter is to provide confidence intervals and hypothesis tests about treatment comparisons and treatment means. We start, in Sect. 4.3, by considering a single treatment comparison or mean, and then, in Sect. 4.4, we develop the techniques needed when more than one treatment comparison or mean is of interest. The number of observations required to achieve confidence intervals of given lengths is calculated in Sect. 4.5. SAS and R commands for confidence intervals and hypothesis tests are provided in Sects. 4.6 and 4.7, respectively.

4.2 Contrasts

In Chap. 3, we defined a contrast to be a linear combination of the parameters $\tau_1, \tau_2, \ldots, \tau_v$ of the form

$$\sum c_i \tau_i, \quad \text{with} \quad \sum c_i = 0.$$

For example, $\tau_u - \tau_s$ is the contrast that compares the effects (as measured by the response variable) of treatments u and s. If $\tau_u - \tau_s = 0$, then treatments u and s affect the response in exactly the same way, and we say that these treatments do not differ. Otherwise, the treatments do differ in the way they affect the response. We showed in Sect. 3.4 that for a completely randomized design and the one-way analysis of variance model (3.3.1), every contrast $\sum c_i \tau_i$ is estimable with least squares estimate

$$\sum c_i \hat{\tau}_i = \sum c_i (\hat{\mu} + \hat{\tau}_i) = \sum c_i \bar{y}_{i.} \tag{4.2.1}$$

and corresponding least squares estimator $\sum c_i \overline{Y}_{i.}$. The variance of the least squares estimator is

$$\text{Var}\left(\sum c_i \overline{Y}_{i.}\right) = \sum c_i^2 \text{Var}(\overline{Y}_{i.}) = \sum c_i^2 (\sigma^2 / r_i) = \sigma^2 \sum (c_i^2 / r_i). \qquad (4.2.2)$$

The first equality uses the fact that the treatment sample means $\overline{Y}_{i.}$ involve different response variables, which in model (3.3.1) are independent. The error variance σ^2 is generally unknown and is estimated by the unbiased estimate msE, giving the estimated variance of the contrast estimator as

$$\widehat{\text{Var}}\left(\sum c_i \overline{Y}_{i.}\right) = msE \sum (c_i^2 / r_i).$$

The *estimated standard error* of the estimator is the square root of this quantity, namely,

$$\sqrt{\widehat{\text{Var}}\left(\sum c_i \overline{Y}_{i.}\right)} = \sqrt{msE \sum (c_i^2 / r_i)}. \qquad (4.2.3)$$

Normalized Contrasts

When several contrasts are to be compared, it is sometimes helpful to be able to measure them all on the same scale. A contrast is said to be *normalized* if it is scaled so that its least squares estimator has variance σ^2. From (4.2.2), it can be seen that a contrast $\Sigma c_i \tau_i$ is normalized by dividing it by $\sqrt{\Sigma c_i^2 / r_i}$. If we write $h_i = c_i / \sqrt{\Sigma c_i^2 / r_i}$, then the least squares estimator $\Sigma h_i \overline{Y}_{i.}$ of the normalized contrast $\Sigma h_i \tau_i$ has the following distribution:

$$\sum h_i \overline{Y}_{i.} \sim N\left(\sum h_i \tau_i, \sigma^2\right), \quad \text{where } h_i = \frac{c_i}{\sqrt{\sum c_i^2 / r_i}}.$$

Normalized contrasts will be used for hypothesis testing (Sect. 4.3.3).

Contrast Coefficients

It is convenient to represent a contrast by listing only the coefficients of the parameters $\tau_1, \tau_2, \ldots, \tau_v$. Thus, $\sum c_i \tau_i = c_1 \tau_1 + c_2 \tau_2 + \cdots + c_v \tau_v$ would be represented by the list of *contrast coefficients*

$$[c_1, \ c_2, \ \ldots, \ c_v].$$

Some types of contrasts are used frequently in practice, and these are identified in Sects. 4.2.1–4.2.4.

4.2.1   Pairwise Comparisons

As the name suggests, *pairwise comparisons* are simple differences $\tau_u - \tau_s$ of pairs of parameters τ_u and τ_s ($u \neq s$). These are of interest when the experimenter wishes to compare each treatment with every other treatment. The list of contrast coefficients for the pairwise difference $\tau_u - \tau_s$ is

$$[0, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0],$$

where the 1 and -1 are in positions u and s, respectively. The least squares estimate of $\tau_u - \tau_s$ is obtained from (4.2.1) by setting $c_u = 1$, $c_s = -1$, and all other c_i equal to zero, giving

$$\hat{\tau}_u - \hat{\tau}_s = \overline{y}_{u.} - \overline{y}_{s.}\,,$$

and the corresponding least squares estimator is $\overline{Y}_{u.} - \overline{Y}_{s.}$. Its estimated standard error is obtained from (4.2.3) and is equal to

$$\sqrt{\widehat{\text{Var}}(\overline{Y}_{u.} - \overline{Y}_{s.})} = \sqrt{msE\left((1/r_u) + (1/r_s)\right)}\,.$$

Example 4.2.1 Battery experiment, continued

Details for the battery experiment were given in Sect. 2.5.2 (p. 24). The experimenter was interested in comparing the life per unit cost of each battery type with that of each of the other battery types. The average lives per unit cost (in minutes/dollar) for the four batteries, calculated from the data in Table 2.8, p. 27, are

$$\overline{y}_{1.} = 570.75\,, \quad \overline{y}_{2.} = 860.50\,, \quad \overline{y}_{3.} = 433.00\,, \quad \overline{y}_{4.} = 496.25\,.$$

The least squares estimates of the pairwise differences are, therefore,

$$\hat{\tau}_1 - \hat{\tau}_2 = -289.75\,, \quad \hat{\tau}_1 - \hat{\tau}_3 = 137.75\,, \quad \hat{\tau}_1 - \hat{\tau}_4 = 74.50\,,$$
$$\hat{\tau}_2 - \hat{\tau}_3 = 427.50\,, \quad \hat{\tau}_2 - \hat{\tau}_4 = 364.25\,, \quad \hat{\tau}_3 - \hat{\tau}_4 = -63.25\,.$$

The estimated pairwise differences suggest that battery type 2 (alkaline, store brand) is vastly superior to the other three battery types in terms of the mean life per unit cost. Battery type 1 (alkaline, name brand) appears better than types 3 and 4, and battery type 4 (heavy duty, store brand) better than type 3 (heavy duty, name brand). We do, however, need to investigate whether or not these perceived differences might be due only to random fluctuations in the data.

In Example 3.4.2 (p. 40), the error variance was estimated to be $msE = 2367.71$. The sample sizes were $r_1 = r_2 = r_3 = r_4 = 4$, and consequently, the estimated standard error for each pairwise comparison is equal to

$$\sqrt{2367.71\left(\frac{1}{4} + \frac{1}{4}\right)} = 34.41 \text{ min/\$}\,.$$

It can be seen that all of the estimated pairwise differences involving battery type 2 are bigger than four times their estimated standard errors. This suggests that the perceived differences in battery type 2 and the other batteries are of sizable magnitudes and are unlikely to be due to random error. We shall formalize these comparisons in terms of confidence intervals in Example 4.4.3 later in this chapter. □

4.2.2 Treatment Versus Control

If the experimenter is interested in comparing the effects of one special treatment with the effects of each of the other treatments, then the special treatment is called the *control*. For example, a pharmaceutical experiment might involve one or more experimental drugs together with a standard drug that has been on the market for some years. Frequently, the objective of such an experiment is to compare the effect of each experimental drug with that of the standard drug but not necessarily with the effects of any of the other experimental drugs. The standard drug is then the control. If we code the control as level 1, and the experimental drugs as levels 2, 3, . . . , v, respectively, then the contrasts of interest are $\tau_2 - \tau_1$, $\tau_3 - \tau_1$, . . . , $\tau_v - \tau_1$. These contrasts are known as *treatment versus control* contrasts. They form a subset of the pairwise differences, so we can use the same formulae for the least squares estimate and the estimated

standard error. The contrast coefficients for the contrast $\tau_i - \tau_1$ are $[-1, 0, \ldots, 0, 1, 0, \ldots, 0]$, where the 1 is in position i.

4.2.3 Difference of Averages

Sometimes the levels of the treatment factors divide naturally into two or more groups, and the experimenter is interested in the *difference of averages* contrast that compares the average effect of one group with the average effect of the other group(s). For example, consider an experiment that is concerned with the effect of different colors of exam paper (the treatments) on students' exam performance (the response). Suppose that treatments 1 and 2 represent the pale colors, white and yellow, whereas treatments 3, 4, and 5 represent the darker colors, blue, green and pink. The experimenter may wish to compare the effects of light and dark colors on exam performance. One way of measuring this is to estimate the contrast $\frac{1}{2}(\tau_1 + \tau_2) - \frac{1}{3}(\tau_3 + \tau_4 + \tau_5)$, which is the difference of the average effects of the light and dark colors. The corresponding contrast coefficients are

$$\left[\frac{1}{2}, \frac{1}{2}, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3} \right].$$

From (4.2.1) and (4.2.3), the least squares estimate would be

$$\frac{1}{2}\bar{y}_{1.} + \frac{1}{2}\bar{y}_{2.} - \frac{1}{3}\bar{y}_{3.} - \frac{1}{3}\bar{y}_{4.} - \frac{1}{3}\bar{y}_{5.}$$

with estimated standard error

$$\sqrt{msE \left(\frac{1}{4r_1} + \frac{1}{4r_2} + \frac{1}{9r_3} + \frac{1}{9r_4} + \frac{1}{9r_5} \right)}.$$

Example 4.2.2 `Battery experiment, continued`

In the battery experiment of Sect. 2.5.2, p. 24, battery types 1 and 2 were alkaline batteries, while types 3 and 4 were heavy duty. In order to compare the running time per unit cost of these two types of batteries, we examine the contrast $\frac{1}{2}(\tau_1 + \tau_2) - \frac{1}{2}(\tau_3 + \tau_4)$. The least squares estimate is

$$\frac{1}{2}(570.75 + 860.50) - \frac{1}{2}(433.00 + 496.25) = 251.00 \text{ min/\$},$$

suggesting that the alkaline batteries are more economical (on average by over four hours per dollar spent). The associated standard error is $\sqrt{msE(4/16)} = 24.32$ min/\$, so the estimated difference in running time per unit cost is over ten times larger than the standard error, suggesting that the observed difference is not just due to random fluctuations in the data.   □

4.2.4 Trends

Trend contrasts may be of interest when the levels of the treatment factor are quantitative and have a natural ordering. For example, suppose that the treatment factor is temperature and its selected levels are 50 °C, 75 °C, 100 °C, coded as 1, 2, 3, respectively. The experimenter may wish to know whether

the value of the response variable increases or decreases as the temperature increases and, if so, whether the rate of change remains constant. These questions can be answered by estimating linear and quadratic trends in the response.

The trend contrast coefficients for v *equally spaced* levels of a treatment factor and *equal sample sizes* are listed in Table A.2 for values of v between 3 and 7. For v treatments, trends up to $(v-1)$th order can be measured. Experimenters rarely use more than four levels for a quantitative treatment factor, since it is unusual for strong quartic and higher-order trends to occur in practice, especially within the narrow range of levels considered in a typical experiment.

Table A.2 does not tabulate contrast coefficients for unequally spaced levels or for unequal sample sizes. The general method of obtaining the coefficients of the trend contrasts involves fitting a regression model to the noncoded levels of the treatment factor. It can be shown that the linear trend contrast coefficients can easily be calculated as

$$c_i = r_i(x_i - \overline{x}_.), \text{ where } \overline{x}_. = (\Sigma r_i x_i)/n, \tag{4.2.4}$$

where r_i is the number of observations taken on the ith uncoded level x_i of the treatment factor, and $n = \Sigma r_i$ is the total number of observations. We are usually interested only in whether or not the linear trend is likely to be negligible, and to make this assessment, the contrast estimate is compared with its standard error. Consequently, we may multiply or divide the calculated coefficients by any integer without losing any information. When the r_i are all equal, the coefficients listed in Appendix A.2 are obtained, possibly multiplied or divided by an integer. Expressions for quadratic and higher-order trend coefficients are more complicated (see Draper and Smith 1998, Chap. 22).

Example 4.2.3 `Heart-lung pump experiment, continued`

The experimenter who ran the heart–lung pump experiment of Example 3.4.1, p. 37, expected to see a linear trend in the data, since he expected the flow rate to increase as the number of revolutions per minute (rpm) of the pump head was increased. The plot of the data in Fig. 3.1 (p. 38) shows the observed flow rates at the five different levels of rpm. From the figure, it might be anticipated that the linear trend is large but higher-order trends are very small.

The five levels of rpm observed were 50, 75, 100, 125, 150, which are equally spaced. Had there been equal numbers of observations at each level, then we could have used the contrast coefficients $[-2, -1, 0, 1, 2]$ for the linear trend contrast and $[2, -1, -2, -1, 2]$ for the quadratic trend contrast as listed in Table A.2 for $v = 5$ levels of the treatment factor. However, here the sample sizes were $r_1 = r_3 = r_5 = 5$, $r_2 = 3$ and $r_4 = 2$. The coefficients for the linear trend are calculated via (4.2.4). Now $n = \Sigma r_i = 20$, and

$$(\Sigma r_i x_i)/n = 20^{-1} \times (5(50) + 3(75) + 5(100) + 2(125) + 5(150)) = 98.75.$$

So, we have

x_i	$r_i(x_i - \overline{x}_.)$
50	$5 \times (50 - 98.75) = -243.75$
75	$3 \times (75 - 98.75) = -71.25$
100	$5 \times (100 - 98.75) = 6.25$
125	$2 \times (125 - 98.75) = 52.50$
150	$5 \times (150 - 98.75) = 256.25$

If the coefficients are multiplied by 4, they are then integers each divisible by 5 so rather than using the calculated coefficients $[-243.75, -71.25, 6.25, 52.50, 256.25]$, we can multiply them by 4/5 and use the linear trend coefficients $[-195, -57, 5, 42, 205]$. The average flow rates (l/min) were calculated as

$$\bar{y}_{1.} = 1.1352, \, \bar{y}_{2.} = 1.7220, \, \bar{y}_{3.} = 2.3268, \, \bar{y}_{4.} = 2.9250, \, \bar{y}_{5.} = 3.5292.$$

The least squares estimate $\Sigma c_i \bar{y}_{i.}$ of the linear contrast is then

$$-195\bar{y}_{1.} - 57\bar{y}_{2.} + 5\bar{y}_{3.} + 42\bar{y}_{4.} + 205\bar{y}_{5.} = 538.45$$

l/min. The linear trend certainly appears to be large. However, before drawing conclusions, we need to compare this trend estimate with its corresponding estimated standard error. The data give $\sum \sum y_{it}^2 = 121.8176$, and we calculate the error sum of squares (3.4.5), p. 39, as $ssE = 0.0208$, giving an unbiased estimate of σ^2 as

$$msE = ssE/(n - v) = 0.0208/(20 - 5) = 0.001387.$$

The estimated standard error of the linear trend estimator is then

$$\sqrt{msE \left(\frac{(-195)^2}{5} + \frac{(-57)^2}{3} + \frac{(5)^2}{5} + \frac{(42)^2}{2} + \frac{(205)^2}{5} \right)} = 4.988.$$

Clearly, the estimate of the linear trend is extremely large compared with its standard error.

Had we normalized the contrast, the linear contrast coefficients would each have been divided by

$$\sqrt{\sum c_i^2/r_i} = \sqrt{\frac{(-195)^2}{5} + \frac{(-57)^2}{3} + \frac{(5)^2}{5} + \frac{(42)^2}{2} + \frac{(205)^2}{5}} = 134.09,$$

and the normalized linear contrast estimate would have been 4.0156. The estimated standard error of all normalized contrasts is $\sqrt{msE} = 0.03724$ for this experiment, so the normalized linear contrast estimate remains large compared with the standard error.   □

4.3   Individual Contrasts and Treatment Means

4.3.1   Confidence Interval for a Single Contrast

In this section, we obtain a formula for a confidence interval for an individual contrast. If confidence intervals for more than one contrast are required, then the multiple comparison methods of Sect. 4.4 should be used instead. We give the formula first, and the derivation afterwards. A $100(1 - \alpha)\%$ confidence interval for the contrast $\Sigma c_i \tau_i$ is

$$\sum c_i \bar{y}_{i.} - t_{n-v,\alpha/2} \sqrt{msE \sum c_i^2/r_i} \, \leq \, \sum c_i \tau_i \tag{4.3.5}$$

$$\leq \sum c_i \bar{y}_{i.} + t_{n-v,\alpha/2} \sqrt{msE \sum c_i^2/r_i}.$$

We can write this more succinctly as

$$\sum c_i \tau_i \ \in \ \left(\sum c_i \bar{y}_{i.} \pm t_{n-v,\alpha/2} \sqrt{msE \sum c_i^2/r_i} \right), \tag{4.3.6}$$

where the symbol $\pm$, which is read as "plus or minus," denotes that the upper limit of the interval is calculated using $+$ and the lower limit using $-$. The symbols "$\sum c_i \tau_i \ \in$" mean that the interval includes the true value of the contrast $\sum c_i \tau_i$ with $100(1-\alpha)\%$ confidence. For future reference, we note that the general form of the above confidence interval is

$$\sum c_i \tau_i \ \in \ \left(\sum c_i \hat{\tau}_i \pm t_{df,\alpha/2} \sqrt{\widehat{\mathrm{Var}}(\Sigma c_i \hat{\tau}_i)} \right), \tag{4.3.7}$$

where df is the number of degrees of freedom for error.

To derive the confidence interval (4.3.5), we will need to use some results about normally distributed random variables. As we saw in the previous section, for the completely randomized design and one-way analysis of variance model (3.3.1), the least squares estimator of the contrast $\sum c_i \tau_i$ is $\sum c_i \bar{Y}_{i.}$, which has variance $\mathrm{Var}(\Sigma c_i \bar{Y}_{i.}) = \sigma^2 \sum c_i^2/r_i$. This estimator is a linear combination of normally distributed random variables and therefore also has a normal distribution. Subtracting the mean and dividing by the standard deviation gives us a random variable

$$\frac{D}{\sigma} = \frac{\sum c_i \bar{Y}_{i.} - \sum c_i \tau_i}{\sigma \sqrt{\sum c_i^2/r_i}}, \tag{4.3.8}$$

which has a $N(0, 1)$ distribution. We estimate the error variance, σ^2, by msE, and from Sect. 3.4.6, p. 39, we know that

$$MSE/\sigma^2 = SSE/(n-v)\sigma^2 \sim \chi^2_{n-v}/(n-v).$$

It can be shown that the random variables D and MSE are independent (see Graybill, 1976), and the ratio of a normally distributed random variable and a chi-squared random variable that are independent has a t-distribution with the same number of degrees of freedom as the chi-squared distribution. Hence, the ratio $D/\sqrt{MSE}$ has a t distribution with $n-v$ degrees of freedom. Using the expression (4.3.8), we can now write down the following probability statement about $D/\sqrt{MSE}$:

$$P\left(-t_{n-v,\alpha/2} \ \le \ \frac{\sum c_i \bar{Y}_{i.} - \sum c_i \tau_i}{\sqrt{MSE \sum c_i^2/r_i}} \ \le \ t_{n-v,\alpha/2} \right) = 1 - \alpha,$$

where $t_{n-v,\alpha/2}$ is the percentile of the t_{n-v} distribution corresponding to a probability of $\alpha/2$ in the right-hand-tail, the value of which can be obtained from Table A.4. Manipulating the two inequalities, the probability statement becomes

$$P\left(\sum c_i \bar{Y}_{i.} - t_{n-v,\alpha/2} \sqrt{MSE \sum c_i^2/r_i} \ \le \ \sum c_i \tau_i \right. \tag{4.3.9}$$

$$\left. \le \ \sum c_i \bar{Y}_{i.} + t_{n-v,\alpha/2} \sqrt{MSE \sum c_i^2/r_i} \right) = 1 - \alpha.$$

Then replacing the estimators by their observed values in this expression gives a $100(1-\alpha)\%$ confidence interval for $\sum c_i \tau_i$ as in (4.3.5).

Example 4.3.1 `Heart-lung pump experiment, continued`

Consider the heart–lung pump experiment of Examples 3.4.1 and 4.2.3, p. 37 and 73. The least squares estimate of the difference in fluid flow at 75 rpm and 50 rpm (levels 2 and 1 of the treatment factor, respectively) is

$$\Sigma c_i \overline{y}_{i.} = \overline{y}_{2.} - \overline{y}_{1.} = 0.5868$$

l/min. Since there were $r_2 = 5$ observations at 75 rpm and $r_1 = 3$ observations at 50 rpm, and $msE = 0.001387$, the estimated standard error of this contrast is

$$\sqrt{msE\,\Sigma c_i^2/r_i} = \sqrt{0.001387\left(\frac{1}{3} + \frac{1}{5}\right)} = 0.0272\ \text{l/min}\,.$$

Using this information, together with $t_{15,0.025} = 2.131$, we obtain from (4.3.6) a 95% confidence interval (in units of l/min) for $\tau_2 - \tau_1$ as

$$(0.5868 \pm (2.131)(0.0272)) \quad = \quad (0.5288, 0.6448)\,.$$

This tells us that with 95% confidence, the fluid flow at 75 rpm of the pump is between 0.53 and 0.64 liters per minute greater than at 50 rpm.   □

  Confidence bounds, or one-sided confidence intervals, can be derived in the same manner as two-sided confidence intervals. For the completely randomized design and one-way analysis of variance model (3.3.1), a $100(1 - \alpha)\%$ *upper confidence bound* for $\sum c_i \tau_i$ is

$$\sum c_i \tau_i < \sum c_i \overline{y}_{i.} + t_{df,\alpha}\sqrt{msE\sum c_i^2/r_i}\,, \tag{4.3.10}$$

and a $100(1 - \alpha)\%$ *lower confidence bound* for $\sum c_i \tau_i$ is

$$\sum c_i \tau_i > \sum c_i \overline{y}_{i.} - t_{df,\alpha}\sqrt{msE\sum c_i^2/r_i}\,, \tag{4.3.11}$$

where $t_{df,\alpha}$ is the percentile of the t distribution with df degrees of freedom and probability α in the right-hand tail.

4.3.2 Confidence Interval for a Single Treatment Mean

For the one-way analysis of variance model (3.3.1), the true mean response $\mu + \tau_s$ of the sth level of a treatment factor was shown in Sect. 3.4 to be estimable with least squares estimator $\overline{Y}_{s.}$. Although one is unlikely to be interested in only one of the treatment means, we can obtain a confidence interval as follows.

  Since $\overline{Y}_{s.} \sim N(\mu + \tau_s, \sigma^2/r_s)$ for model (3.3.1), we can follow the same steps as those leading to (4.3.6) and obtain a $100(1 - \alpha)\%$ confidence interval for $\mu + \tau_s$ as

$$\mu + \tau_s \in (\overline{y}_{s.} \pm t_{df,\alpha/2}\sqrt{msE/r_s})\,. \tag{4.3.12}$$

Example 4.3.2 `Heart-lung pump experiment, continued`

Suppose that the experimenter had required a 99% confidence interval for the true average fluid flow ($\mu + \tau_3$) for the heart–lung pump experiment of Example 3.4.1, p. 37, when the revolutions per minute of the pump are set to 100 rpm. Using (4.3.12) and $r_3 = 5$, $\overline{y}_{3.} = 2.3268$, $msE = 0.001387$, $n - v = 20 - 5$, and $t_{15, 0.005} = 2.947$, the 99% confidence interval for $\mu + \tau_3$ is

$$\mu + \tau_3 \ \in \ (2.3268 \pm (2.947)(0.01666)) \ = \ (2.2777, 2.3759) \, .$$

So, with 99% confidence, the true average flow rate at 100 rpm of the pump is believed to be between 2.28 and 2.38 l/min. □

4.3.3 Hypothesis Test for a Single Contrast or Treatment Mean

The outcome of a hypothesis test can be deduced from the corresponding confidence interval in the following way. The null hypothesis $H_0 : \Sigma c_i \tau_i = h$ will be rejected at significance level α in favor of the two-sided alternative hypothesis $H_A : \Sigma c_i \tau_i \neq h$ if the corresponding confidence interval for $\Sigma c_i \tau_i$ fails to contain h. For example, the 95% confidence interval for $\tau_2 - \tau_1$ in Example 4.3.1 does not contain zero, so the hypothesis $H_0 : \tau_2 - \tau_1 = 0$ (that the flow rates are the same at 50 and 75 rpm) would be rejected at significance level $\alpha = 0.05$ in favor of the alternative hypothesis (that the flow rates are not equal).

We can make this more explicit, as follows. Suppose we wish to test the hypothesis $H_0 : \Sigma c_i \tau_i = 0$ against the alternative hypothesis $H_A : \Sigma c_i \tau_i \neq 0$. The interval (4.3.6) fails to contain 0 if the absolute value of $\Sigma c_i \overline{y}_{i.}$ is bigger than $t_{n-v, \alpha/2} \sqrt{msE \, \Sigma c_i^2 / r_i}$. Therefore, the rule for testing the null hypothesis against the alternative hypothesis is

$$\text{reject } H_0 \text{ if } \left| \frac{\sum c_i \overline{y}_{i.}}{\sqrt{msE \sum c_i^2 / r_i}} \right| > t_{n-v, \alpha/2} \, , \tag{4.3.13}$$

where $|\ |$ denotes absolute value. We call such rules *decision rules*. If H_0 is rejected, then H_A is automatically accepted. The test statistic can be squared, so that the decision rule becomes

$$\text{reject } H_0 \text{ if } \frac{\left(\sum c_i \overline{y}_{i.} \right)^2}{msE \sum c_i^2 / r_i} > t_{n-v, \alpha/2}^2 = F_{1, n-v, \alpha} \, ,$$

and the F distribution can be used instead of the t distribution. Notice that the test statistic is the square of the normalized contrast estimate divided by *msE*. We call the quantity

$$ssc = \frac{\left(\sum c_i \overline{y}_{i.} \right)^2}{\sum c_i^2 / r_i} \tag{4.3.14}$$

the *sum of squares for the contrast*, or *contrast sum of squares* (even though it is the "sum" of only one squared term). The decision rule can be more simply expressed as

$$\text{reject } H_0 \text{ if } \frac{ssc}{msE} > F_{1, n-v, \alpha} \, . \tag{4.3.15}$$

For future reference, we can see that the general form of *ssc/msE* is

$$\frac{ssc}{msE} = \frac{\left(\sum c_i \hat{\tau}_i\right)^2}{\widehat{\text{Var}}\left(\sum c_i \hat{\tau}_i\right)}. \tag{4.3.16}$$

The above test is a two-tailed test, since the null hypothesis will be rejected for both large and small values of the contrast. One-tailed tests can be derived also, as follows.

The decision rule for the test of $H_0 : \Sigma c_i \tau_i = 0$ against the one-sided alternative hypothesis $H_A : \sum c_i \tau_i > 0$ is

$$\text{reject } H_0 \text{ if } \frac{\sum c_i \bar{y}_{i.}}{\sqrt{msE \sum c_i^2 / r_i}} > t_{n-v,\alpha}. \tag{4.3.17}$$

The outcome of this test can be deduced from the appropriate one-sided confidence bound. In particular, the null hypothesis will be rejected at significance level α if the corresponding $100(1-\alpha)\%$ lower confidence bound for $\sum c_i \tau_i$ in Eq. (4.3.11) is above zero so excludes zero.

Similarly, for the one-sided alternative hypothesis $H_A : \sum c_i \tau_i < 0$, the decision rule is

$$\text{reject } H_0 \text{ if } \frac{\sum c_i \bar{y}_{i.}}{\sqrt{msE \sum c_i^2 / r_i}} < -t_{n-v,\alpha}. \tag{4.3.18}$$

Here the null hypothesis will be rejected at significance level α if the corresponding $100(1-\alpha)\%$ upper confidence bound for $\sum c_i \tau_i$ in Eq. (4.3.10) is below zero so excludes zero.

If the hypothesis test concerns a single treatment mean, for example, $H_0 : \mu + \tau_s = 0$, then the decision rules (4.3.13)–(4.3.18) are modified by setting c_s equal to one and all the other c_i equal to zero.

Example 4.3.3 `Filter experiment`

Lorenz et al. (1982) describe an experiment that was carried out to determine the relative performance of seven membrane filters in supporting the growth of bacterial colonies. The seven filter types are regarded as the seven levels of the treatment factor and are coded 1, 2, ..., 7. Filter types 1, 4, and 7 were received presterilized. Several different types of data were collected, but the only data considered here are the colony counts of fecal coliforms from a sample of Olentangy River water (August 1980) that grew on each filter. Three filters of each type were observed and the average colony counts[1] were

$$\bar{y}_{1.} = 36.0, \ \bar{y}_{2.} = 18.0, \ \bar{y}_{3.} = 27.7, \ \bar{y}_{4.} = 28.0, \ \bar{y}_{5.} = 28.3, \ \bar{y}_{6.} = 37.7, \ \bar{y}_{7.} = 30.3.$$

The mean squared error was $msE = 21.6$. Suppose we wish to test the hypothesis that the presterilized filters do not differ from the nonpresterilized filters in terms of the average colony counts, against a two-sided alternative hypothesis that they do differ. The hypothesis of interest involves a difference of averages contrast, that is,

$$H_0 : \frac{1}{3}(\tau_1 + \tau_4 + \tau_7) - \frac{1}{4}(\tau_2 + \tau_3 + \tau_5 + \tau_6) = 0.$$

From (4.3.15), the decision rule is to reject H_0 if

$$\frac{ssc}{msE} = \frac{\left[\frac{1}{3}(\overline{y}_{1.} + \overline{y}_{4.} + \overline{y}_{7.}) - \frac{1}{4}(\overline{y}_{2.} + \overline{y}_{3.} + \overline{y}_{5.} + \overline{y}_{6.})\right]^2}{msE\left[\frac{(\frac{1}{3})^2}{3} + \frac{(\frac{1}{3})^2}{3} + \frac{(\frac{1}{3})^2}{3} + \frac{(-\frac{1}{4})^2}{3} + \frac{(-\frac{1}{4})^2}{3} + \frac{(-\frac{1}{4})^2}{3} + \frac{(-\frac{1}{4})^2}{3}\right]} > F_{1,14,\alpha}\,.$$

Selecting a probability of a Type I error equal to $\alpha = 0.05$, this becomes

$$\text{reject } H_0 \text{ if } \frac{(3.508)^2}{(21.6)(0.1944)} = 2.931 \; > \; F_{1,14,0.05}\,.$$

Since $F_{1,14,0.05} = 4.6$, there is not sufficient evidence to reject the null hypothesis, and we conclude that the presterilized filters do not differ significantly from the nonpresterilized filters when α is set at 0.05.

   Notice that the null hypothesis would be rejected if the probability of a Type I error is set a little higher than $\alpha = 0.10$, since $F_{1,14,0.10} = 3.10$. Thus, if these experimenters are willing to accept a high risk of incorrectly rejecting the null hypothesis, they would be able to conclude that there is a difference between the presterilized and the nonpresterilized filters.

   A 95% confidence interval for this difference can be obtained from (4.3.6) as follows:

$$\frac{1}{3}(\tau_1 + \tau_4 + \tau_7) - \frac{1}{4}(\tau_2 + \tau_3 + \tau_5 + \tau_6) \in \left(3.508 \pm t_{14,0.025}\sqrt{(21.6)(0.1944)}\right),$$

and since $t_{14,0.025} = 2.145$, the interval becomes

$$(3.508 \pm (2.145)(2.0492)) = (-0.888, 7.904)\,,$$

where the measurements are average colony counts. The interval contains zero, which agrees with the hypothesis test at $\alpha = 0.05$.   □

4.3.4   Equivalence of Tests and Confidence Intervals (Optional)

There is a stronger relationship between hypothesis tests and confidence intervals (including both 1- and 2-sided confidence intervals) than was described in Sect. 4.3.3. As already discussed, the outcome of a hypothesis test at significance level α can be deduced from the corresponding $100(1 - \alpha)\%$ confidence interval. Correspondingly, though less well known, one can conclude from a hypothesis test that the true value of the parameter is in the corresponding confidence interval, by virtue of rejecting all values outside the interval, providing more specific test conclusions than simply whether or not one rejects the null hypothesis and so believes the alternative.

   To illustrate this, consider a two-tailed level-α test of the null hypothesis $H_0 : \sum c_i \tau_i = 0$ against the alternative hypothesis $H_A : \sum c_i \tau_i \neq 0$. Under standard practice, only the null hypothesis H_0 is tested at significance level α. If H_0 is rejected in favor of H_A, one simply eliminates zero as a possible value of the treatment contrast, and the hypothesis testing procedure guarantees that the probability of making a mistake by rejecting $H_0 : \sum c_i \tau_i = 0$ when it is true is at most α.

   Expanding upon standard practice, suppose one not only tests H_0; rather, suppose one conducts a standard two-tailed level-α test of the null hypothesis $H_{0b} : \sum c_i \tau_i = b$ against the alternative

hypothesis $H_{Ab} : \sum c_i \tau_i \neq b$ for *each* real number b. Then the probably of rejecting $H_0 : \sum c_i \tau_i = 0$ if it is true is still controlled to be α. Moreover, even though this expanded testing procedure involves conducting an infinite number of tests rather than only one, the probability of making *any* false rejections—namely, of falsely rejecting any true null hypothesis H_{0b}—is still at most α. This follows from the *partitioning principle*, (see Finner and Strassburger 2002, and references therein). In particular, because the sets $\{b\}$ partition the set of real numbers, H_{0b} is only true for exactly one value of b, b^* say. So, one can only make a mistake by rejecting the only true null hypothesis H_{0b^*}, and the probability of rejecting H_{0b^*} is α. Thus, in terms of error rates, there is no additional cost in testing infinitely many hypothesis H_{0b} instead of just one.

Furthermore, as we know, the null hypothesis H_{0b} will be rejected at level-α precisely for those values b outside the $100(1-\alpha)\%$ confidence interval for $\sum c_i \tau_i$. In other words, all values of $\sum c_i \tau_i$ outside of the $100(1-\alpha)\%$ confidence interval are rejected at simultaneous significance level α. Hence, one can conclude from this extended test that the true value of $\sum c_i \tau_i$ is in the corresponding $100(1-\alpha)\%$ confidence interval for $\sum c_i \tau_i$. This is true whether or not one rejects H_0, providing a more specific conclusion than simply rejecting the null hypothesis or not.

For example, if one does reject $H_0 : \sum c_i \tau_i = 0$ at significance level α, then one can conclude with Type I error probability, α not only that $\sum c_i \tau_i \neq 0$ but also more specifically that the true value of $\sum c_i \tau_i$ is in the corresponding $100(1-\alpha)\%$ confidence interval for $\sum c_i \tau_i$, where this confidence interval will consist only of positive values if H_0 is rejected and $\sum c_i \hat{\tau}_i > 0$, or only of negative values if H_0 is rejected and $\sum c_i \hat{\tau}_i < 0$. On the other hand, if one fails to reject H_0, one can still conclude that the true value of $\sum c_i \tau_i$ is in the corresponding $100(1-\alpha)\%$ confidence interval for $\sum c_i \tau_i$, but this confidence interval will include zero as a possible value of the treatment contrast.

The analogous equivalence exists between one-sided tests and corresponding confidence bounds. Consider for example the standard level-α test of $H_0 : \Sigma c_i \tau_i = 0$ against the one-sided alternative hypothesis $H_A : \sum c_i \tau_i > 0$. More broadly, one can conduct a standard one-tailed α-level test of the null hypothesis $H_{0b} : \sum c_i \tau_i = b$ against the alternative hypothesis $H_{Ab} : \sum c_i \tau_i > b$ for each real number b. In so doing, $H_{0b} : \sum c_i \tau_i = b$ will be rejected for exactly those values of b that are below the $100(1-\alpha)\%$ lower confidence bound for $\sum c_i \tau_i$ given in Eq. (4.3.11). In other words, the values of $\sum c_i \tau_i$ rejected at level α are exactly the values outside of the $100(1-\alpha)\%$ (one-sided) confidence interval. Consequently, whether or not H_0 is rejected, one can conclude that the true value of $\sum c_i \tau_i$ is above the $100(1-\alpha)\%$ lower confidence bound for $\sum c_i \tau_i$, and one will reject H_0 and conclude $\sum c_i \tau_i > 0$ exactly when the lower confidence bound is positive.

Similarly, for testing $H_0 : \Sigma c_i \tau_i = 0$ against $H_A : \sum c_i \tau_i < 0$ at level α, one can expand this by conducting a standard one-tailed level-α test of $H_{0b} : \sum c_i \tau_i = b$ against $H_{Ab} : \sum c_i \tau_i < b$ for each real number b. Then the values of $\sum c_i \tau_i$ rejected at level α are exactly the values above the $100(1-\alpha)\%$ upper confidence bound given in Eq. (4.3.10). Consequently, whether or not H_0 is rejected, one can conclude that the true value of $\sum c_i \tau_i$ is below its $100(1-\alpha)\%$ upper confidence bound. Also, one will reject H_0 and conclude $\sum c_i \tau_i < 0$ exactly when the upper confidence bound is negative.

The equivalence between testing and confidence intervals illustrated above applies quite broadly, including for example to one-step multiple comparison procedures such as those considered in the next section (as discussed by Voss 2008, 2010). The partitioning principle also facilitates the construction of more complicated confidence sets corresponding to stepwise multiple tests (see Stefansson et al. 1988).

4.4   Methods of Multiple Comparisons

4.4.1   Multiple Confidence Intervals

Often, the most useful analysis of experimental data involves the calculation of a number of different confidence intervals, one for each of several contrasts or treatment means. The confidence level for a single confidence interval is based on the probability, like (4.3.9), that the random interval will be "correct" (meaning that the random interval will contain the true value of the contrast or function).

It is shown below that when several confidence intervals are calculated, the probability that they are all simultaneously correct can be alarmingly small. Similarly, when several hypotheses are to be tested, the probability that at least one hypothesis is incorrectly rejected can be uncomfortably high. Much research has been done over the years to find ways around these problems. The resulting techniques are known as *methods of multiple comparison*, the intervals are called *simultaneous confidence intervals*, and the tests are called *simultaneous hypothesis tests*.

Suppose an experimenter wishes to calculate m confidence intervals, each having a $100(1 - \alpha^*)\%$ confidence level. Then each interval will be individually correct with probability $1 - \alpha^*$. Let S_j be the event that the jth confidence interval will be correct and $\overline{S}_j$ the event that it will be incorrect ($j = 1, \ldots, m$). Then, using the standard rules for probabilities of unions and intersections of events, it follows that

$$P(S_1 \cap S_2 \cap \cdots \cap S_m) = 1 - P(\overline{S}_1 \cup \overline{S}_2 \cup \cdots \cup \overline{S}_m) \,.$$

This says that the probability that all of the intervals will be correct is equal to one minus the probability that at least one will be incorrect. If $m = 2$,

$$
\begin{aligned}
P(\overline{S}_1 \cup \overline{S}_2) &= P(\overline{S}_1) + P(\overline{S}_2) - P(\overline{S}_1 \cap \overline{S}_2) \\
&\leq P(\overline{S}_1) + P(\overline{S}_2) \,.
\end{aligned}
$$

A similar result, which can be proved by mathematical induction, holds for any number m of events, that is,

$$P(\overline{S}_1 \cup \overline{S}_2 \cup \cdots \cup \overline{S}_m) \leq \sum_j P(\overline{S}_j) \,,$$

with equality if the events $\overline{S}_1, \overline{S}_2, \ldots, \overline{S}_m$ are mutually exclusive. Consequently,

$$P(S_1 \cap S_2 \cap \cdots \cap S_m) \geq 1 - \sum_j P(\overline{S}_j) = 1 - m\alpha^* \,; \qquad (4.4.19)$$

that is, the probability that the m intervals will simultaneously be correct is at least $1 - m\alpha^*$. The probability $m\alpha^*$ is called the *overall significance level* or *experimentwise error rate*. A typical value for α^* for a single confidence interval is 0.05, so the probability that six confidence intervals each calculated at a 95% individual confidence level will simultaneously be correct is at least 0.7. Although "at least" means "bigger than or equal to," it is not known in practice how much bigger than 0.7 the probability might actually be. This is because the degree of overlap between the events $\overline{S}_1, \overline{S}_2, \ldots, \overline{S}_m$ is generally unknown. The probability "at least 0.7" translates into an *overall confidence level* of "at least 70%" when the responses are observed. Similarly, if an experimenter calculates ten confidence intervals each having individual confidence level 95%, then the simultaneous confidence level for the

ten intervals is at least 50%, which is not very informative. As m becomes larger the problem becomes worse, and when $m \geq 20$, the overall confidence level is at least 0%, clearly a useless assertion!

Similar comments apply to the hypothesis testing situation. If hypotheses for m different contrasts are to be tested, each at significance level α^*, then the probability that at least one hypothesis is incorrectly rejected is at most $m\alpha^*$.

Various methods have been developed to ensure that the overall confidence level is not too small and the overall significance level is not too high. Some methods are completely general, that is, they can be used for any set of estimable functions, while others have been developed for very specialized purposes such as comparing each treatment with a control. Which method is best depends on which contrasts are of interest and the number of contrasts to be investigated. In this section, four methods are discussed that control the overall confidence level and overall significance level. The terms *preplanned contrasts* and *data snooping* occur in the summary of methods and the subsequent subsections. These have the following meanings. Before the experiment commences, the experimenter will have written out a checklist, highlighted the contrasts and/or treatment means that are of special interest, and designed the experiment in such a way as to ensure that these are estimable with as small variances as possible. These are the preplanned contrasts and means. After the data have been collected, the experimenter usually looks carefully at the data to see whether anything unexpected has occurred. One or more unplanned contrasts may turn out to be the most interesting, and the conclusions of the experiment may not be as anticipated. Allowing the data to suggest additional interesting contrasts is called data snooping.

The following summary is written in terms of confidence intervals, but it also applies to hypothesis tests. A shorter confidence interval corresponds to a more powerful hypothesis test. The block designs mentioned in the summary will be discussed in Chaps. 10 and 11.

Summary of Multiple Comparison Methods

1. **Bonferroni method for preplanned comparisons**
   Applies to any m preplanned estimable contrasts or functions of the parameters. Gives shorter confidence intervals than the other methods listed if m is small. Can be used for any design. Cannot be used for data snooping.

2. **Scheffé method for all comparisons**
   Applies to any m estimable contrasts or functions of the parameters. Gives shorter intervals than Bonferroni's method if m is large. Allows data snooping. Can be used for any design.

3. **Tukey method for all pairwise comparisons**
   Best for all pairwise comparisons. Can be used for completely randomized designs, randomized block designs, and balanced incomplete block designs. Is believed to be applicable (conservative) for other designs as well. Can be extended to include all contrasts, but Scheffé's method is generally better for these.

4. **Dunnett method for treatment-versus-control comparisons**
   Best for all treatment-versus-control contrasts. Can be used for completely randomized designs, randomized block designs, and balanced incomplete block designs.

Details of confidence intervals obtained by each of the above methods are given in Sects. 4.4.2–4.4.6. The terminology "a set of simultaneous $100(1 - \alpha)\%$ confidence intervals" will always refer to the fact that the *overall* confidence level for a set of contrasts or treatment means is (at least) $100(1 - \alpha)\%$. Each of the four methods discussed gives confidence intervals of the form

$$\sum_i c_i \tau_i \in \left(\sum_i c_i \hat{\tau}_i \pm w \sqrt{\widehat{\mathrm{Var}(\Sigma c_i \hat{\tau}_i)}} \right), \tag{4.4.20}$$

where w, which we call the *critical coefficient*, depends on the method, on v, on the number of confidence intervals calculated, and on the number of error degrees of freedom. The term

$$msd = w \sqrt{\widehat{\mathrm{Var}(\Sigma c_i \hat{\tau}_i)}},$$

which is added and subtracted from the least squares estimate in (4.4.20), is called the *minimum significant difference*, because if the estimate is larger than *msd*, the confidence interval excludes zero, and the contrast is significantly different from zero.

4.4.2   Bonferroni Method for Preplanned Comparisons

The inequality (4.4.19) shows that if m simultaneous confidence intervals are calculated for preplanned contrasts, and if each confidence interval has confidence level $100(1 - \alpha^*)\%$, then the overall confidence level is greater than or equal to $100(1 - m\alpha^*)\%$. Thus, an experimenter can ensure that the overall confidence level is at least $100(1 - \alpha)\%$ by setting $\alpha^* = \alpha/m$. This is known as the Bonferroni method for simultaneous confidence intervals. Replacing α by α/m in the formula (4.3.6), p. 75, for an individual confidence interval, we obtain a formula for a set of simultaneous $100(1 - \alpha)\%$ confidence intervals for m preplanned contrasts $\Sigma c_i \tau_i$ in a completely randomized design with the one-way analysis of variance model (3.3.1), as

$$\sum_i c_i \tau_i \in \left(\sum_i c_i \overline{y}_{i.} \pm t_{n-v, \alpha/(2m)} \sqrt{msE \sum_i c_i^2 / r_i} \right), \tag{4.4.21}$$

where the critical coefficient, w_B, is

$$w_B = t_{n-v, \alpha/(2m)}.$$

Since $\alpha/(2m)$ is likely to be an atypical value, the percentiles $t_{n-v, \alpha/(2m)}$ may need to be obtained by use of a computer package, or by approximate interpolation between values in Table A.4, or by using the following approximate formula due to Peiser (1943):

$$t_{df, \alpha/(2m)} \approx z_{\alpha/(2m)} + (z_{\alpha/(2m)}^3 + z_{\alpha/(2m)})/(4(df)), \tag{4.4.22}$$

where *df* is the error degrees of freedom (equal to $n - v$ in the present context), and where $z_{\alpha/(2m)}$ is the percentile of the standard normal distribution corresponding to a probability of $\alpha/(2m)$ in the right hand tail. The standard normal distribution is tabulated in Table A.3 and covers the entire range of values for $\alpha/(2m)$. When m is very large, $\alpha/(2m)$ is very small, possibly resulting in extremely wide simultaneous confidence intervals. In this case, the Scheffé or Tukey methods described in the following subsections would be preferred.

If some of the m simultaneous intervals are for true mean responses $\mu + \tau_s$, then the required intervals are of the form (4.3.12), p. 76, with α replaced by α/m, that is,

$$\mu + \tau_s \in \left(\overline{y}_{s.} \pm t_{n-v, \alpha/(2m)} \sqrt{msE/r_s} \right). \tag{4.4.23}$$

Similarly, replacing α by α/m in (4.3.15), a set of m null hypotheses, each of the form

$$H_0 : \sum_{i=1}^{v} c_i \tau_i = 0 \,,$$

can be tested against their respective two-sided alternative hypotheses at overall significance level α using the set of decision rules each of the form

$$\text{reject } H_0 \text{ if } \frac{ssc}{msE} > F_{1,df,\alpha/m} \,. \tag{4.4.24}$$

Each null hypothesis is rejected if the corresponding confidence interval (4.4.21) excludes zero, and each confidence interval consists of exactly those values that would not be rejected by a two-tailed test.

Note that Bonferroni's method can be use only for *preplanned* contrasts and means. An experimenter who looks at the data and then proceeds to calculate simultaneous confidence intervals for the few contrasts that look interesting has effectively calculated a very large number of intervals. This is because the interesting contrasts are usually those that seem to be significantly different from zero, and a rough mental calculation of the estimates of a large number of contrasts has to be done to identify these interesting contrasts. Scheffé's method should be used for contrasts that were selected after the data were examined.

Example 4.4.1 `Filter experiment, continued`

The filter experiment was described in Example 4.3.3, p. 78. Suppose that before the data had been collected, the experimenters had planned to calculate a set of simultaneous 90% confidence intervals for the following $m = 3$ contrasts. These contrasts have been selected based on the details of the original study described by Lorenz et al. (1982).

(i) $\frac{1}{3}(\tau_1 + \tau_4 + \tau_7) - \frac{1}{4}(\tau_2 + \tau_3 + \tau_5 + \tau_6)$. This contrast measures the difference in the average effect of the presterilized and the nonpresterilized filter types. This was used in Example 4.3.3 to illustrate a hypothesis test for a single contrast.

(ii) $\frac{1}{2}(\tau_1 + \tau_7) - \frac{1}{5}(\tau_2 + \tau_3 + \tau_4 + \tau_5 + \tau_6)$. This contrast measures the difference in the average effects of two filter types with gradated pore size and five filter types with uniform pore size.

(iii) $\frac{1}{6}(\tau_1 + \tau_2 + \tau_4 + \tau_5 + \tau_6 + \tau_7) - \tau_3$. This contrast is the difference in the average effect of the filter types that are recommended by their manufacturers for bacteriologic analysis of water and the single filter type that is recommended for sterility testing of pharmaceutical or cosmetic products.

From Example 4.3.3, we know that

$$\bar{y}_{1.} = 36.0, \quad \bar{y}_{2.} = 18.0, \quad \bar{y}_{3.} = 27.7, \quad \bar{y}_{4.} = 28.0, \quad \bar{y}_{5.} = 28.3,$$
$$\bar{y}_{6.} = 37.7, \quad \bar{y}_{7.} = 30.3, \quad r_i = 3, \quad msE = 21.6.$$

The formula for each of the three preplanned simultaneous 90% confidence intervals is given by (4.4.21) and involves the critical coefficient $w_B = t_{14,(0.1)/6} = t_{14,0.0167}$, which is not available in Table A.4. Either the value can be calculated from a computer program, or an approximate value can be obtained from formula (4.4.22) as

$$t_{14,0.0167} \approx 2.128 + (2.128^3 + 2.128)/(4 \times 14) = 2.338.$$

The minimum significant difference for each of the three simultaneous 90% confidence intervals is

$$msd \; = \; 2.338\sqrt{(21.6) \sum c_i^2/3} \; = \; 6.2735\sqrt{\sum c_i^2} \, .$$

Thus, for the first interval, we have

$$msd \; = \; 6.2735\sqrt{3\left(\frac{1}{9}\right) + 4\left(\frac{1}{16}\right)} \; = \; 4.791 \, ,$$

giving the interval as

$$\frac{1}{3}(\tau_1 + \tau_4 + \tau_7) - \frac{1}{4}(\tau_2 + \tau_3 + \tau_5 + \tau_6) \in (3.508 \pm 4.791) \;\; = \;\; (-1.283, \; 8.299) \, .$$

Calculating the minimum significant differences separately for the other two confidence intervals leads to

$$\frac{1}{2}(\tau_1 + \tau_7) - \frac{1}{5}(\tau_2 + \tau_3 + \tau_4 + \tau_5 + \tau_6) \in (-0.039, \; 10.459) \, ;$$

$$\frac{1}{6}(\tau_1 + \tau_2 + \tau_4 + \tau_5 + \tau_6 + \tau_7) - \tau_3 \in (-4.759, \; 8.793) \, .$$

Notice that all three intervals include zero, although the second is close to excluding it. Thus, at overall significance level $\alpha = 0.10$, we would fail to reject the hypothesis that there is no difference in average colony counts between the presterilized and nonpresterilized filters, nor between filter 3 and the others, nor between filters with gradated and uniform pore sizes. At a slightly higher significance level, we would reject the hypothesis that the filters with gradated pore size have the same average colony counts as those with uniform pore size. The same conclusion would be obtained if (4.4.24) were used to test simultaneously, at overall level $\alpha = 0.10$, the hypotheses that each of the three contrasts is zero. The confidence interval, whether utilized directly or obtained as the conclusion of the test, has the added benefit that it provides more specific conclusions. For example, we can say with overall 90% confidence that on average, the filters with gradated pore size give rise to colony counts up to 10.4 greater than the filters with uniform pore sizes.   □

4.4.3   Scheffé Method of Multiple Comparisons

The main drawbacks of the Bonferroni method of multiple comparisons are that the m contrasts to be examined must be preplanned and the confidence intervals can become very wide if m is large. Scheffé's method, on the other hand, provides a set of simultaneous $100(1 - \alpha)\%$ confidence intervals whose widths are determined only by the number of treatments and the number of observations in the experiment, no matter how many contrasts are of interest. The two methods are compared directly later in this section.

Scheffé's method is based on the fact that every possible contrast $\Sigma c_i \tau_i$ can be written as a linear combination of the set of $(v - 1)$ treatment versus control contrasts, $\tau_2 - \tau_1, \tau_3 - \tau_1, \ldots, \tau_v - \tau_1$. (We leave it to the reader to check that this is true.) Once the experimental data have been collected, it is possible to find a $100(1 - \alpha)\%$ *confidence region* for these $v - 1$ treatment-versus-control contrasts.

The confidence region not only determines confidence bounds for each treatment-versus-control contrast, it determines bounds for *every* possible contrast $\Sigma c_i \tau_i$ and, in fact, for *any number* of contrasts, while the overall confidence level remains fixed. The mathematical details are given by Scheffé (1959).

For v treatments in a completely randomized design and the one-way analysis of variance model (3.3.1), a set of simultaneous $100(1 - \alpha)\%$ confidence intervals for all contrasts $\Sigma c_i \tau_i$ is given by

$$\sum_i c_i \tau_i \in \left(\sum_i c_i \bar{y}_{i.} \pm \sqrt{(v-1)F_{v-1,n-v,\alpha}} \sqrt{msE \sum_i c_i^2/r_i} \right) . \tag{4.4.25}$$

Notice that this is the same form as the general formula (4.4.20), p. 83, where the critical coefficient w is

$$w_S = \sqrt{(v-1)F_{v-1,n-v,\alpha}} \ .$$

If confidence intervals for the treatment means $\mu + \tau_i$ are also of interest, the critical coefficient w_S needs to be replaced by

$$w_S^* = \sqrt{v F_{v,n-v,\alpha}} .$$

The reason for the increase in the numerator degrees of freedom is that any of the functions $\mu + \tau_i$ can be written as a linear combination of the $v - 1$ treatment versus control contrasts and one additional function $\mu + \tau_1$. For the completely randomized design and model (3.3.1), a set of simultaneous $100(1 - \alpha)\%$ confidence intervals for any number of true mean responses and contrasts is therefore given by

$$\sum_i c_i \tau_i \in \left(\sum_i c_i \bar{y}_{i.} \pm \sqrt{v F_{v,n-v,\alpha}} \sqrt{msE \sum_i c_i^2/r_i} \right)$$

together with

$$\mu + \tau_s \in \left(\bar{y}_{i.} \pm \sqrt{v F_{v,n-v,\alpha}} \sqrt{msE/r_s} \right) . \tag{4.4.26}$$

Example 4.4.2 Filter experiment, continued

If we look at the observed average colony counts,

$$\bar{y}_{1.} = 36.0, \ \ \bar{y}_{2.} = 18.0, \ \ \bar{y}_{3.} = 27.7, \ \ \bar{y}_{4.} = 28.0,$$
$$\bar{y}_{5.} = 28.3, \ \ \bar{y}_{6.} = 37.7, \ \ \bar{y}_{7.} = 30.3,$$

for the filter experiment of Examples 4.3.3 and 4.4.1 (p. 78 and 84), filter type 2 appears to give a much lower count than the other types. One may wish to recalculate each of the three intervals in Example 4.4.1 with filter type 2 excluded. It might also be of interest to compare the filter types 1 and 6, which showed the highest average colony counts, with the other filters. These are *not preplanned* contrasts. They have become interesting only after the data have been examined, and therefore we need to use Scheffé's method of multiple comparisons. In summary, we are interested in the following twelve contrasts:

$$\tfrac{1}{3}(\tau_1 + \tau_4 + \tau_7) - \tfrac{1}{3}(\tau_3 + \tau_5 + \tau_6), \ \ \ \ \tfrac{1}{2}(\tau_1 + \tau_7) - \tfrac{1}{4}(\tau_3 + \tau_4 + \tau_5 + \tau_6),$$
$$\tfrac{1}{5}(\tau_1 + \tau_4 + \tau_5 + \tau_6 + \tau_7) - \tau_3,$$
$$\tau_1 - \tau_3, \ \ \ \ \tau_1 - \tau_4, \ \ \ \ \tau_1 - \tau_5, \ \ \ \ \tau_1 - \tau_6, \ \ \ \ \tau_1 - \tau_7,$$
$$\tau_6 - \tau_3, \ \ \ \ \tau_6 - \tau_4, \ \ \ \ \tau_6 - \tau_5, \ \ \ \ \tau_6 - \tau_7.$$

The formula for a set of Scheffé 90% simultaneous confidence intervals is given by (4.4.25) with $\alpha = 0.10$. Since $v = 7$, $n = 21$, and $msE = 21.6$ for the filter experiment, the minimum significant difference for each interval becomes

$$msd = \sqrt{6F_{6,14,0.10}}\sqrt{21.6 \ \Sigma c_i^2/3} = 9.837\sqrt{\Sigma c_i^2} \ .$$

The twelve simultaneous 90% confidence intervals are then

$$\frac{1}{3}(\tau_1 + \tau_4 + \tau_7) - \frac{1}{3}(\tau_3 + \tau_5 + \tau_6)$$
$$\in \ \left((31.43 - 31.23) \ \pm \ 9.837\sqrt{3\left(\frac{1}{9}\right) + 3\left(\frac{1}{9}\right)}\right)$$
$$= \ (-7.83, 8.23) \, ,$$

$$\frac{1}{2}(\tau_1 + \tau_7) - \frac{1}{4}(\tau_3 + \tau_4 + \tau_5 + \tau_6) \in (-5.79, 11.24) \, ,$$

$$\frac{1}{5}(\tau_1 + \tau_4 + \tau_5 + \tau_6 + \tau_7) - \tau_3 \in (-6.42, 15.14) \, ,$$

$$\begin{aligned}
\tau_1 - \tau_3 &\in \ (-5.61, 22.21) \, , \quad \tau_6 - \tau_3 \in (-3.91, 23.91) \, , \\
\tau_1 - \tau_4 &\in \ (-5.91, 21.91) \, , \quad \tau_6 - \tau_4 \in (-4.21, 23.61) \, , \\
\tau_1 - \tau_5 &\in \ (-6.21, 21.61) \, , \quad \tau_6 - \tau_5 \in (-4.51, 23.31) \, , \\
\tau_1 - \tau_6 &\in (-15.61, 12.21) \, , \quad \tau_6 - \tau_7 \in (-6.51, 21.31) \, , \\
\tau_1 - \tau_7 &\in \ (-8.21, 19.61) \, .
\end{aligned}$$

These intervals are all fairly wide and all include zero. Consequently, at overall error rate $\alpha = 0.1$, we are unable to infer that any of the contrasts are significantly different from zero.   □

Relationship Between Analysis of Variance and the Scheffé Method

The analysis of variance test and the Scheffé method of multiple comparisons are equivalent in the following sense. The analysis of variance test will reject the null hypothesis $H_0 : \tau_1 = \tau_2 = \cdots = \tau_v$ at significance level α if there is at least one confidence interval among the infinite number of Scheffé simultaneous $100(1 - \alpha)\%$ confidence intervals for all contrasts $\Sigma c_i \tau_i$ that excludes zero. However, the intervals that exclude zero may not be among those for the interesting contrasts being examined.

Other methods of multiple comparisons do not relate to the analysis of variance test in this way. It is possible when using one of the other multiple comparison methods that one or more intervals in a simultaneous $100(1 - \alpha)\%$ set may exclude 0, while the analysis of variance test of H_0 is *not* rejected at significance level α. Hence, if specific contrasts of interest have been identified in advance of running the experiment and a method of multiple comparisons other than Scheffé's method is to be used, then it is sensible to analyze the data using only the multiple comparison procedure.

4.4.4   Tukey Method for All Pairwise Comparisons

In some experiments, confidence intervals may be required only for pairwise difference contrasts. Tukey, in 1953, proposed a method that is specially tailored to handle this situation and that gives shorter intervals for pairwise differences than do the Bonferroni and Scheffé methods.

For the completely randomized design and the one-way analysis of variance model (3.3.1), Tukey's simultaneous confidence intervals for all pairwise comparisons $\tau_i - \tau_s$, $i \neq s$, with overall confidence level at least $100(1 - \alpha)\%$ is given by

$$\tau_i - \tau_s \in \left((\overline{y}_{i.} - \overline{y}_{s.}) \pm w_T \sqrt{msE \left(\frac{1}{r_i} + \frac{1}{r_s} \right)} \right), \tag{4.4.27}$$

where the critical coefficient w_T is

$$w_T = q_{v,n-v,\alpha}/\sqrt{2},$$

and where $q_{v,n-v,\alpha}$ is tabulated in Appendix A.8. When the sample sizes are equal ($r_i = r$; $i = 1, \ldots, v$), the overall confidence level is *exactly* $100(1 - \alpha)\%$. When the sample sizes are unequal, the confidence level is *at least* $100(1 - \alpha)\%$.

The derivation of (4.4.27) is as follows. For equal sample sizes, the formula for Tukey's simultaneous confidence intervals is based on the distribution of the statistic

$$Q = \frac{\max\{T_i\} - \min\{T_i\}}{\sqrt{MSE/r}},$$

where $T_i = \overline{Y}_{i.} - (\mu + \tau_i)$ for the one-way analysis of variance model (3.3.1), and where $\max\{T_i\}$ is the maximum value of the random variables $T_1, T_2, \ldots, T_v$ and $\min\{T_i\}$ the minimum value. Since the $\overline{Y}_{i.}$'s are independent, the numerator of Q is the range of v independent $N(0, \sigma^2/r)$ random variables, and is standardized by the estimated standard deviation. The distribution of Q is called the *Studentized range distribution*. The percentile corresponding to a probability of α in the right-hand tail of this distribution is denoted by $q_{v,n-v,\alpha}$, where v is the number of treatments being compared, and $n - v$ is the number of degrees of freedom for error. Therefore,

$$P \left(\frac{\max\{T_i\} - \min\{T_i\}}{\sqrt{MSE/r}} \leq q_{v,n-v,\alpha} \right) = 1 - \alpha.$$

Now, if $\max\{T_i\} - \min\{T_i\}$ is less than or equal to $q_{v,n-v,\alpha}\sqrt{MSE/r}$, then it must be true that $|T_i - T_s| \leq q_{v,n-v,\alpha}\sqrt{MSE/r}$ for *every* pair of random variables $T_i, T_s, i \neq s$. Using this fact and the above definition of T_i, we have

$$1 - \alpha = P \left(-q_{v,n-v,\alpha}\sqrt{MSE/r} \leq (\overline{Y}_{i.} - \overline{Y}_{s.}) - (\tau_i - \tau_s) \right.$$
$$\left. \leq q_{v,n-v,\alpha}\sqrt{MSE/r}, \text{ for all } i \neq s \right).$$

Replacing $\overline{Y}_{i.}$ by its observed value $\overline{y}_{i.}$, and MSE by the observed value msE, a set of simultaneous $100(1 - \alpha)\%$ confidence intervals for all pairwise differences $\tau_i - \tau_s$, $i \neq s$, is given by

$$\tau_i - \tau_s \in \left((\overline{y}_{i.} - \overline{y}_{s.}) \pm q_{v,n-v,\alpha}\sqrt{msE/r} \right),$$

which can be written in terms of the critical coefficient as

$$\tau_i - \tau_s \in \left((\overline{y}_{i.} - \overline{y}_{s.}) \pm w_T \sqrt{msE \left(\frac{1}{r} + \frac{1}{r} \right)} \right). \tag{4.4.28}$$

More recently, Hayter ([1984](#)) showed that the same form of interval can be used for unequal sample sizes as in (4.4.27), and that the overall confidence level is then at least $100(1 - \alpha)\%$.

Example 4.4.3 `Battery experiment, continued`

In the battery experiment of Example [4.2.1](#) (p. 71), we considered the pairwise differences in the life lengths per unit cost of $v = 4$ different battery types, and we obtained the least squares estimates

$$\hat{\tau}_1 - \hat{\tau}_2 = -289.75\,, \quad \hat{\tau}_1 - \hat{\tau}_3 = 137.75\,, \quad \hat{\tau}_1 - \hat{\tau}_4 = 74.50\,,$$
$$\hat{\tau}_2 - \hat{\tau}_3 = 427.50\,, \quad \hat{\tau}_2 - \hat{\tau}_4 = 364.25\,, \quad \hat{\tau}_3 - \hat{\tau}_4 = -63.25\,.$$

The standard error was $\sqrt{msE(\frac{1}{4} + \frac{1}{4})} = 34.41$, and the number of error degrees of freedom was $n - v = (16 - 4) = 12$. From Table A.8, $q_{4,12,0.05} = 4.20$, so $w_T = 4.20/\sqrt{2}$, and the minimum significant difference is

$$msd = (4.20/\sqrt{2})\,(34.41) = 102.19\,.$$

Therefore, using Tukey's method, the simultaneous 95% confidence intervals for the pairwise comparisons of lifetimes per unit cost of the different battery types are

$$\tau_1 - \tau_2 \in (-289.75 \pm 102.19) = (-391.94, -187.56),$$
$$\tau_1 - \tau_3 \in (137.75 \pm 102.19) = (35.56, 239.94),$$
$$\tau_1 - \tau_4 \in (-27.69, 176.69), \qquad \tau_2 - \tau_3 \in (325.31, 529.69),$$
$$\tau_2 - \tau_4 \in (262.06, 466.44), \qquad \tau_3 - \tau_4 \in (-165.44, 38.94).$$

Four of these intervals exclude zero, and one can conclude (at an overall 95% confidence level) that battery type 2 (alkaline, store brand) has the highest lifetime per unit cost, and battery type 3 (heavy duty, name brand) has lower lifetime per unit cost than does battery type 1 (alkaline, name brand). The intervals show us that with overall 95% confidence, battery type 2 is between 188 and 391 minute per dollar better than battery type 1 (the name brand alkaline battery) and even more economical than the heavy-duty brands.   □

Example 4.4.4 `Bonferroni, Scheffé and Tukey methods compared`

Suppose that $v = 5$, $n = 35$, and $\alpha = 0.05$, and that only the 10 pairwise comparisons $\tau_i - \tau_s$, $i \neq s$, are of interest to the experimenter and these were specifically selected prior to the experiment (i.e., were preplanned). If we compare the critical coefficients for the three methods, we obtain

$$\text{Bonferroni}: w_B \;=\; t_{30,.025/10} \;=\; 3.02,$$
$$\text{Scheffé}: \quad w_S \;=\; \sqrt{4\,F_{4,30,.05}} \;=\; 3.28,$$
$$\text{Tukey}: \quad w_T \;=\; \tfrac{1}{\sqrt{2}}q_{5,30,.05} \;=\; 2.91.$$

Since w_T is less than w_B, which is less than w_S for this example, the Tukey intervals will be shorter than the Bonferroni intervals, which will be shorter than the Scheffé intervals.   □

4.4.5   Dunnett Method for Treatment-Versus-Control Comparisons

In 1955, Dunnett developed a method of multiple comparisons that is specially designed to provide a set of simultaneous confidence intervals for preplanned treatment-versus-control contrasts $\tau_i - \tau_1$ ($i = 2, \ldots, v$), where level 1 corresponds to the control treatment. The intervals are shorter than those given by the Scheffé, Tukey, and Bonferroni methods, but the method should not be used for any other type of contrasts.

The formulae for the simultaneous confidence intervals are based on the joint distribution of the estimators $\overline{Y}_{i.} - \overline{Y}_{1.}$ of $\tau_i - \tau_1$ ($i = 2, \ldots, v$). This distribution is a special case of the multivariate t distribution and depends on the correlation between $\overline{Y}_{i.} - \overline{Y}_{1.}$ and $\overline{Y}_{s.} - \overline{Y}_{1.}$. For the completely randomized design, with equal numbers of observations $r_2 = \cdots = r_v = r$ on the experimental treatments and $r_1 = c$ observations on the control treatment, the correlation is

$$\rho = r/(c + r).$$

In many experiments, the same number of observations will be taken on the control and experimental treatments, in which case $\rho = 0.5$. However, the shortest confidence intervals for comparing $v - 1$ experimental treatments with a control treatment are generally obtained when c/r is chosen to be close to $\sqrt{v-1}$. Since we have tabulated the multivariate t-distribution only with correlation $\rho = 0.5$, we will discuss only the case $c = r$. Other tables can be found in the book of Hochberg and Tamhane (1987), and intervals can also be obtained via some computer packages (see Sects. 4.6.2 and 4.7.2 for the SAS and R software, respectively).

If the purpose of the experiment is to determine which of the experimental treatments give a significantly higher response than the control treatment, then one-sided confidence bounds should be used. For a completely randomized design with equal sample sizes and the one-way analysis of variance model (3.3.1), Dunnett's simultaneous one-sided $100(1 - \alpha)\%$ confidence bounds for treatment-versus-control contrasts $\tau_i - \tau_1$ ($i = 2, 3, \ldots, v$) are

$$\tau_i - \tau_1 \geq (\overline{y}_{i.} - \overline{y}_{1.}) - w_{D1} \sqrt{msE\left(\frac{2}{r}\right)}, \tag{4.4.29}$$

where the critical coefficient is

$$w_{D1} = t_{v-1,n-v,\alpha}^{(0.5)}$$

and where $t_{v-1,n-v,\alpha}^{(0.5)}$ is the percentile of the maximum of a multivariate t-distribution with common correlation 0.5 and $n - v$ degrees of freedom, corresponding to a Type I error probability of α in the right-hand tail. The critical coefficient is tabulated in Table A.9. If the right hand side of (4.4.29) is positive, we infer that the ith experimental treatment gives a larger response than the control.

If the purpose is to determine which of the experimental treatments give a significantly lower response than the control, then the inequality is reversed, and the confidence bound becomes

$$\tau_i - \tau_1 \leq (\overline{y}_{i.} - \overline{y}_{1.}) + w_{D1} \sqrt{msE\left(\frac{2}{r}\right)}. \tag{4.4.30}$$

If the right-hand side is negative, we infer that the ith experimental treatment gives a smaller response than the control.

To determine which experimental treatments are better than the control *and* which ones are worse, two-sided intervals of the general form (4.4.20) are used as for the other multiple comparison methods. For the completely randomized design, one-way analysis of variance model (3.3.1), and equal sample sizes, the formula is

$$\tau_i - \tau_1 \in \left(\bar{y}_{i.} - \bar{y}_{1.} \pm w_{D2} \sqrt{msE\left(\frac{2}{r}\right)} \right), \tag{4.4.31}$$

where the critical coefficient is

$$w_{D2} = |t|_{v-1,n-v,\alpha}^{(0.5)}$$

and is the upper critical value for the maximum of the absolute values of a multivariate t-distribution with correlation 0.5 and $n - v$ error degrees of freedom, corresponding to the chosen value of α in the right-hand tail. The critical coefficients for equal sample sizes are provided in Table A.10.

For future reference, the general formula for Dunnett's two-sided simultaneous $100(1 - \alpha)\%$ confidence intervals for treatment versus control contrasts $\tau_i - \tau_1$ $(i = 2, 3, \ldots, v)$ is

$$\tau_i - \tau_1 \in \left((\hat{\tau}_i - \hat{\tau}_1) \pm w_{D2} \sqrt{\widehat{\text{Var}}(\hat{\tau}_i - \hat{\tau}_1)} \right), \tag{4.4.32}$$

and, for one-sided confidence bounds, we replace w_{D2} by w_{D1} and replace "$\in$" by "$\leq$" or "$\geq$." The critical coefficients are

$$w_{D2} = |t|_{v-1,df,\alpha}^{(0.5)} \text{ and } w_{D1} = t_{v-1,df,\alpha}^{(0.5)}$$

for two-sided and one-sided intervals, respectively, where *df* is the number of error degrees of freedom.

Example 4.4.5 `Soap experiment, continued`

Suppose that as a preplanned objective of the soap experiment of Sect. 2.5.1, p. 20, the experimenter had wanted simultaneous 99% confidence intervals comparing the weight losses of the deodorant and moisturizing soaps (levels 2 and 3) with that of the regular soap (level 1). Then it is appropriate to use Dunnett's method as given in (4.4.31). From Sect. 3.7.2, $r_1 = r_2 = r_3 = 4$, $msE = 0.0772$, $\hat{\tau}_2 - \hat{\tau}_1 = 2.7350$, and $\hat{\tau}_3 - \hat{\tau}_1 = 2.0275$. From Table A.10, $w_{D2} = |t|_{v-1,n-v,\alpha}^{(0.5)} = |t|_{2,9,0.01}^{(0.5)} = 3.63$, so the minimum significant difference is

$$msd = 3.63 \sqrt{msE(2/4)} = 0.713 .$$

Hence, the simultaneous 99% confidence intervals are

$$\tau_2 - \tau_1 \in (2.7350 \pm 0.713) \approx (2.022, 3.448)$$

and

$$\tau_3 - \tau_1 \in (2.0275 \pm 0.713) \approx (1.314, 2.741) .$$

One can conclude from these intervals (with overall 99% confidence) that the deodorant soap (soap 2) loses between 2 and 3.4 g more weight on average than does the regular soap, and the moisturizing soap loses between 1.3 and 2.7 g more weight on average than the regular soap. We leave it to the reader to verify that neither the Tukey nor the Bonferroni method would have been preferred for these contrasts (see Exercise 7). $\square$

4.4.6 Combination of Methods

The Bonferroni method is based on the fact that if m individual confidence intervals are obtained, each with confidence level $100(1 - \alpha^*)\%$, then the overall confidence level is at least $100(1 - m\alpha^*)\%$. The same fact can be used to combine the overall confidence levels arising from more than one multiple comparison procedure.

In Example 4.4.1 (p. 84), the Bonferroni method was used to calculate simultaneous 90% confidence intervals for $m = 3$ preplanned contrasts. In Example 4.4.2 (p. 86), the analysis was continued by calculating simultaneous 90% Scheffé intervals for twelve other contrasts. The overall error rate for these two sets of intervals combined is therefore at most $0.1 + 0.1 = 0.2$, giving an overall, or "experimentwise," confidence level of at least $100(1 - 0.2)\% = 80\%$ for all fifteen intervals together.

Different possible strategies for multiple comparisons should be examined when outlining the analysis at step (g) of the checklist (Sect. 2.2, p. 7). Suppose that in the above example the overall level for all intervals (both planned and otherwise) had been required to be at least 90%. We examine two possible strategies that could have been used. First, the confidence levels for the Bonferroni and Scheffé contrasts could have been adjusted, dividing $\alpha = 0.10$ into two pieces, α_1 for the preplanned contrasts and α_2 for the others, where $\alpha_1 + \alpha_2 = 0.10$. This strategy would have resulted in intervals that were somewhat wider than the above for all of the contrasts. Alternatively, Scheffé's method could have been used with $\alpha = 0.10$ for all of the contrasts including the three preplanned contrasts. This strategy would have resulted in wider intervals for the three preplanned contrasts but not for the others. Both strategies would result in an overall, or experimentwise, confidence level of 90% instead of 80%.

4.4.7 Methods Not Controlling Experimentwise Error Rate

We have introduced four methods of multiple comparisons, each of which allows the experimenter to control the overall confidence level, and the same methods can be used to control the experimentwise error rate when multiple hypotheses are to be tested. There exist other multiple comparison procedures that are more powerful (i.e., that more easily detect a nonzero contrast) but do not control the overall confidence level nor the experimentwise error rate. While some of these are used quite commonly, we do not advocate their use. Such procedures include Duncan's multiple range test, Fisher's protected LSD procedure, and the Newman–Keuls method. (For more details, see Hsu 1996.)

4.5 Sample Sizes

Before an experiment can be run, it is necessary to determine the number of observations that should be taken on each level of each treatment factor (step (h) of the checklist in Sect. 2.2, p. 7). In Sect. 3.6.2, a method was presented to calculate the sample sizes needed to achieve a specified power of the test of the hypothesis $H_0 : \tau_1 = \cdots = \tau_v$. In this section we show how to determine the sample sizes to achieve confidence intervals of specified lengths.

The lengths of confidence intervals decrease as sample sizes increase. Consequently, if the length of an interval is specified, it should be possible to calculate the required sample sizes, especially when these are equal. However, there is a problem. Since the experimental data have not yet been collected, the value of the mean squared error is not known. As in Sect. 3.6.2, if the value of the mean squared error can be reasonably well be guessed at, either from previous experience or from a pilot study, then a trial and error approach to the problem can be followed, as illustrated in the next example.

Example 4.5.1 `Bean-soaking experiment`

Suppose we were to plan an experiment to compare the effects of $v = 5$ different soaking times on the growth rate of mung bean seeds. The response variable will be the length of a shoot of a mung bean seed 48 hours after soaking. Suppose that a pilot experiment has indicated that the mean square for error is likely to be not more than $10\,\text{mm}^2$, and suppose that we would like a set of 95% simultaneous confidence intervals for pairwise differences of the soaking times, with each interval no wider than $6\,\text{mm}$ (that is, the half width or minimum significant difference should be no greater than $3\,\text{mm}$).

The formula for each of the simultaneous confidence intervals for pairwise comparisons using Tukey's method of multiple comparisons is given by (4.4.27) p. 88. For equal sample sizes, the interval half width, or minimum significant difference, is required to be at most $3\,\text{mm}$; that is, we require

$$msd = w_T \sqrt{10 \left(\frac{1}{r} + \frac{1}{r} \right)} \leq 3\,,$$

where $w_T = q_{5,5r-5,.05}/\sqrt{2}$ or, equivalently,

$$q^2_{5,5r-5,.05} \leq 0.9r\,.$$

Adopting a trial-and-error approach, we guess a value for r, say $r = 10$. Then, from Table A.8, we find $q^2_{5,45,.05} \approx 4.03^2 = 16.24$, which does not satisfy the requirement that $q^2 \leq 0.9r = 9$. A larger value for r is needed, and we might try $r = 20$ next. The calculations are most conveniently laid out in table form, as follows.

r	$5r-5$	$q^2_{5,5r-5,0.05}$	$0.9r$	*Action*
10	45	$4.03^2 = 16.24$	9.00	Increase r
20	95	$3.95^2 = 15.60$	18.00	Decrease r
15	70	$3.97^2 = 15.76$	13.50	Increase r
18	85	$3.96^2 = 15.68$	16.20	Decrease r
17	80	$3.96^2 = 15.68$	15.30	

If $r = 17$ observations are taken on each of the five soaking times, and if the mean square for error is approximately $10\,\text{mm}^2$ in the main experiment, then the 95% Tukey simultaneous confidence intervals for pairwise comparisons will be a little over the required $6\,\text{mm}$ in length. If $r = 18$ observations are taken, the interval will be a little shorter than the $6\,\text{mm}$ required. If the cost of the experiment is high, then $r = 17$ would be selected; otherwise, $r = 18$ might be preferred.

Trial and error procedures such as that illustrated in Example 4.5.1 for Tukey's method of multiple comparisons can be used for any of the other multiple comparison methods to obtain the approximate sample sizes required to meet the objectives of the experiment. The same type of calculation can be done for unequal sample sizes, provided that the relative sizes are specified, for example $r_1 = 2r_2 = 2r_3 = 2r_4$.

Unless more information is desired on some treatments than on others, or unless costs or variances are unequal, it is generally advisable to select equal sample sizes whenever possible. Choosing equal sample sizes produces two benefits! Confidence intervals for pairwise comparisons are all the same length, which makes them easier to compare, and the multiple comparison and analysis of variance procedures are less sensitive to an incorrect assumption of normality of the error variables.

Quite often, the sample size calculation will reveal that the required number of observations is too large to meet the budget or the time restrictions of the experiment. There are several possible remedies:

(a) Refine the experimental procedure to reduce the likely size of *msE*,

(b) Omit one or more treatments,

(c) Allow longer confidence intervals,

(d) Allow a lower confidence level.

4.6   Using SAS Software

In this section we illustrate how to use the SAS software to generate information for confidence intervals and hypothesis tests for individual contrasts and means and also for the multiple comparison procedures. We use the data from the battery experiment of Sect. 2.5.2 (p. 24).

A sample SAS program to analyze the data is given in Table 4.1. As in Chap. 3, line numbers have been included for reference but are not part of the SAS program. A data set BATTERY, with variables TYPE, LPUC and ORDER, is created from the statements in lines 1–9. The treatment factor is the type of battery TYPE, the response variable is the life per unit cost LPUC, and the one-way analysis of variance model (3.3.1) was used for the analysis. Lines 10–12 generate the analysis of variance table shown in the top of Fig. 4.1.

4.6.1   Inferences on Individual Contrasts

The SAS statements ESTIMATE and CONTRAST are part of the GLM procedure and are used for making inferences concerning specific contrasts.

The ESTIMATE statements (lines 13–15 of Table 4.1) generate information for constructing confidence intervals or conducting hypothesis tests for individual contrasts. Each of the three ESTIMATE statements includes a user-selected contrast name in single quotes, together with the name of the factor for which the effects of levels are to be compared, and the coefficients of the contrast to be estimated.

Table 4.1 SAS program for the battery experiment: contrasts and multiple comparisons

```
Line   SAS Program
  1    DATA BATTERY;
  2      INPUT TYPE LPUC ORDER;
  3      LINES;
  4      1 611  1
  5      2 923  2
  6      1 537  3
  7      :   :   :
  8      3 413 16
  9    ;
 10    PROC GLM;
 11      CLASS TYPE;
 12      MODEL LPUC = TYPE;
 13      ESTIMATE 'DUTY'      TYPE  1  1 -1 -1 / DIVISOR = 2;
 14      ESTIMATE 'BRAND'     TYPE  1 -1  1 -1 / DIVISOR = 2;
 15      ESTIMATE 'INTERACTN' TYPE  1 -1 -1  1 / DIVISOR = 2;
 16      CONTRAST 'BRAND'     TYPE  1 -1  1 -1;
 17      LSMEANS TYPE / ADJUST = TUKEY CL PDIFF ALPHA = 0.01;
```

Fig. 4.1 Analysis of variance and output from the `CONTRASTS` and `ESTIMATE` statements

The GLM Procedure

Dependent Variable: LPUC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	427915.2500	142638.4167	60.24	<.0001
Error	12	28412.5000	2367.7083		
Corrected Total	15	456327.7500			

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
BRAND	1	124609.0000	124609.0000	52.63	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
DUTY	251.000000	24.3295516	10.32	<.0001
BRAND	-176.500000	24.3295516	-7.25	<.0001
INTERACTN	-113.250000	24.3295516	-4.65	0.0006

If the contrast coefficients are to be divided by a constant, this is indicated by means of the `DIVISOR` option. The information generated by these statements is shown in the bottom section of Fig. 4.1. The columns show the contrast name, the contrast estimate $\Sigma c_i \overline{y}_{i.}$, the standard error $\sqrt{msE(\Sigma c_i^2 / r_i)}$ for the estimate, the value of the t-statistic for testing the null hypothesis that the contrast is zero (see (4.3.13), p. 77), and the corresponding p-value for a two-tailed test. For each of the contrasts shown in Fig. 4.1, the p-value is at most 0.0006, indicating that all three contrasts are significantly different from zero for any choice of *individual* significance level α^* greater than 0.0006. The overall and individual significance levels should be selected prior to analysis. The parameter estimates and standard errors can be used to construct confidence intervals by hand, using the critical coefficient for the selected multiple comparison methods (see also Sect. 4.6.2).

The `CONTRAST` statement in line 16 of Table 4.1 generates the information shown in the middle portion of Fig. 4.1 that is needed in (4.3.15), p. 77, for testing the single null hypothesis that the brand contrast is zero versus the alternative hypothesis that it is not zero. The "`F Value`" of 52.63 is the square of the "`t Value`" of -7.25 (up to rounding error) for the brand contrast generated by the `ESTIMATE` statement, the two tests (4.3.13) and (4.3.15) being equivalent.

4.6.2   Multiple Comparisons

The `LSMEANS` statement (line 17) in the `GLM` procedure in Table 4.1 can be used to generate the observed least squares means $\overline{y}_{i.}$ for each level of a factor, and to implement the multiple comparisons procedures introduced in Sect. 4.4. Inclusion of the options `ADJUST=TUKEY`, `PDIFF` and `CL` causes the SAS software to use the Tukey method to compare the effects of each pair of levels, providing both p-values for simultaneous testing and confidence limits for simultaneous estimation of all pairwise comparisons. Individual confidence intervals for each treatment mean are also provided as a consequence of the `CL` option. The option `ALPHA=0.01` sets the confidence level at 99%, both for Tukey's

Fig. 4.2 Tukey's method
for the battery experiment

method for the simultaneous pairwise comparisons and for the individual confidence intervals for each
treatment mean. Part of the corresponding SAS output is given in Fig. 4.2.

Other methods of multiple comparisons can also be requested as options in the `LSMEANS` statement
of the `GLM` procedure. For example, the options `ADJUST=BON` and `ADJUST=SCHEFFE` request
all pairwise comparisons using the methods of Bonferroni and Scheffé, respectively. The option
`ADJUST=DUNNETT` requests Dunnett's 2-sided method of comparing all treatments with a con-
trol, the lowest treatment level serving as the control by default. To explicitly specify level 1, say,
as the control, replace `PDIFF` with `PDIFF=CONTROL('1')`. Similarly, replacing `PDIFF` with
`PDIFF=CONTROLU('1')` requests simultaneous lower bounds for the treatment-versus-control con-
trasts $\tau_i - \tau_1$ by Dunnett's method and is useful for "upper-tailed"alternative hypotheses—namely, for
showing which treatments have a larger effect than the control treatment (coded 1). Likewise, the option
`PDIFF=CONTROLL('1')` provides upper bounds useful for "lower-tailed" alternatives—namely, for
showing which treatments have a smaller effect than the control treatment (coded 1).

4.7   Using R Software

In this section we illustrate how to use the R software to generate information for confidence intervals
and hypothesis tests for individual contrasts and means and also for the multiple comparison procedures.
We use the data from the battery experiment of Sect. 2.5.2 (p. 24). The treatment factor is type of battery
`Type`, the response variable is the life per unit cost `LPUC`, and the one-way analysis of variance
model (3.3.1) was used for the analysis.

A sample R program to analyze the data is given in Table 4.2. Line numbers have been included
for the sake of reference but they are not part of the R program. The data are read from file by the
statement in line 1 of the R code. In line 2, the data set is augmented with a new variable, `fType`,
created by converting the numerical variable `Type` to a factor variable. The `head` command in line 3
displays the first 3 lines of the data set. The function `aov` in line 4 fits the one-way analysis of variance
model (3.3.1) to the data, saving the results as the object `model1` for use by subsequent R functions.

Table 4.2   R program for the battery experiment

```
Line   R Code
  1    battery.data = read.table("data/battery.txt", header=T)
  2    battery.data$fType = factor(battery.data$Type)
  3    head(battery.data, 3)

  4    model1 = aov(LPUC ~ fType, data=battery.data) # Fit aov model
  5    anova(model1)  # Display 1-way ANOVA

  6    # Individual contrasts: estimates, CIs, tests
  7    library(lsmeans)
  8    lsmType = lsmeans(model1, ~ fType) # Compute and save lsmeans
  9    levels(battery.data$fType)
 10    summary(contrast(lsmType, list(Duty=c( 1, 1,-1,-1)/2,
 11                                   Brand=c( 1,-1, 1,-1)/2,
 12                                      DB=c( 1,-1,-1, 1)/2)),
 13          infer=c(T,T), level=0.95, side="two-sided")

 14    # Multiple comparisons
 15    confint(lsmType, level=0.90) # Display lsmeans and 90

 16    # Tukey's method
 17    summary(contrast(lsmType, method="pairwise", adjust="tukey"),
 18          infer=c(T,T), level=0.99, side="two-sided")

 19    # Dunnett's method
 20    summary(contrast(lsmType, method="trt.vs.ctrl", adjust="mvt", ref=1),
 21          infer=c(T,T), level=0.99, side="two-sided")
```

The `anova(model1)` function in line 5 generates the analysis of variance data shown in the top of Table 4.3.

4.7.1   Inferences on Individual Contrasts

The `lsmeans` package, loaded in line 7, provides the functionality for computing least squares means and using these for inferences on treatment contrasts. The `lsmeans` statement (line 8) uses the results of the previously fitted model (line 4) to compute least squares means for each battery type (i.e. for each level of `fType`), saving the results as `lsmType`. The `levels` command in line 9 displays the levels of the factor `fType` in order: `"1" "2" "3" "4"`. Using the least squares means saved in line 8, the `summary` and `contrast` functions of the `lsmeans` package (lines 10–13 of Table 4.2) are coupled to generate least squares estimates, tests, and confidence intervals for specified treatment contrasts. For each of these contrasts, the coefficients correspond to the respective levels of `fType` displayed by line 9. In particular, the `contrast` function inputs the least squares means `lsmType` for each battery type plus a list of contrasts, including a name and the coefficients for each, and would generate the information in the middle of Table 4.3 except the confidence limits. The confidence limits are obtained by wrapping the `contrast` function in the `summary` function, for which the option `infer=c(T,T)` requests confidence intervals and tests. The confidence level is optionally specified to be 95% (the default). Two-sided confidence intervals and tests (the default) are also optionally specified, whereas `side="<"` would request upper confidence limits and specify one-sided alternative hypotheses corresponding to the contrasts being less than zero, for example.

Table 4.3 R output: analysis of variance, individual contrasts, and Tukey's method

```
> anova(model1) # Display 1-way ANOVA

Analysis of Variance Table

Response: LPUC
          Df Sum Sq Mean Sq F value  Pr(>F)
fType      3 427915  142638    60.2 1.7e-07
Residuals 12  28412    2368

> # Individual contrasts: estimates, CIs, tests
> library(lsmeans)
> lsmType = lsmeans(model1, ~ fType) # Compute and save lsmeans
> levels(battery.data$fType)

[1] "1" "2" "3" "4"

> summary(contrast(lsmType, list(Duty=c( 1, 1,-1,-1)/2,
+                                Brand=c( 1,-1, 1,-1)/2,
+                                DB=c( 1,-1,-1, 1)/2)),
+         infer=c(T,T), level=0.95, side="two-sided")

 contrast estimate    SE df lower.CL upper.CL t.ratio p.value
 Duty       251.00 24.33 12   197.99   304.01  10.317  <.0001
 Brand     -176.50 24.33 12  -229.51  -123.49  -7.255  <.0001
 DB        -113.25 24.33 12  -166.26   -60.24  -4.655  0.0006

Confidence level used: 0.95

> # Tukey's method
> summary(contrast(lsmType, method="pairwise", adjust="tukey"),
+         infer=c(T,T), level=0.99, side="two-sided")

 contrast estimate     SE df  lower.CL upper.CL t.ratio p.value
 1 - 2     -289.75 34.407 12 -423.6021 -155.898  -8.421  <.0001
 1 - 3      137.75 34.407 12    3.8979  271.602   4.004  0.0082
 1 - 4       74.50 34.407 12  -59.3521  208.352   2.165  0.1882
 2 - 3      427.50 34.407 12  293.6479  561.352  12.425  <.0001
 2 - 4      364.25 34.407 12  230.3979  498.102  10.586  <.0001
 3 - 4      -63.25 34.407 12 -197.1021   70.602  -1.838  0.3035

Confidence level used: 0.99
Conf-level adjustment: tukey method for comparing a family of 4 estimates
P value adjustment: tukey method for comparing a family of 4 estimates
```

Lines 6–13 of the R code are reproduced in the middle of Table 4.3, along with the corresponding output. The output for each listed contrast includes the contrast name, the estimate $\Sigma c_i \bar{y}_{i.}$, the standard error $\sqrt{msE(\Sigma c_i^2 / r_i)}$ of the estimate, the number of error degrees of freedom, the 95% confidence interval for the treatment contrast, the value of the t-statistic for testing the null hypothesis that the contrast is zero (see (4.3.13) p. 77), and the corresponding p-value for a two-tailed test. For each of the contrasts shown in Table 4.3, the p-value is less than 0.0006, indicating that all three contrasts are significantly different from zero for any choice of *individual* significance level α^* greater than 0.0006. The overall and individual significance levels should be selected prior to analysis. For multiple

comparisons including non-pairwise comparisons, the contrast estimates and standard errors could be used to construct confidence intervals by hand, using the critical coefficient for the selected multiple comparison methods (see Sect. 4.6.2). Pairwise comparisons will be illustrated in the next section.

4.7.2 Multiple Comparisons

Multiple comparisons procedures introduced in Sect. 4.4 are implemented by the R code in lines 14–21 of Table 4.2. The least squares means package `lsmeans`, loaded in line 7, provides functions to generate the observed least squares means $\overline{y}_{i.}$ for each level of a factor, and also to implement the multiple comparisons procedures introduced in Sect. 4.4. In line 8, `lsmeans` uses the information stored in `model1` to compute the least squares mean for each battery type, saving the least squares means as `lsmType` for subsequent use. The `confint` function in line 15 would display the least squares means and corresponding individual 90% confidence intervals for the treatment means (not shown).

Multiple comparison methods can be implemented by coupling the `summary` and `contrast` functions, as illustrated for Tukey's method in lines 16–18. These code lines and the corresponding output are shown in the bottom of Table 4.3. The option `method="pairwise"` requests all pairwise comparisons. The `contrast` statement embedded in line 17 would apply Tukey's method to test whether each pairwise comparison is zero. One also gets the corresponding Tukey confidence intervals by including the `summary` function and its options, where `infer=c(T,T)` requests confidence intervals as well as tests, `level=0.99` sets the confidence level, and `side="two-sided"` requests two-sided confidence intervals and tests. Tukey's method and two-sided inferences are the defaults for all pairwise comparisons, so `adjust="tukey"` and `side="two-sided"` are redundant here, but one can replace `"tukey"` with `"scheffe"` or `"bonferroni"` to apply the corresponding method, or with `"none"` for no multiple comparisons adjustment. The default confidence level is 95%. Using `method="revpairwise"` reverses the order of the pairwise comparisons, considering $\tau_j - \tau_i$ rather than $\tau_i - \tau_j$.

Implementation of Dunnett's method for all treatment-versus-control comparisons is similar and illustrated by lines 19–21 of Table 4.2. The option `method="trt.vs.ctrl"` yields all treatment-versus-control comparisons (not shown). Dunnett's method uses critical values from the multivariate t-distribution, corresponding to `adjust="mvt"`. These critical values are computed by simulation, so the results vary slightly from run to run unless a simulation seed is specified. Also, if the number of treatments is large, implementation of R functions for the multivariate t-distribution may be slow or simply not work, so the default option `adjust="dunnettx"` provides an approximation of Dunnett's method for two-sided confidence intervals that runs faster and dependably, though it is only applicable when the contrast estimates have pairwise correlations of 0.5 such as in the equireplicate case. The first level of the factor, `"1"`, which happens to be level 1 in this case, is the control by default; the syntax `ref=1` illustrates how to specify the first (or any) level as the control. Also, `"two-sided"` is the default for confidence intervals and tests, but one can specify `side="<"` for the one-sided alternative $H_A : \tau_i < \tau_1$ and the corresponding upper confidence bound for $\tau_i - \tau_1$, or `side=">"` for the alternative $H_A : \tau_i > \tau_1$ and the corresponding lower confidence bound for $\tau_i - \tau_1$.

For additional functionality for multiple comparisons procedures, see the multiple comparisons package `multcomp`.

Exercises

1. **Buoyancy experiment**

   Consider conducting an experiment to investigate the question, "Is the buoyancy of an object in water affected by different concentrations of salt in the water?"

   (a) Complete steps (a)–(d) of the checklist (p. 7) in detail. Specify any preplanned contrasts or functions that should be estimated. State, with reasons, which, if any, methods of multiple comparisons will be used.
   (b) Run a small pilot experiment to obtain a preliminary estimate of σ^2.
   (c) Finish the checklist.

2. **Cotton-spinning experiment, continued**

   For the cotton-spinning experiment of Sect. 2.3, p. 13, identify any contrasts or functions that you think might be interesting to estimate. For any contrasts that you have selected, list the corresponding contrast coefficients.

3. **Meat cooking experiment, continued**

   The meat cooking experiment was described in Exercise 14 of Chap. 3, and the data were given in Table 3.14, p. 68.

   (a) Compare the effects of the six treatments, pairwise, using Scheffé's method of multiple comparisons and a 95% overall confidence level.
   (b) Consider $\mu + (\tau_1 + \tau_4)/2$, $\mu + (\tau_2 + \tau_5)/2$, and $\mu + (\tau_3 + \tau_6)/2$. What do these represent? Make pairwise comparisons of these three expressions, using Scheffé's method of multiple comparisons and a 95% overall confidence level for all treatment contrasts. Interpret the results.

4. **Reaction time experiment**

   (L. Cai, T. Li, Nishant, and A. van der Kouwe, 1996)

   The experiment was run to compare the effects of auditory and visual cues on speed of response of a human subject. A personal computer was used to present a "stimulus" to a subject, and the reaction time required for the subject to press a key was monitored. The subject was warned that the stimulus was forthcoming by means of an auditory or a visual cue. The experimenters were interested in the effects on the subjects' reaction time of the auditory and visual cues and also in different elapsed times between cue and stimulus. Thus, there were two different treatment factors: "cue stimulus" at two levels "auditory" or "visual," and "elapsed time between cue and stimulus" at three levels "five," "ten," or "fifteen" seconds. This gave a total of six treatment combinations, which can be coded as

$$
\begin{array}{ll}
1 = \text{auditory, 5 sec} & 4 = \text{visual, 5 sec} \\
2 = \text{auditory, 10 sec} & 5 = \text{visual, 10 sec} \\
3 = \text{auditory, 15 sec} & 6 = \text{visual, 15 sec}
\end{array}
$$

Table 4.4 Reaction times, in seconds, for the reaction time experiment—(order of collection in parentheses)

			Treatments			
1	2	3	4	5	6
0.204 (9)	0.167 (3)	0.202 (13)	0.257 (7)	0.283 (6)	0.256 (1)
0.170 (10)	0.182 (5)	0.198 (16)	0.279 (14)	0.235 (8)	0.281 (2)
0.181 (18)	0.187 (12)	0.236 (17)	0.269 (15)	0.260 (11)	0.258 (4)

The results of a pilot experiment, involving only one subject, are shown in Table 4.4. The reaction times were measured by the computer and are shown in seconds. The order of observation is shown in parentheses.

(a) Identify a set of contrasts that you would find particularly interesting in this experiment. (Hint: A comparison between the auditory treatments and the visual treatments might be of interest). These are your preplanned contrasts.
(b) Plot the data. What does the plot suggest about the treatments?
(c) Test the hypothesis that the treatments do not have different effects on the reaction time against the alternative hypothesis that they do have different effects.
(d) Calculate a set of simultaneous 90% confidence intervals for your preplanned contrasts, using a method or methods of your choice. State your conclusions.

5. **Trout experiment, continued**

Exercise 15 of Chap. 3 (p. 67) concerns a study of the effects of four levels of sulfamerazine (0, 5, 10, 15 g per 100 lb of fish) on the hemoglobin content of trout blood. An analysis of variance test rejected the hypothesis that the four treatment effects are the same at significance level $\alpha = 0.01$.

(a) Compare the four treatments using Tukey's method of pairwise comparisons and a 99% overall confidence level.
(b) Compare the effect of no sulfamerazine on the hemoglobin content of trout blood with the average effect of the other three levels. The overall confidence level of all intervals in parts (a) and (b) should be at least 98%.

6. **Battery experiment, continued**

In Example 4.4.3 (page 89), Tukey's method is used to obtain a set of 95% simultaneous confidence intervals for the pairwise differences $\tau_i - \tau_s$. Verify that this method gives shorter confidence intervals than would either of the Bonferroni or Scheffé methods (for $v = 4$ and $r = 4$).

7. **Soap experiment, continued**

The soap experiment was described in Sect. 2.5.1, p. 20, and an analysis was given in Sect. 3.7.2, p. 50.

(a) Suppose that the experimenter had been interested only in the contrast $\tau_1 - \frac{1}{2}(\tau_2 + \tau_3)$, which compares the weight loss for the regular soap with the average weight loss for the other two soaps. Calculate a confidence interval for this single contrast.

(b) Test the hypothesis that the regular soap has the same average weight loss as the average of the other two soaps. Do this via your confidence interval in part (a) and also via (4.3.13) and (4.3.15).

(c) In Example 4.4.5 (p. 91), Dunnett's method was used for simultaneous 99% confidence intervals for two preplanned treatment-versus-control contrasts. Would either or both of the Bonferroni and Tukey methods have given shorter intervals?

(d) Which method would be the best if all pairwise differences are required? Calculate a set of simultaneous 99% confidence intervals for all of the pairwise differences. Why are the intervals longer than those in part (c)?

8. **Trout experiment, continued**

(a) For the trout experiment in Exercise 15 of Chap. 3 (see p. 67), test the hypotheses that the linear and quadratic trends in hemoglobin content of trout blood due to the amount of sulfamerazine added to the diet is negligible. State the overall significance level of your tests.

(b) Regarding the absence of sulfamerazine in the diet as the control treatment, calculate simultaneous 99% confidence intervals for the three treatment-versus-control comparisons. Which method did you use and why?

(c) What is the overall confidence level of the intervals in part (b) together with those in Exercise 5? Is there a better strategy than using three different procedures for the three sets of intervals? Explain.

9. **Battery experiment, continued**

Suppose the battery experiment of Sect. 2.5.2 (p. 24) is to be repeated. The experiment involved four treatments, and the error standard deviation is estimated from that experiment to be about 48.66 minutes per dollar (minute/dollar).

(a) Calculate a 90% upper confidence limit for the error variance σ^2.

(b) How large should the sample sizes be in the new experiment if Tukey's method of pairwise comparisons is to be used and it is desired to obtain a set of 95% simultaneous confidence intervals of length at most 100 minutes per dollar?

(c) How large should the sample sizes be in the new experiment if Scheffé's method is to be used to obtain a set of 95% simultaneous confidence intervals for various contrasts and if the confidence interval for the duty contrast is to be of length at most 100 minute per dollar?

10. **Trout experiment, continued**

Consider again the trout experiment in Exercise 15 of Chap. 3.

(a) Suppose the experiment were to be repeated. Suggest the largest likely value for the error mean square *msE.*

(b) How many observations should be taken on each treatment so that the length of each interval in a set of simultaneous 95% confidence intervals for pairwise comparisons should be at most 2 g per 100 ml?

Checking Model Assumptions

5.1 Introduction

Throughout the two previous chapters, we discussed experiments whose data could be described by the one-way analysis of variance model (3.3.1), that is,

$$Y_{it} = \mu + \tau_i + \epsilon_{it},$$
$$\epsilon_{it} \sim N(0, \sigma^2),$$
$$\epsilon_{it}\text{'s are mutually independent},$$
$$t = 1, \ldots, r_i, \quad i = 1, \ldots, v.$$

This model implies that the response variables Y_{it} are mutually independent and have a normal distribution with mean $\mu + \tau_i$ and variance σ^2, that is, $Y_{it} \sim N(\mu + \tau_i, \sigma^2)$. For a given experiment, the model is selected in step (f) of the checklist using any available knowledge about the experimental situation, including the anticipated major sources of variation, the measurements to be made, the type of experimental design selected, and the results of any pilot experiment. However, it is not until the data have been collected that the adequacy of the model can be checked. Even if a pilot experiment has been used to help select the model, it is still important to check that the chosen model is a reasonable description of the data arising from the main experiment.

Methods of checking the model assumptions form the subject of this chapter, together with some indications of how to proceed if the assumptions are not valid. We begin by presenting a general strategy, including the order in which model assumptions should be checked. For checking model assumptions, we rely heavily on residual plots. We do so because while examination of residual plots is more subjective than would be testing for model lack-of-fit, the plots are often more informative about the nature of the problem, the consequences, and the corrective action.

5.2 Strategy for Checking Model Assumptions

In this section we discuss strategy and introduce the notions of residuals and residual plots. A good strategy for checking the assumptions about the model is to use the following sequence of checks.

- *Check the form of the model*—are the mean responses for the treatments adequately described by $E(Y_{it}) = \mu + \tau_i, i = 1, \ldots, v$?

- *Check for outliers*—are there any unusual observations (outliers)?
- *Check for independence*—do the error variables ϵ_{it} appear to be independent?
- *Check for constant variance*—do the error variables ϵ_{it} have similar variances for each treatment?
- *Check for normality*—do the error variables ϵ_{it} appear to be a random sample from a normal distribution?

For all of the fixed-effects models considered in this book, these same assumptions should be checked, except that $E(Y_{it})$ differs from model to model. The assumptions of independence, equal variance, and normality are the error assumptions mentioned in Chap. 3.

5.2.1 Residuals

The assumptions on the model involve the error variables, $\epsilon_{it} = Y_{it} - E(Y_{it})$, and can be checked by examination of the *residuals*. The it th residual $\hat{e}_{it}$ is defined as the observed value of $Y_{it} - \hat{Y}_{it}$, where $\hat{Y}_{it}$ is the least squares estimator of $E[Y_{it}]$, that is,

$$\hat{e}_{it} = y_{it} - \hat{y}_{it} \,.$$

For the one-way analysis of variance model (3.3.1), $E[Y_{it}] = \mu + \tau_i$, so the it th residual is

$$\hat{e}_{it} = y_{it} - (\hat{\mu} + \hat{\tau}_i) = y_{it} - \overline{y}_{i.} \,.$$

While one can simply use the residuals, we prefer to work with the *standardized residuals*, since standardization facilitates the identification of outliers. The standardization we use is achieved by dividing the residuals by their standard deviation, that is, by $\sqrt{ssE/(n-1)}$. The standardized residuals,

$$z_{it} = \frac{\hat{e}_{it}}{\sqrt{ssE/(n-1)}} \,,$$

then have sample variance equal to 1.0. Residuals standardized in this simplistic way are *scaled residuals*. Readers may prefer to use *Studentized residuals*, obtained by dividing each residual by its estimated standard error, either including or excluding the corresponding observation from the model fit. However, there is little distinction between these various approaches for analysis of variance models for data that is balanced or nearly so.

If the assumptions on the model are correct, the standardized error variables ϵ_{it}/σ are independently distributed with a $N(0, 1)$ distribution, so the observed values $e_{it}/\sigma = (y_{it} - (\mu + \tau_i))/\sigma$ would constitute independent observations from a standard normal distribution. Although the standardized residuals are dependent and involve estimates of both e_{it} and σ, their behavior should be similar. Consequently, methods for evaluating the model assumptions using the standardized residuals look for deviations from patterns that would be expected of independent observations from a standard normal distribution.

5.2.2 Residual Plots

A *residual plot* is a plot of the standardized residuals z_{it} against the levels of another variable, the choice of which depends on the assumption being checked. In Fig. 5.1, we show a plot of the standardized

Table 5.1 Data for the trout experiment

Code	Hemoglobin (grams per 100 ml)										$\overline{y}_{i.}$
1	6.7	7.8	5.5	8.4	7.0	7.8	8.6	7.4	5.8	7.0	7.20
2	9.9	8.4	10.4	9.3	10.7	11.9	7.1	6.4	8.6	10.6	9.33
3	10.4	8.1	10.6	8.7	10.7	9.1	8.8	8.1	7.8	8.0	9.03
4	9.3	9.3	7.2	7.8	9.3	10.2	8.7	8.6	9.3	7.2	8.69

Source: Gutsell (1951). Copyright © 1951 International Biometric Society. Reprinted with permission

residuals against the levels of the treatment factor for the trout experiment. Plots like this are useful for evaluating the assumption of constant error variance as well as the adequacy of the model.

Example 5.2.1 Constructing a residual plot: trout experiment

The trout experiment was described in Exercise 15 of Chap. 3. There was one treatment factor (grams of sulfamerazine per 100 lb of fish) with four levels coded 1, 2, 3, 4, each observed $r = 10$ times. The response variable was grams of hemoglobin per 100 ml of trout blood. The $n = 40$ data values are reproduced in Table 5.1 together with the treatment means.

Using the one-way analysis of variance model (3.3.1), it can be verified that $ssE = 56.471$. The residuals $\hat{e}_{it} = y_{it} - \overline{y}_{i.}$ and the standardized residuals $z_{it} = \hat{e}_{it}/\sqrt{ssE/(n-1)}$ are shown in Table 5.2. For example, the observation $y_{11} = 6.7$ yields the residual

$$\hat{e}_{11} = 6.7 - 7.2 = -0.5$$

and the standardized residual

$$z_{11} = -0.5/\sqrt{56.471/39} = -0.42$$

to two decimal places.

A plot of the standardized residuals against treatments is shown in Fig. 5.1. The residuals sum to zero for each treatment since $\Sigma_t(y_{it} - \overline{y}_{i.}) = 0$ for each $i = 1, \ldots, v$. The standardized residuals seem fairly well scattered around zero, although the spread of the residuals for treatment 2 seems a little larger than the spread for the other three treatments. This could be interpreted as a sign of unequal variances of the error variables or that the data values having standardized residuals 2.14 and −2.43

Fig. 5.1 Plot of standardized residuals for the trout experiment
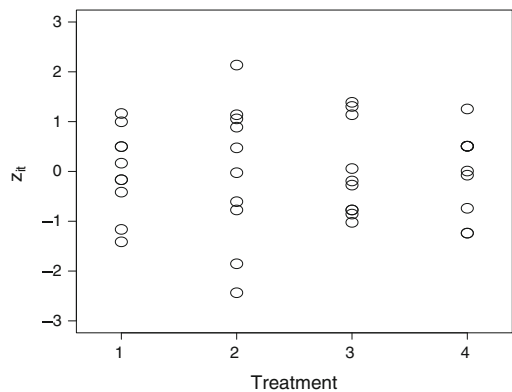
Table 5.2   Residuals and standardized residuals for the trout experiment

Treatment	Residuals				
1	−0.50	0.60	−1.70	1.20	−0.20
	0.60	1.40	0.20	−1.40	−0.20
2	0.57	−0.93	1.07	−0.03	1.37
	2.57	−2.23	−2.93	−0.73	1.27
3	1.37	−0.93	1.57	−0.33	1.67
	0.07	−0.23	−0.93	−1.23	−1.03
4	0.61	0.61	−1.49	−0.89	0.61
	1.51	0.01	−0.09	0.61	−1.49
Treatment	Standardized residuals				
1	−0.42	0.50	−1.41	1.00	−0.17
	0.50	1.16	0.17	−1.16	−0.17
2	0.47	−0.77	0.89	−0.02	1.14
	2.14	−1.85	−2.43	−0.61	1.06
3	1.14	−0.77	1.30	−0.27	1.39
	0.06	−0.19	−0.77	−1.02	−0.86
4	0.51	0.51	−1.24	−0.74	0.51
	1.25	0.01	−0.07	0.51	−1.24

are outliers, or it could be attributed to chance variation. Methods for checking for outliers and equality of variances will be discussed in Sects. 5.4 and 5.6, respectively.   □

5.3   Checking the Fit of the Model

The first assumption to be checked is the assumption that the model $E(Y_{it})$ for the mean response is correct. One purpose of running a pilot experiment is to choose a model that is a reasonable description of the data. If this is done, the model assumption checks for the main experiment should show no problems. If the model for mean response does not adequately fit the data, then there is said to be model *lack of fit*. If this occurs and if the model is changed accordingly, then any stated confidence levels and significance levels will only be approximate. This should be taken into account when decisions are to be made based on the results of the experiment.

In general, the fit of the model is checked by plotting the standardized residuals versus the levels of each independent variable (treatment factor, block factor, or covariate) included in the model. Lack of fit is indicated if the residuals exhibit a nonrandom pattern about zero in any such plot, being too often positive for some levels of the independent variable and too often negative for others.

For model (3.3.1), the only independent variable included in the model is the treatment factor. Since the residuals sum to zero for each level of the treatment factor, lack of fit would only be detected if there were a number of unusually large or small observations. However, lack of fit can also be detected by plotting the standardized residuals against the levels of factors that were omitted from the model. For example, for the trout experiment, if the standardized residuals were plotted against the age of the corresponding fish and if the plot were to show a pattern, then it would indicate that age should have been included in the model as a covariate. A similar idea is discussed in Sect. 5.5 with respect to checking for independence.

5.4   **Checking for Outliers**

An *outlier* is an observation that is much larger or much smaller than expected. This is indicated by a residual that has an unusually large positive or negative value. Outliers are fairly easy to detect from a plot of the standardized residuals versus the levels of the treatment factors. Any outlier should be investigated. Sometimes such investigation will reveal an error in recording the data, and this can be corrected. Otherwise, outliers may be due to the error variables not being normally distributed, or having different variances, or an incorrect specification of the model.

If all of the model assumptions hold, including normality, then approximately 68% of the standardized residuals should be between -1 and $+1$, approximately 95% between -2 and $+2$, and approximately 99.7% between -3 and $+3$. If there are more outliers than expected under normality, then the true confidence levels are lower than stated and the true significance levels are higher.

Example 5.4.1 Checking for outliers: battery experiment

In the battery experiment of Sect. 2.5.2 (p. 24), four observations on battery life per unit cost were collected for each of four battery types. Figure 5.2 shows the standardized residuals plotted versus battery type for the data as originally entered into the computer for analysis using model (3.3.1). This plot shows two related anomalies. There is one apparent outlier for battery type 2, the residual value being -2.98. Also, *all* of the standardized residuals for the other three battery types are less than one in magnitude. This is many more than the 68% expected.

An investigation of the outlier revealed a data entry error for the corresponding observation—a life length of 473 minutes was typed, but the recording sheet for the experiment showed the correct value to be 773 minutes. The unit cost for battery type 2 was $0.935 per battery, yielding the erroneous value of 506 minutes per dollar for the life per unit cost, rather than the correct value of 827. After correcting the error, the model was fitted again and the standardized residuals were replotted, as shown in Fig. 5.3.

Observe how correcting the single data entry error corrects both problems observed in Fig. 5.2. Not only is there no outlier, but the distribution of the 16 standardized residuals about zero is as one might anticipate for independent observations from a standard normal distribution—about a third of the standardized residuals exceed one in magnitude, and all are less than two in magnitude. The two anomalies are related, since correcting the data entry error makes *ssE* smaller and the standardized residuals correspondingly larger.   □

For an outlier like that shown in Fig. 5.2, the most probable cause of the problem is a measurement error, a recording error, or a transcribing error. When an outlier is detected, the experimenter should

Fig. 5.2 Original residual plot for the battery experiment
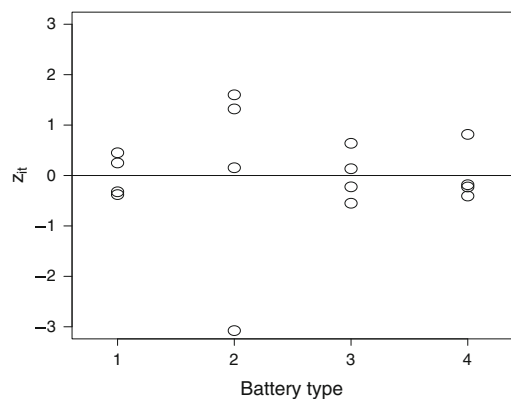
Fig. 5.3 Residual plot
after data correction for the
battery experiment
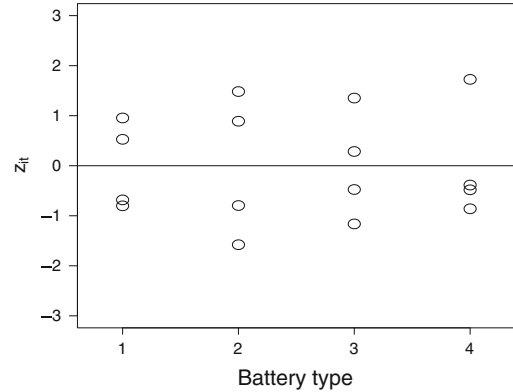
look at the original recording sheet to see whether the original data value has been copied incorrectly
at some stage. If the error can be found, then it can be corrected. When no obvious cause can be found
for an outlier, the data value should not automatically be discarded, since it may be an indication of an
occasional erratic behavior of a treatment. For example, had it not been due to a typographical error,
the outlier for battery type 2 in the previous example might have been due to a larger variability in the
responses for battery type 2.

The experimenter has to decide whether to include the unusual value in the analysis or whether
to omit it. First, the data should be reanalyzed without the outlying value. If the conclusions of the
experiment remain the same, then the outlier can safely be left in the analysis. If the conclusions change
dramatically, then the outlier is said to be *influential*, and the experimenter must make a judgment as to
whether the outlying observation is likely to be an experimental error or whether unusual observations
do occur from time to time. If the experimenter decides on the former, then the analysis should be
reported without the outlying observation. If the experimenter decides on the latter, then the model is
not adequate to describe the experimental situation, and a more complicated model would be needed.

5.5   Checking Independence of the Error Terms

Since the checks for the constant variance and normality assumptions assume that the error terms are
independent, a check for independence should be made next. The most likely cause of nonindepen-
dence in the error variables is the similarity of experimental units close together in time or space.
The independence assumption is checked by plotting the standardized residuals against the order in
which the corresponding observations were collected and against any spatial arrangement of the cor-
responding experimental units. If the independence assumption is satisfied, the residuals should be
randomly scattered around zero with no discernible pattern. Such is the case for Fig. 5.4 for the battery
experiment. If the plot were to exhibit a strong pattern, then this would indicate a serious violation of
the independence assumption, as illustrated in the following example.

Example 5.5.1 `Checking independence: balloon experiment`

The experimenter who ran the balloon experiment in Exercise 12 of Chap. 3 was concerned about
lack of independence of the observations. She had used a single subject to blow up all the balloons
in the experiment, and the subject had become an expert balloon blower before the experiment was
finished! Having fitted the one-way analysis of variance model (3.3.1) to the data (Table 3.13), she
plotted the standardized residuals against the time order in which the balloons were inflated. The plot

Fig. 5.4 Residual plot for the battery experiment
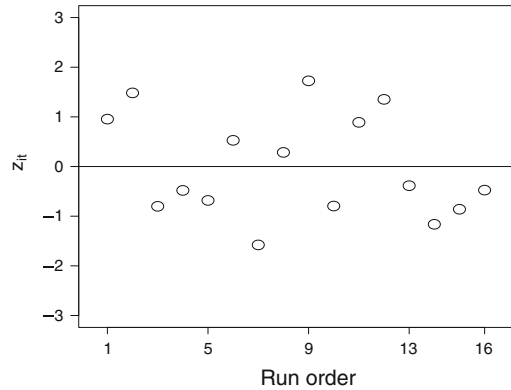
Fig. 5.5 Residual plot for the balloon experiment
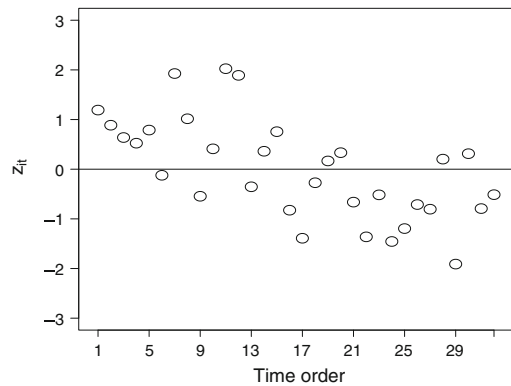
is shown in Fig. 5.5. There appears to be a strong downward drift in the residuals as time progresses. The observations are clearly *dependent*.   □

If an analysis is conducted under the assumptions of model (3.3.1) when, in fact, the error variables are dependent, the statistical conclusions may be distorted. For example, if errors corresponding to observations on the same treatment are positively correlated, but errors associated with different treatments are independently distributed, this artificially increases the power of tests, causing the true significance levels of tests under model (3.3.1) to be higher than stated, and causing the true confidence levels of confidence intervals to be lower than stated. Conversely, if groups of observations on different treatments (analogous to observations in the same block) have positively correlated errors, but errors associated with other pairs of observations (analogous to observations in different blocks) are independent, this tends to inflate the mean squared error and deflate test power, causing the true significance levels of tests under model (3.3.1) to be lower than stated, and causing the true confidence levels of confidence intervals to be higher than stated. The problem of dependent errors can be difficult to correct and a different model would need to be used (e.g. Chap. 17). If there is a clear trend in the residual plot, such as the linear trend in Fig. 5.5, it may be possible to add terms into the model to represent a time or space effect. For example, a more complex model that might be adequate for the balloon experiment is

$$Y_{it} = \mu + \tau_i + \gamma x_{it} + \epsilon_{it}$$
$$\epsilon_{it} \sim N(0, \sigma^2)$$
$$\epsilon_{it}{'}\text{s are mutually independent}$$
$$t = 1, 2, \ldots, r_i; \quad i = 1, \ldots, v,$$

where the variable x_{it} denotes the time at which the observation was taken and γ is a linear time trend parameter that must be estimated. Such a model is called an *analysis of covariance* model and will be studied in Chap. 9. The assumptions for analysis of covariance models are checked using the same types of plots as discussed in this chapter. In addition, the standardized residuals should also be plotted against the values of x_{it}.

Had the experimenter in the balloon experiment anticipated a run order effect, she could have selected an analysis of covariance model prior to the experiment. Alternatively, she could have grouped the observations into blocks of, say, eight observations. Notice that each group of eight residuals in Fig. 5.5 looks somewhat randomly scattered. As mentioned earlier in this chapter, when the model is changed after the data have been examined, then stated confidence levels and significance levels using that same data are inaccurate.

If a formal test of independence is desired, the most commonly used test is that of Durbin and Watson (1951) for time-series data (see Neter et al. 1996, pp. 504–510).

5.6   Checking the Equal Variance Assumption

If the independence assumption appears to be satisfied, then the equal-variance assumption should be checked. Studies have shown that if the sample sizes $r_1, \ldots, r_v$ are chosen to be equal, then unless one variance is considerably larger than the others, the significance level of hypothesis tests and confidence levels of the associated confidence intervals remain close to the stated values. However, if the sample sizes are unequal, and if the treatment factor levels which are more highly variable in response happen to have been observed fewer times (i.e. if smaller r_i coincide with larger $\mathrm{Var}(\epsilon_{it}) = \sigma_i^2$), then the statistical procedures are generally quite liberal, and the experimenter has a greater chance of making a Type I error in testing than anticipated, and also, the true confidence level of a confidence interval is lower than intended. On the other hand, if the large r_i coincide with large σ_i^2, then the procedures are conservative (significance levels are lower than stated and confidence levels are higher). Thus, unless there is good knowledge of which treatment factor levels are the more variable, an argument can be made that *the sample sizes should be chosen to be equal*.

5.6.1   Detection of Unequal Variances

The most common pattern of nonconstant variance is that in which the error variance increases as the mean response increases. This situation is suggested when the plot of the standardized residuals versus the fitted values resembles a megaphone in shape, as in Fig. 5.6. In such a case, one can generally find a transformation of the data, known as a variance-stabilizing transformation, which will correct the problem (see Sect. 5.6.2).

If the residual plot indicates unequal variances but not the pattern of Fig. 5.6 (or its mirror image), then a variance-stabilizing transformation is generally not available. Approximate and somewhat less powerful methods of data analysis such as those discussed in Sect. 5.6.3 must then be applied.

An unbiased estimate of the error variance σ_i^2 for the ith treatment is the sample variance of the residuals for the ith treatment, namely

Fig. 5.6 Megaphone-shaped residual plot
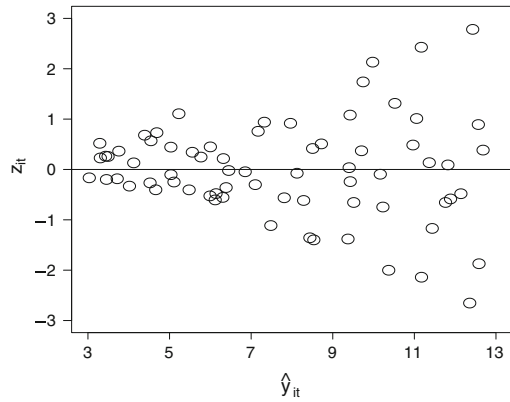
$$s_i^2 = \frac{1}{r_i - 1} \sum_{t=1}^{r_i} \hat{e}_{it}^2 = \frac{1}{r_i - 1} \sum_{t=1}^{r_i} (y_{it} - \hat{\mu} - \hat{\tau}_i)^2 \tag{5.6.1}$$

$$= \frac{1}{r_i - 1} \sum_{t=1}^{r_i} (y_{it} - \overline{y}_{i.})^2 \,.$$

There do exist tests for the equality of variances, but they tend to have low power unless there are large numbers of observations on each treatment factor level. Also, the tests tend to be very sensitive to nonnormality. (The interested reader is referred to Neter et al. 1996, p. 763).

A rule of thumb that we shall apply is that the usual analysis of variance F-test and the methods of multiple comparisons discussed in Chap. 4 are appropriate, provided that the ratio of the largest of the v treatment variance estimates to the smallest, $s_{\max}^2/s_{\min}^2$, does not exceed three. The rule of thumb is based on simulation studies suggesting that the methods of analysis are appropriate, provided that the largest ratio of actual variances, $\sigma_{\max}^2/\sigma_{\min}^2$, does not exceed three. Since the actual variances are unknown in practice, we are basing our rule of thumb on the estimates s_i^2 of the variances. Be aware, however, that *it is possible, and perhaps even likely, for the ratio of extreme variance estimates $s_{\max}^2/s_{\min}^2$ to exceed three, even when the model assumptions are correct, making the rule of thumb conservative*.

Example 5.6.1 `Comparing variances: trout experiment`

Figure 5.1 (p. 105) shows a plot of the standardized residuals against the levels of the treatment factor for the trout experiment. The plot suggests that the variance of the error variables for treatment 2 might be larger than the variances for the other treatments. Using the data in Table 3.15, we obtain

i	1	2	3	4
$\overline{y}_{i.}$	7.20	9.33	9.03	8.69
s_i^2	1.04	2.95	1.29	1.00

so $s_{\max}^2/s_{\min}^2 = 2.95$, which satisfies our rule of thumb, but only just. Both the standard analysis using model (3.3.1) and an approximate analysis that does not require equal variances will be discussed in Example 5.6.3.   □

5.6.2　Data Transformations to Equalize Variances

Finding a transformation of the data to equalize the variances of the error variables involves finding some function $h(y_{it})$ of the data so that the model

$$h(Y_{it}) = \mu^* + \tau_i^* + \epsilon_{it}^*$$

holds and $\epsilon_{it}^* \sim N(0, \sigma^2)$ and the ϵ_{it}^*'s are mutually independent for all $t = 1, \ldots, r_i$ and $i = 1, \ldots, v$. An appropriate transformation can generally be found if there is a clear relationship between the error variance $\sigma_i^2 = \mathrm{Var}(\epsilon_{it})$ and the mean response $E[Y_{it}] = \mu + \tau_i$, for $i = 1, \ldots, v$. If the variance and the mean increase together, as suggested by the megaphone-shaped residual plot in Fig. 5.6, or if one increases as the other decreases, then the relationship between σ_i^2 and $\mu + \tau_i$ is often of the form

$$\sigma_i^2 = k(\mu + \tau_i)^q \,, \tag{5.6.2}$$

where k and q are constants. In this case, the function $h(y_{it})$ should be chosen to be

$$h(y_{it}) = \begin{cases} (y_{it})^{1-(q/2)} & \text{if } q \neq 2, \\ \ln(y_{it}) & \text{if } q = 2 \text{ and all } y_{it}\text{'s are nonzero,} \\ \ln(y_{it} + 1) & \text{if } q = 2 \text{ and some } y_{it}\text{'s are zero.} \end{cases} \tag{5.6.3}$$

Here "ln" denotes the natural logarithm, which is the logarithm to the base e. Usually, the value of q is not known, but a reasonable approximation can be obtained empirically as follows. Substituting the least squares estimates for the parameters into Eq. (5.6.2) and taking logs of both sides gives

$$\ln(s_i^2) = \ln(k) + q(\ln(\overline{y}_{i.})) \,.$$

Therefore, the slope of the line obtained by plotting $\ln(s_i^2)$ against $\ln(\overline{y}_{i.})$ gives an estimate for q. This will be illustrated in Example 5.6.2.

The value of q is sometimes suggested by theoretical considerations. For example, if the normal distribution assumed in the model is actually an approximation to the Poisson distribution, then the variance would be equal to the mean, and $q = 1$. The square-root transformation $h(y_{it}) = (y_{it})^{1/2}$ would then be appropriate. The binomial distribution provides another commonly occurring case for which an appropriate transformation can be obtained theoretically. If each Y_{it} has a binomial distribution with mean mp and variance $mp(1 - p)$, then a variance-stabilizing transformation is

$$h(y_{it}) = \sin^{-1} \sqrt{y_{it}/m} = \arcsin\left(\sqrt{y_{it}/m}\right) \,.$$

When a transformation is found that equalizes the variances, then it is necessary to check or recheck the other model assumptions, since a transformation that cures one problem could cause others. If there are no problems with the other model assumptions, then analysis can proceed using the techniques of the previous two chapters, but using the transformed data $h(y_{it})$.

Example 5.6.2 Choosing a transformation: battery experiment

In Sect. 2.5.2, the response variable considered for the battery experiment was "battery life per unit cost," and a plot of the residuals versus the fitted values looks similar to Fig. 5.3 and shows fairly constant error variances.

Table 5.3   Life data for the battery experiment

Battery	Lifetime (minutes)				$\overline{y}_{i.}$	s_i^2
1	602	529	534	585	562.50	1333.71
2	863	743	773	840	804.75	3151.70
3	232	255	200	215	225.50	557.43
4	235	282	238	228	245.75	601.72

Fig. 5.7 Residual plot for the battery life data
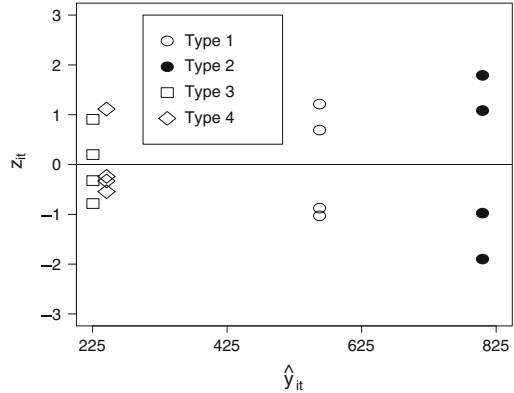
Suppose, however, that the response variable of interest had been "battery life" regardless of cost. The corresponding data are given in Table 5.3. The battery types are

1 = alkaline, name brand

2 = alkaline, store brand

3 = heavy duty, name brand

4 = heavy duty, store brand

Figure 5.7 shows a plot of the standardized residuals versus the fitted values. Variability seems to be increasing modestly with mean response, suggesting that a transformation can be found to stabilize the error variance. The ratio of extreme variance estimates is $s_{\max}^2/s_{\min}^2 = s_2^2/s_3^2 = 3151.70/557.43 \approx$ 5.65. Hence, based on the rule of thumb, a variance stabilizing transformation should be used. Using the treatment sample means and variances from Table 5.3, we have

i	$\overline{y}_{i.}$	$\ln(\overline{y}_{i.})$	s_i^2	$\ln(s_i^2)$
1	562.50	6.3324	1333.71	7.1957
2	804.75	6.6905	3151.70	8.0557
3	225.50	5.4183	557.43	6.3233
4	245.75	5.5043	601.72	6.3998

Figure 5.8 shows a plot of $\ln(s_i^2)$ against $\ln(\overline{y}_{i.})$. This plot is nearly linear, so the slope will provide an estimate of q in (5.6.2). A line can be drawn by eye or by the regression methods of Chap. 8. Both methods give a slope approximately equal to $q = 1.25$. From Eq. (5.6.3) a variance-stabilizing transformation is

$$h(y_{it}) = (y_{it})^{0.375} .$$

Fig. 5.8 Plot of $\ln(s_i^2)$ versus $\ln(\overline{y}_{i\cdot})$ for the battery life data
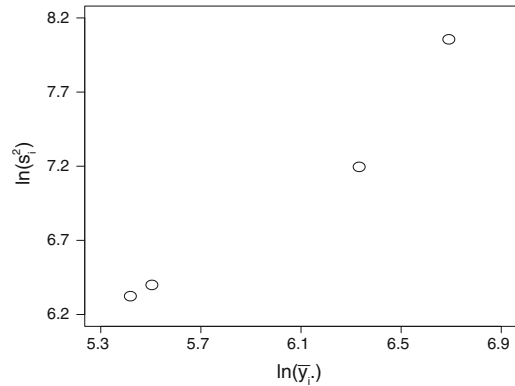
Table 5.4 Transformed life data $\sqrt{y_{it}}$ for the battery experiment

Brand		$x_{it} = h(y_{it}) = \sqrt{y_{it}}$			$\overline{x}_{i\cdot}$	s_i^2
1	24.536	23.000	23.108	24.187	23.708	0.592
2	29.377	27.258	27.803	28.983	28.355	0.982
3	15.232	15.969	14.142	14.663	15.001	0.614
4	15.330	16.793	15.427	15.100	15.662	0.587

Since $(y_{it})^{0.375}$ is close to $(y_{it})^{0.5}$, and since the square root of the data values is perhaps more meaningful than $(y_{it})^{0.375}$, we will try taking the square root transformation. The square roots of the data are shown in Table 5.4.

The transformation has stabilized the variances considerably, as evidenced by $s_{\max}^2/s_{\min}^2 = 0.982/0.587 \approx 1.67$. Checks of the other model assumptions for the transformed data also reveal no severe problems. The analysis can now proceed using the transformed data. The stated significance level and confidence levels will now be approximate, since the model has been changed based on the data. For the transformed data, $msE = 0.6936$. Using Tukey's method of multiple comparisons to compare the lives of the four battery types (regardless of cost) at an overall confidence level of 99%, the minimum significant difference obtained from Eq. (4.4.28) is

$$msd = q_{4,12,0.01}\sqrt{msE/4} = 5.50\sqrt{0.6936/4} = 2.29 \,.$$

Comparing *msd* with the differences in the sample means $\overline{x}_{i\cdot}$ of the transformed data in Table 5.4, we can conclude that at an overall 99% level of confidence, all pairwise differences are significantly different from zero except for the comparison of battery types 3 and 4. Furthermore, it is reasonable to conclude that type 2 (alkaline, store brand) is best, followed by type 1 (alkaline, name brand). However, any more detailed interpretation of the results is muddled by use of the transformation, since the comparisons use mean values of $\sqrt{\text{life}}$. A more natural transformation, which also provided approximately equal error variances, was used in Sect. 2.5.2. There, the response variable was taken to be "life per unit cost," and confidence intervals were able to be calculated in meaningful units. □

5.6.3   Analysis with Unequal Error Variances

An alternative to transforming the data to equalize the error variances is to use a method of data analysis that is designed for nonconstant variances. Such a method will be presented for constructing confidence intervals. The method is approximate and tends to be less powerful than the methods of Chap. 4 with transformed data. However, the original data units are maintained, and the analysis can be used whether or not a variance-stabilizing transformation is available.

Without the assumption of equal variances for all treatments, the one-way analysis of variance model (3.3.1) is

$$
\begin{aligned}
Y_{it} &= \mu + \tau_i + \epsilon_{it}\,, \\
\epsilon_{it} &\sim N(0, \sigma_i^2)\,, \\
&\epsilon_{it}\text{'s are mutually independent}\,, \\
&t = 1, \ldots, r_i, \quad i = 1, \ldots, v.
\end{aligned}
$$

For this model, each contrast $\Sigma c_i \tau_i$ in the treatment parameters remains estimable, but the least squares estimator $\Sigma c_i \hat{\tau}_i = \Sigma c_i \overline{Y}_{i.}$ now has variance $\mathrm{Var}(\Sigma c_i \overline{Y}_{i.}) = \Sigma c_i^2 \sigma_i^2 / r_i$. If we estimate σ_i^2 by s_i^2 as given in (5.6.1), then

$$
\frac{\Sigma c_i \hat{\tau}_i - \Sigma c_i \tau_i}{\sqrt{\widehat{\mathrm{Var}}(\Sigma c_i \hat{\tau}_i)}}
$$

has approximately a t-distribution with df degrees of freedom, where

$$
\widehat{\mathrm{Var}}(\Sigma c_i \hat{\tau}_i) = \sum \frac{c_i^2}{r_i} s_i^2 \quad \text{and} \quad df = \frac{(\Sigma c_i^2 s_i^2 / r_i)^2}{\sum \frac{(c_i^2 s_i^2 / r_i)^2}{(r_i - 1)}}\,. \tag{5.6.4}
$$

Then an approximate $100(1 - \alpha)\%$ confidence interval for a single treatment contrast $\Sigma c_i \tau_i$ is

$$
\sum c_i \tau_i \in \left(\sum c_i \hat{\tau}_i \pm w \sqrt{\widehat{\mathrm{Var}}(\Sigma c_i \hat{\tau}_i)} \right), \tag{5.6.5}
$$

where $w = t_{df, \alpha/2}$ and $\Sigma c_i \hat{\tau}_i = \Sigma c_i \overline{y}_{i.}$, all sums being from $i = 1$ to $i = v$. The formulae for $\widehat{\mathrm{Var}}(\Sigma c_i \hat{\tau}_i)$ and df in (5.6.4), often called *Satterthwaite's approximation*, are due to Smith (1936), Welch (1938), and Satterthwaite (1946). The approximation is best known for use in inferences on a pairwise comparison $\tau_h - \tau_i$ of the effects of two treatments, in which case, for samples each of size r, (5.6.4) reduces to

$$
\widehat{\mathrm{Var}}(\hat{\tau}_h - \hat{\tau}_i) = \frac{s_h^2}{r} + \frac{s_i^2}{r} \quad \text{and} \quad df = \frac{(r - 1)(s_h^2 + s_i^2)^2}{s_h^4 + s_i^4}\,. \tag{5.6.6}
$$

Satterthwaite's approach can be extended to multiple comparison procedures by changing the critical coefficient w appropriately and computing $\Sigma c_i \hat{\tau}_i$ and df separately for each contrast. For Tukey's method, for example, the critical coefficient in (5.6.5) is $w_T = q_{v, df, \alpha} / \sqrt{2}$; this variation on Tukey's method is the *Games–Howell method* due to Games and Howell (1976). Simulation studies by Dunnett (1980) have shown this Games–Howell method to maintain approximately the specified error rate, though in a few circumstances it can be modestly liberal (true α slightly larger than the stated value).

Fig. 5.9 Residual plot for the trout experiment
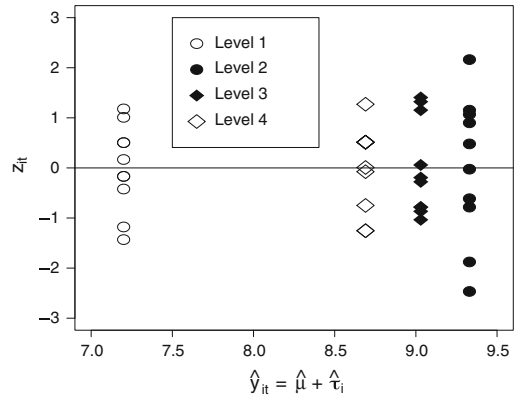
Example 5.6.3 `Satterthwaite's approximation: trout experiment`

In Example 5.6.1, it was shown that the ratio of the maximum to the minimum error variance for the trout experiment satisfies the rule of thumb, but only just. The standardized residuals are plotted against the fitted values in Fig. 5.9. The data for treatment 2 are the most variable and have the highest mean response, but there is no clear pattern of variability increasing as the mean response increases. In fact, it can be verified that a plot of $\ln(s_i^2)$ against $\ln(\overline{y}_{i.})$ is not very close to linear, suggesting that a transformation will not be successful in stabilizing the variances.

To obtain simultaneous approximate 95% confidence intervals for pairwise comparisons in the treatment effects by Tukey's method using Satterthwaite's approximation, we use Eqs. (5.6.5) and (5.6.6) with $r = 10$. The minimum significant difference for pairwise comparison $\tau_h - \tau_i$ is

$$ msd = \frac{1}{\sqrt{2}} q_{4,df,0.05} \sqrt{\frac{s_h^2}{r} + \frac{s_i^2}{r}} \, , $$

the size of which depends upon which pair of treatments is being compared. From Example 5.6.1, we have

$$ s_1^2 = 1.04, \;\; s_2^2 = 2.95, \;\; s_3^2 = 1.29, \;\; s_4^2 = 1.00 \, . $$

The values of $\sqrt{\widehat{\mathrm{Var}}(\hat{\tau}_h - \hat{\tau}_i)} = \sqrt{s_h^2/r + s_i^2/r}$ are listed in Table 5.5. Comparing the values of *msd* with the values of $\overline{y}_{h.} - \overline{y}_{i.}$ in Table 5.5, we can conclude with simultaneous approximate 95% confidence

Table 5.5 Approximate values for Tukey's multiple comparisons for the trout experiment

(h, i)	$\sqrt{s_h^2/r + s_i^2/r}$	df	$q_{4,df,0.05}$	msd	$\overline{y}_{h.} - \overline{y}_{i.}$
(2, 3)	0.651	$15.6 \approx 16$	4.05	1.86	0.30
(2, 4)	0.629	$14.5 \approx 15$	4.08	1.82	0.64
(2, 1)	0.631	$14.6 \approx 15$	4.08	1.82	2.13
(3, 4)	0.478	$17.7 \approx 18$	4.00	1.35	0.34
(3, 1)	0.483	$17.8 \approx 18$	4.00	1.37	1.83
(4, 1)	0.452	$18.0 \approx 18$	4.00	1.28	1.49

that each of treatments 2, 3, and 4 yields statistically significantly higher mean response than does treatment 1.

Since $s^2_{max}/s^2_{min} = 2.95$, we could accept the rule of thumb and apply Tukey's method (4.4.28) for equal variances. The minimum significant difference for each pairwise comparison would then be

$$msd = q_{4,36,0.05}\sqrt{msE/10} = 3.82\sqrt{1.5685/10} \approx 1.51 \,.$$

Comparing this with the values of $\overline{y}_{h.} - \overline{y}_{i.}$ in Table 5.5, the same conclusion is obtained as in the analysis using Satterthwaite's approximation, namely, treatment 1 has significantly lower mean response than do treatments 2, 3, and 4. The three confidence intervals involving treatment 2, having length $2(msd)$, would be slightly wider using Satterthwaite's approximation, and the other three confidence intervals would be slightly narrower. Where there is so little difference in the two methods of analysis, the standard analysis would usually be preferred.   □

5.7   Checking the Normality Assumption

The assumption that the error variables have a normal distribution is checked using a *normal probability plot*, which is a plot of the standardized residuals against their normal scores. *Normal scores* are percentiles of the standard normal distribution, and we will show how to obtain them after providing motivation for the normal probability plot.

If a given linear model is a reasonable description of a set of data without any outliers, and if the error assumptions are satisfied, then the standardized residuals would look similar to n independent observations from the standard normal distribution. In particular, the qth smallest standardized residual would be approximately equal to the $100[q/(n+1)]$th percentile of the standard normal distribution. Consequently, when the model assumptions hold, a plot of the qth smallest standardized residual against the $100[q/(n+1)]$th percentile of the standard normal distribution for each $q = 1, 2, \ldots, n$ would show points roughly on a straight line through the origin with slope equal to 1.0. However, if any of the model assumptions fail, and in particular if the normality assumption fails, then the normal probability plot shows a nonlinear pattern.

Blom, in 1958, recommended that the standardized residuals be plotted against the $100[(q - 0.375)/(n + 0.25)]$th percentiles of the standard normal distribution rather than the $100[q/(n+1)]$th percentiles, since this gives a slightly straighter line. These percentiles are called *Blom's normal scores*.

Blom's qth normal score is the value ξ_q for which

$$P(Z \le \xi_q) = (q - 0.375)/(n + 0.25),$$

where Z is a standard normal random variable. Hence, Blom's qth normal score is

$$\xi_q = \Phi^{-1}[(q - 0.375)/(n + 0.25)] \,, \tag{5.7.7}$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution. The normal scores possess a symmetry about zero, that is, the jth smallest and the jth largest scores are always equal in magnitude but opposite in sign.

The normal scores are easily obtained and normal probability plots are easily generated using most statistical packages, as illustrated in Sects. 5.8 and 5.9 for SAS and R software, respectively. Alternatively, the normal scores can be calculated as shown in Example 5.7.1 using Table A.3 for the standard normal distribution.

Table 5.6 Normal scores: battery experiment

z_{it}	ξ_q	$\sqrt{y_{it}}$	Battery
-1.47	-1.77	27.258	2
-1.15	-1.28	14.142	3
-0.95	-0.99	23.000	1
-0.80	-0.76	23.108	1
-0.76	-0.57	15.100	4
-0.74	-0.40	27.803	2
-0.45	-0.23	14.663	3
-0.45	-0.08	15.330	4
-0.32	0.08	15.427	4
0.31	0.23	15.232	3
0.64	0.40	24.187	1
0.84	0.57	28.983	2
1.11	0.76	24.536	1
1.30	0.99	15.969	3
1.37	1.28	29.377	2
1.52	1.77	16.793	4

Example 5.7.1 Computing normal scores: battery experiment

To illustrate the normal probability plot and the computation of normal scores, consider the battery life data (regardless of cost) that were transformed in Example 5.6.2 to equalize the variances. The transformed observations, standardized residuals, and normal scores are listed in Table 5.6, in order of increasing size of the residuals. In the battery experiment there were $n = 16$ observations in total. The first normal score that corresponds to the smallest residual ($q = 1$) is

$$\xi_1 = \Phi^{-1}[(1 - 0.375)/(16 + 0.25)] = \Phi^{-1}(0.0385)\,.$$

Thus, the area under the standard normal curve to the left of ξ_1 is 0.0385. Using a table for the standard normal distribution or a computer program, this value is

$$\Phi^{-1}(0.0385) = -1.77\,.$$

By symmetry, the largest normal score is 1.77. The other normal scores are calculated in a similar fashion, and the corresponding normal probability plot is shown in Fig. 5.10. We discuss the interpretation of this plot below.   □

For inferences concerning treatment means and contrasts, the assumption of normality needs only to be approximately satisfied. Interpretation of a normal probability plot, such as that in Fig. 5.10, requires some basis of comparison. The plot is not completely linear. Such plots always exhibit some sampling variation even if the normality assumption is satisfied. Since it is difficult to judge a straight line for small samples, normal probability plots are useful only if there are at least 15 standardized residuals being plotted. A plot for 50 standardized residuals that are known to have a normal distribution is shown in plot (a) of Fig. 5.11 and can be used as a benchmark of what might be expected when the assumption of normality is satisfied.

Small deviations from normality do not badly affect the stated significance levels, confidence levels, or power. If the sample sizes are equal, the main case for concern is that in which the distribution has

Fig. 5.10 Normal probability plot for the square root battery data
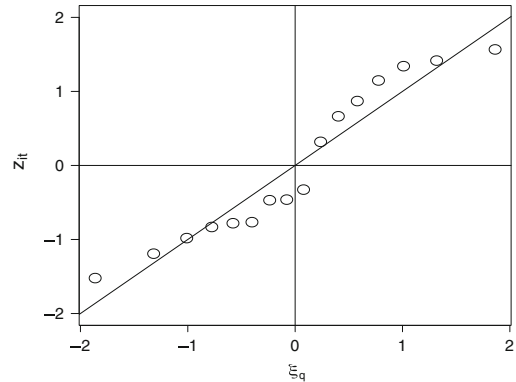
Fig. 5.11 Normal probability plots for two distributions

(a) Normal distribution   (b) Distribution with heavy tails
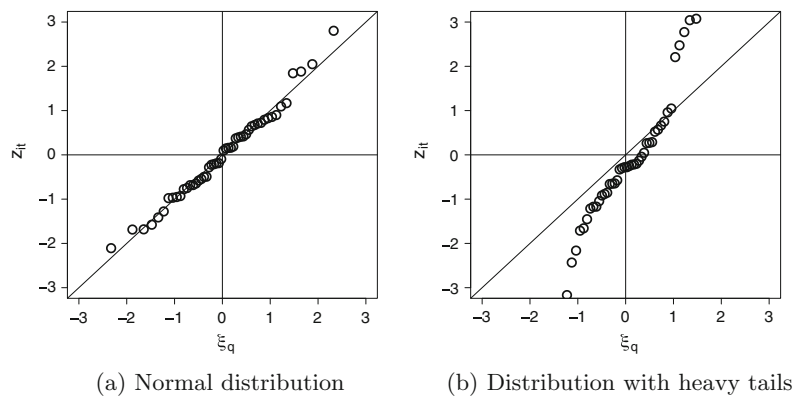
heavier tails than the normal distribution, as in plot (b) of Fig. 5.11. The apparent outliers are caused by the long tails of the nonnormal distribution, and a model based on normality would not be adequate to represent such a set of data. If this is the case, then use of nonparametric methods of analysis should be considered (as described, for example, by Hollander and Wolfe 2013). Sometimes, a problem of nonnormality can be cured by taking a transformation of the data, such as $\ln(y_{it})$. However, it should be remembered that any transformation could cause a problem of unequal variances where none existed before. If the equal variance assumption does not hold for a given set of data, then a separate normal probability plot should be generated for each treatment instead of one plot using all n residuals (provided that there are sufficient data values).

The plot for the transformed battery life data shown in Fig. 5.10 is less linear than the benchmark plot, but it does not exhibit the extreme behavior of plot (b) of Fig. 5.11 for the heavy-tailed nonnormal distribution. Consequently, the normality assumption can be taken to be approximately satisfied, and the stated confidence and significance levels will be approximately correct.

5.8   Using SAS Software

5.8.1   Residual Plots

We now illustrate use of the SAS software to generate the various plots used in this chapter. In the following sections, we will check the assumptions on the one-way analysis of variance model (3.3.1) for the data of the mung bean experiment described in Example 5.8.1 below.

Table 5.7 Data for the mung bean experiment

Treatment	Shoot length in mm (Order of observation in parentheses)			
1	1.5 (14)	1.1 (15)	1.3 (18)	0.9 (30)
	8.5 (35)	10.6 (39)	3.5 (42)	7.4 (43)
2	0.0 (3)	0.6 (4)	9.5 (7)	11.3 (12)
	12.6 (17)	8.1 (27)	7.8 (29)	7.3 (37)
3	5.2 (16)	0.4 (23)	3.6 (31)	2.8 (36)
	12.3 (45)	14.1 (46)	0.3 (47)	1.8 (48)
4	13.2 (1)	14.8 (11)	10.7 (13)	13.8 (20)
	9.6 (24)	0.0 (34)	0.6 (40)	8.2 (44)
5	5.1 (5)	3.3 (21)	0.2 (26)	3.9 (28)
	7.0 (32)	9.5 (33)	11.1 (38)	6.2 (41)
6	11.6 (2)	2.3 (6)	6.7 (8)	2.5 (9)
	10.6 (10)	10.8 (19)	15.9 (22)	9.0 (25)

Example 5.8.1 `Mung bean experiment`

An experiment was run in 1993 by K.H. Chen, Y.F. Kuo, R. Sengupta, J. Xu, and L.L. Yu to compare watering schedules and growing mediums for mung bean seeds. There were two treatment factors: "amount of water" with three levels (1, 2, and 3 teaspoons of water per day) and "growing medium" having two levels (tissue and paper towel, coded 1 and 2). We will recode the six treatment combinations as $1 = 11, 2 = 12, 3 = 21, 4 = 22, 5 = 31, 6 = 32$.

Forty-eight beans of approximately equal weights were randomly selected for the experiment. These were all soaked in water in a single container for two hours. After this time, the beans were placed in separate containers and randomly assigned to a treatment (water/medium) combination in such a way that eight containers were assigned to each treatment combination. The 48 containers were placed on a table in a random order. The shoot lengths of the beans were measured (in mm) after one week. The data are shown in Table 5.7 together with the order in which they were collected.   □

A SAS program that generates the residual plots for the mung bean experiment is shown in Table 5.8. The program uses the SAS procedures `GLM`, `PRINT`, and `SGPLOT`, all of which were introduced in Sect. 3.8.

The values of the factors `ORDER` (order of observation), `WATER`, `MEDIUM`, and the response variable `LENGTH` are entered into the data set `MUNGBEAN` using the `INPUT` statement. The treatment combinations are then recoded, with the levels of `TRTMT` representing the recoded levels 1–6.

The `OUTPUT` statement in the `GLM` procedure calculates and saves the predicted values $\widehat{y}_{it}$ as the variable `YPRED` and two copies of the residuals $\hat{e}_{it}$ as the variables `E` and `Z` in a new data set named `MUNGBN2`. The data set `MUNGBN2` also contains all of the variables in the original data set `MUNGBEAN`. The residuals stored as the variable `Z` are then standardized using the procedure `STANDARD` by dividing each residual by $\sqrt{ssE/(n-1)}$. This is done by requesting the procedure `STANDARD` to achieve a standard deviation of 1.0. The variables `E` and `Z` then represent the residuals and standardized residuals, respectively.

The procedure `RANK` is used to compute Blom's normal scores. The procedure orders the standardized residuals from smallest to largest and calculates their ranks. (The qth smallest residual has rank q.) The values of the variable `NSCORE` calculated by this procedure are the normal scores for the values of `Z`. The `PRINT` procedure prints all the values of the variables created so far. Some representative output is shown in Fig. 5.12. The `PRINT` statement can be omitted if this information is not wanted.

Table 5.8 SAS program to generate residual plots: mung bean experiment

```
DATA MUNGBEAN;
  INPUT ORDER WATER MEDIUM LENGTH;
  TRTMT = 2*(WATER-1) + MEDIUM;
  LINES;
   1   2   2   13.2
   2   3   2   11.6
   3   1   2    0.0
   :   :   :    :
  48   2   1    1.8
;
PROC GLM;
  CLASS TRTMT;
  MODEL LENGTH = TRTMT;
  OUTPUT OUT=MUNGBN2 PREDICTED=YPRED RESIDUAL=E RESIDUAL=Z;
;
PROC STANDARD STD=1.0; VAR Z;
PROC RANK NORMAL=BLOM; VAR Z; RANKS NSCORE;
PROC PRINT;
;
* Plotting standardized residuals versus run order;
PROC SGPLOT;
  SCATTER X=ORDER Y=Z;
  XAXIS LABEL = 'Order';
  YAXIS LABEL = 'Standardized Residuals';
  REFLINE 0 / AXIS=Y;
;
* Plotting standardized residuals versus normal scores;
PROC SGPLOT;
  SCATTER X=NSCORE Y=Z;
  XAXIS VALUES = (-4 to 4 by 2) LABEL = 'Normal Scores';
  YAXIS LABEL = 'Standardized Residuals';
  REFLINE 0 / AXIS=Y;
  REFLINE 0 / AXIS=X;
```

Fig. 5.12 Output from
`PROC PRINT`

The SAS System

Obs	ORDER	WATER	MEDIUM	LENGTH	TRTMT	YPRED	E	Z	NSCORE
1	1	2	2	13.2	4	8.8625	4.3375	0.98205	0.92011
2	2	3	2	11.6	6	8.6750	2.9250	0.66224	0.57578
3	3	1	2	0.0	2	7.1500	-7.1500	-1.61882	-1.60357
4	4	1	2	0.6	2	7.1500	-6.5500	-1.48297	-1.43862
5	5	3	1	5.1	5	5.7875	-0.6875	-0.15566	-0.18284

Fig. 5.13 Plot of z_{it} versus run order: mung bean experiment
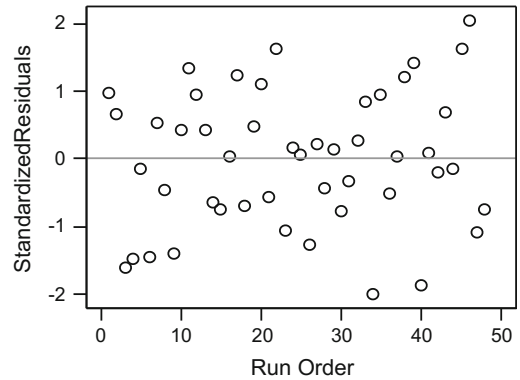
Fig. 5.14 Plot of z_{it} versus normal score: mung bean experiment
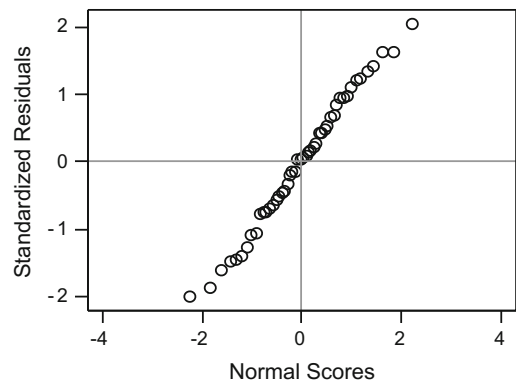
Plots of the standardized residuals z_{it} against treatments, predicted values, run order, and normal scores may be of interest. For illustration, the last two of these are requested using the SGPLOT procedure. Vertical and horizontal reference lines at zero may be included as appropriate via the REFLINE statements.

For the mung bean experiment, a plot of the standardized residuals against the order in which the observations are collected is shown in Fig. 5.13, and a plot of standardized residuals against normal scores is shown in Fig. 5.14. Neither of these plots indicates any serious problems with the assumptions on the model.

A plot of the standardized residuals against the predicted values (not shown) suggests that treatment variances are not too unequal, but that there could be outliers associated with one or two of the treatments. The first nine lines of the SAS program in Table 5.9, through the first PRINT procedure, produced the first four columns of output of Fig. 5.15. From this, the rule of thumb can be checked that the sample variances should not differ by more than a factor of 3. It can be verified that the ratio of the maximum and minimum variances is under 2.7 for this experiment.
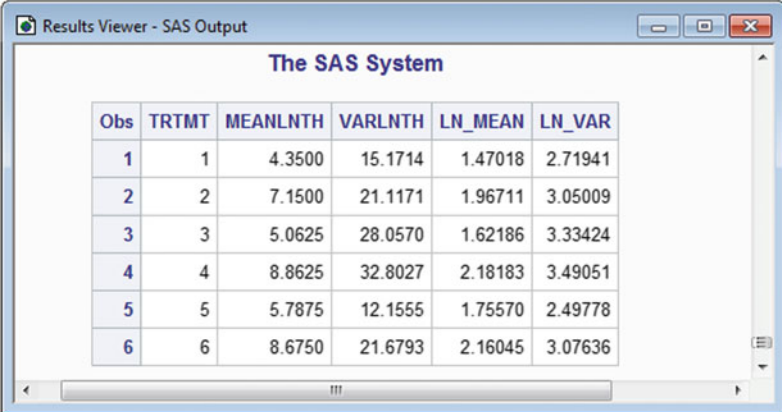
When the equal-variance assumption does not appear to be valid, the experimenter may choose to use an analysis based on Satterthwaite's approximation (see Sect. 5.8.3), using formulas involving the treatment sample variances such as those in Fig. 5.15. A normal probability plot such as that of Fig. 5.14 would not be relevant; rather, the normality assumption needs to be checked for each treatment separately. This can be done by generating a separate normal probability plot for each treatment (provided that the sample sizes are sufficiently large). To obtain the plots, first obtain the normal scores separately for each treatment by including a BY TRTMT statement in the SORT and

Table 5.9 SAS program to plot $\ln(s_i^2)$ against $\ln(\bar{y}_{i.})$: mung bean experiment

```
DATA MUNGBEAN; SET MUNGBEAN;
PROC SORT; BY TRTMT;
PROC MEANS NOPRINT MEAN VAR; BY TRTMT;
  VAR LENGTH;
  OUTPUT OUT=MUNGBN3 MEAN=MEANLNTH VAR=VARLNTH;
PROC PRINT;
  VAR TRTMT MEANLNTH VARLNTH;
DATA MUNGBN3; SET MUNGBN3;
  LN_MEAN=LOG(MEANLNTH); LN_VAR=LOG(VARLNTH);
PROC PRINT;
  VAR TRTMT MEANLNTH VARLNTH LN_MEAN LN_VAR;
PROC SGPLOT;
  SCATTER X = LN_MEAN Y = LN_VAR;
  XAXIS VALUES = (1.4 to 2.2 by .2) LABEL = 'ln(mean)';
  YAXIS VALUES = (2.5 to 3.5 by .2) LABEL = 'ln(var)';
```

Fig. 5.15 Treatment sample means and variances: mung bean experiment

Obs	TRTMT	MEANLNTH	VARLNTH	LN_MEAN	LN_VAR
1	1	4.3500	15.1714	1.47018	2.71941
2	2	7.1500	21.1171	1.96711	3.05009
3	3	5.0625	28.0570	1.62186	3.33424
4	4	8.8625	32.8027	2.18183	3.49051
5	5	5.7875	12.1555	1.75570	2.49778
6	6	8.6750	21.6793	2.16045	3.07636

RANK procedures. Then, instead of SGPLOT, use the SGPANEL procedure and PANELBY TRTMT to produce a panel of plots—one for each treatment. Sample program lines are as follows.

```
PROC SORT; BY TRTMT;
PROC RANK NORMAL=BLOM; BY TRTMT;
  VAR Z; RANKS NSCORE;
PROC SGPANEL; PANELBY TRTMT;
  SCATTER X=NSCORE Y=Z;
```

5.8.2   Transforming the Data

If a variance-stabilizing transformation is needed, a plot of $\ln(s_i^2)$ against $\ln(\bar{y}_{i.})$ can be achieved via the program in Table 5.9 (shown for the mung bean experiment). These statements can be added to those in Table 5.8 either before the GLM procedure or at the end of the program.

The SORT procedure and the BY statement sort the observations in the original data set MUNGBEAN using the values of the variable TRTMT. This is required by the subsequent MEANS procedure with the NOPRINT option, which computes the mean and variance of the variable LENGTH separately for each

Fig. 5.16   Plot of $\ln(s_i^2)$ against $\ln(\bar{y}_i.)$: mung bean experiment
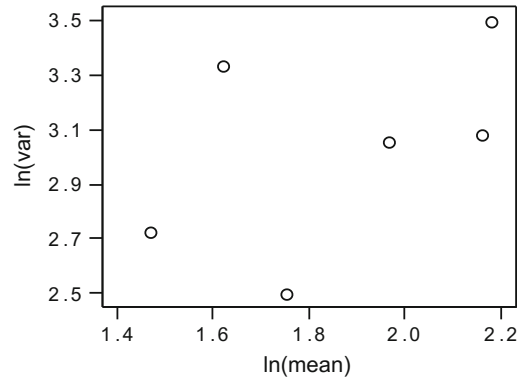
treatment, without printing the results. The OUTPUT statement creates a data set named MUNGBN3, with one observation for each treatment, and with the two variables MEANLNTH and VARLNTH containing the sample mean lengths and sample variances for each treatment. Two new variables LN_MEAN and LN_VAR are created.

These are the natural logarithm, or log base e, of the sample mean and variance of length for each treatment. The PRINT procedure prints the values of the variables TRTMT, MEANLNTH, VARLNTH, LN_MEAN, LN_VAR. The output is in Fig. 5.15.

Finally, the SGPLOT procedure generates the plot of $\ln(s_i^2)$ against $\ln(\bar{y}_i.)$, shown in Fig. 5.16. The values do not fall along a straight line, so a variance-stabilizing transformation of the type given in Eq. (5.6.3) does not exist for this data set. However, since the ratio of the maximum to the minimum variance is less than 3.0, a transformation is not vital, according to our rule of thumb.

If an appropriate transformation is identified, then the transformed variable can be created from the untransformed variable in a DATA step of a SAS program, just as the variables LN_MEAN and LN_VAR were created in the data set MUNGBN3 by transforming the variables MEANLNTH and VARLNTH, respectively. Alternatively, the transformation can be achieved after the INPUT statement in the same way as the factor TRTMT was created. SAS statements useful for the variance-stabilizing transformations of Eq. (5.6.3) include:

Transformation	SAS Statement
$h = \ln(y)$	H = LOG(Y);
$h = \sin^{-1}(y)$	H = ARSIN(Y);
$h = y^p$	H = Y**P;

5.8.3   Implementing Satterthwaite's Method

In Example 5.6.3, given indications of unequal variances in the trout experiment, simultaneous approximate 95% confidence intervals for pairwise comparisons were computed using the Games–Howell method—namely, using Satterthwaite's approximation in conjunction with Tukey's method. This method can be implemented in SAS software using PROC MIXED—a procedure that will be introduced in greater detail in later chapters. Appropriate statements are given in Table 5.10. The REPEATED statement relaxes the model assumption of equal variances, allowing for separate variance estimates s_i^2 at each level of sulfa. Correspondingly, the model is fit by restricted maximum likelihood estimation rather than ordinary least squares (see Chap. 19 for more on restricted maximum likelihood estimation). The collective options in the MODEL and LSMEANS statements implement Tukey's method,

Table 5.10 SAS program for multiple comparisons with unequal variances: trout experiment

```
DATA TROUT;
  INPUT SULFA HEMO;
  LINES;
  1  6.7
  1  7.8
  1  5.5
  :   :
  4  7.2
;
PROC MIXED;
  CLASS SULFA;
  MODEL HEMO = SULFA / DDFM=SATTERTH;
  REPEATED / GROUP=SULFA;
  LSMEANS SULFA / ADJDFE=ROW ADJUST=TUKEY;
```


The SAS System

Obs	Effect	SULFA	_SULFA	Estimate	StdErr	DF	Adjustment	Adjp	Alpha	AdjLower	AdjUpper
1	SULFA	1	2	-2.1300	0.6312	14.6	Tukey-Kramer	0.0199	0.05	-3.9545	-0.3055
2	SULFA	1	3	-1.8300	0.4824	17.8	Tukey-Kramer	0.0067	0.05	-3.1949	-0.4651
3	SULFA	1	4	-1.4900	0.4515	18	Tukey-Kramer	0.0190	0.05	-2.7662	-0.2138
4	SULFA	2	3	0.3000	0.6508	15.6	Tukey-Kramer	0.9664	0.05	-1.5672	2.1672
5	SULFA	2	4	0.6400	0.6283	14.5	Tukey-Kramer	0.7415	0.05	-1.1785	2.4585
6	SULFA	3	4	0.3400	0.4785	17.7	Tukey-Kramer	0.8916	0.05	-1.0146	1.6946

Fig. 5.17 Approximate multiple comparisons allowing for unequal variances: trout experiment

using Satterthwaite's method to compute the number of degrees of freedom separately for each pairwise comparison. Some of the corresponding multiple comparisons output is shown in Fig. 5.17. The estimates, standard errors, and degrees of freedom match the values in Table 5.5, and the adjusted confidence limits correspond to the values $\bar{y}_{h.} - \bar{y}_{i.} \pm msd$ computable from the estimates and msd values in Table 5.5.

5.9 Using R Software

5.9.1 Residual Plots

We now illustrate use of the R software to generate the various plots used in this chapter. In the following sections, we will check the assumptions on the one-way analysis of variance model (3.3.1) for the data of the mung bean experiment described in Example 5.8.1, p. 120. The experiment was conducted to compare the effects of two treatment factors—"amount of water" (1, 2, or 3 teaspoons of water per day) and "growing medium" (tissue and paper towel, coded 1 and 2)—on the growth of mung beans. The response variable was shoot lengths of the beans measured (in mm) after one week. The experiment was a completely randomized design with eight replicates, and the experimental units

were 48 containers placed in random order on a table. The data were provided in Table 5.7, with the six treatment combinations recoded as $1 = 11, 2 = 12, 3 = 21, 4 = 22, 5 = 31, 6 = 32$.

An R program that generates the residual plots for the mung bean experiment is shown in Table 5.11, with the first three lines of data displayed. After reading the data, the program uses the R function `aov`, introduced in Sect. 3.9.3, to fit model (3.3.1), saving related information as the object `model`. Consequently, the fitted values and residuals are available as the columns `ypred = fitted(model)` and `e = resid(model)`, respectively. The function `sd(e)` computes the sample standard deviation of the residuals, so the column `z = e/sd(e)` contains the standardized residuals. Semi-colons separate commands on the same line. Blom's normal scores are computed by Eq. (5.7.7), p. 117, using the column `q = rank(e)` of ranks of the residuals and the standard normal quantile (inverse cumulative distribution) function `qnorm`, and are saved as the column `nscore`. The qth smallest residual has rank q and yields the qth smallest normal score. Creating these four new variables within the brackets of the statement

```
mung.data = within(mung.data, {...})
```

Table 5.11 R program to generate residual plots: mung bean experiment

```
# R code and output
mung.data = read.table("data/mungbean.txt", header=T)
model1 = aov(Length ~ factor(Trtmt), data=mung.data)

# Compute predicted values, residuals, standardized residuals, normal scores
mung.data = within(mung.data, {
   # Compute predicted, residual, and standardized residual values
   ypred = fitted(model1); e = resid(model1); z = e/sd(e);
   # Compute Blom's normal scores
   n = length(e); q = rank(e); nscore = qnorm((q-0.375)/(n+0.25)) })

# Display first 3 lines of mung.data, 4 digits per variable
print(head(mung.data, 3), digits=4)

    Order Water Medium Length Trtmt  nscore  q  n       z      e ypred
  1     1     2      2   13.2     4  0.9201 40 48  0.9820  4.337 8.863
  2     2     3      2   11.6     6  0.5758 35 48  0.6622  2.925 8.675
  3     3     1      2    0.0     2 -1.6036  3 48 -1.6188 -7.150 7.150

# Generate residual plots
plot(z ~ Trtmt, data=mung.data, ylab="Standardized Residuals", las=1)
  abline(h=0)  # Horizontal line at zero
plot(z ~ Order, data=mung.data, ylab="Standardized Residuals", las=1)
  abline(h=0)
plot(z ~ ypred, data=mung.data, ylab="Standardized Residuals", las=1)
  abline(h=0)
plot(z ~ nscore, data=mung.data, ylab="Standardized Residuals", las=1)
  qqline(mung.data$z)  # Line through 1st and 3rd quantile points
# A simpler way to generate the normal probability plot
qqnorm(mung.data$z); qqline(mung.data$z)
```

enables their creation from variables in the data set `mung.data` and their addition to the data set. Alternatively, the normal scores could be obtained by replacing the three statements for n, q and nscore with the single statement

```
nscore = qqnorm(z)$x
```

though the resulting normal scores are only Blom's normal scores for 10 or fewer residuals, with `nscore = qnorm((q-0.5)/n)` otherwise. Here, `qqnorm` is a plotting function to be discussed shortly that generates a normal probability plot with normal scores on the x axis.

Plots of the standardized residuals z_{it} against treatments, run order, predicted values, and normal scores are generated by the four `plot` function calls. For each of the first three plots, the statement `abline(h=0)` causes inclusion of a horizontal reference line at zero. For the normal probability plot, the statement `qqline(mung.data$z)` causes inclusion of a line through the first and third quantile-quantile points of z and nscore—namely, through the point corresponding to the first quantile of each variable, and through the point corresponding to their third quantiles.

The last three lines of code illustrate an alternative, simpler method of generating the normal probability plot, using the function `qqnorm(z)`. This function generates a normal probability plot, plotting the standardized residuals z against the normal scores—namely, the quantiles of the standard normal distribution. This function uses Blom's normal scores for 10 or fewer z-values, and uses normal scores equal to the $100[(q-0.5)/n]$th percentiles of the standard normal distribution otherwise. These normal scores, corresponding to the x-axis of the plot, can be saved by the command `nscore = qqnorm(z)$x` as noted above. As will be seen, using the `qqnorm` function is convenient if separate normal probability plots are needed for each treatment.

For the mung bean experiment, a plot of the standardized residuals against the order in which the observations are collected is shown in Fig. 5.18, and a plot of standardized residuals against normal

Fig. 5.18 Plot of z_{it} versus order: mung bean experiment
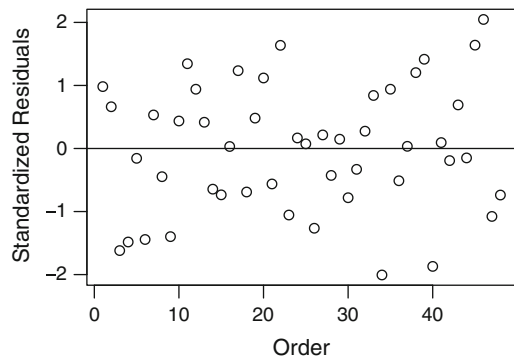
Fig. 5.19 Plot of z_{it} versus normal score: mung bean experiment
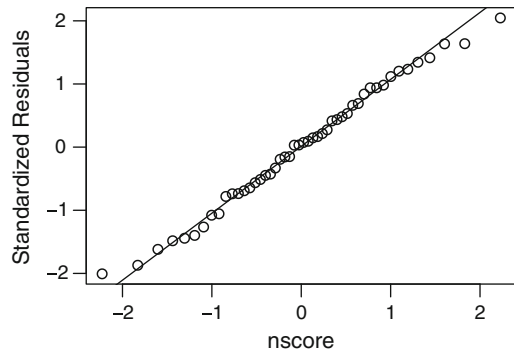
Table 5.12 R program to plot $\ln(s_i^2)$ against $\ln(\bar{y}_{i\cdot})$: mung bean experiment

```
# R Code and Output
mung.data = read.table("data/mungbean.txt", header=T)

# Compute sample means and variances and their natural logs by trtmt
MeanLnth = by(mung.data$Length, mung.data$Trtmt, mean) # Sample means
VarLnth = by(mung.data$Length, mung.data$Trtmt, var) # Sample variances
LnMean = log(MeanLnth)   # Column of ln sample means
LnVar = log(VarLnth)   # Column of ln sample variances
Trtmt = c(1:6)   # Column of trtmt levels
stats = cbind(Trtmt, MeanLnth, VarLnth, LnMean, LnVar) # Column bind
stats   # Display the stats data

    Trtmt MeanLnth VarLnth LnMean  LnVar
  1     1   4.3500  15.171 1.4702 2.7194
  2     2   7.1500  21.117 1.9671 3.0501
  3     3   5.0625  28.057 1.6219 3.3342
  4     4   8.8625  32.803 2.1818 3.4905
  5     5   5.7875  12.156 1.7557 2.4978
  6     6   8.6750  21.679 2.1604 3.0764

plot(LnVar ~ LnMean, las=1)
```

scores is shown in Fig. 5.19. Neither of these plots indicates any serious problems with the assumptions on the model.

A plot of the standardized residuals against the predicted values (not shown) suggests that treatment variances are not too unequal, but that there could be outliers associated with one or two of the treatments. In the R program in Table 5.12, the second block of code computes the sample statistics displayed subsequently by treatment. The `by` function is used to compute the (sample) mean and variance of `Length` by `Trtmt`, saving the results in the columns `MeanLnth` and `VarLnth`, respectively. Then the natural log of each value is computed, saving the log sample means and log sample variances in the columns `LnMean` and `LnVar`, respectively. The `cbind` function column-binds these four columns with another containing the treatment labels, saving them as `stats`, which is then displayed. Given the displayed information, the rule of thumb can be checked that the sample variances should not differ by more than a factor of 3. It can be verified that the ratio of the maximum and minimum variances is under 2.7 for this experiment.

When the equal-variance assumption does not appear to be valid, the experimenter may choose to use an analysis based on Satterthwaite's approximation (see Sect. 5.9.3), using formulas involving the treatment sample variances such as those in Table 5.12. A normal probability plot such as that of Fig. 5.19 would not be relevant, but the normality assumption needs to be checked for each treatment separately. This can be done by generating a separate normal probability plot for each treatment (provided that the sample sizes are sufficiently large). These separate plots are generated by the following single line of R code.

```
by(mung.data$z, mung.data$Trtmt, qqnorm) # Generate NPPlots by Trtmt
```
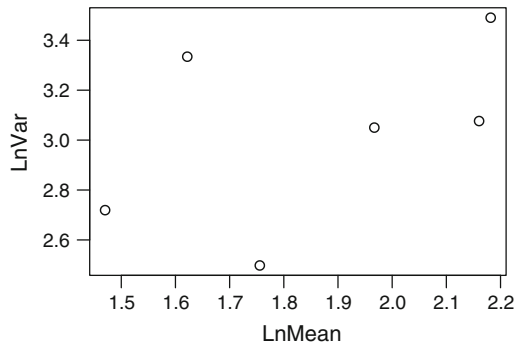
Fig. 5.20   Plot of $\ln(s_i^2)$ against $\ln(\bar{y}_{i.})$; mung bean experiment

The by function applies the function qqnorm to the variable z at each Trtmt level. Alternatively, these plots can be generated one-by-one using the following example for treatment 1, where the main option adds a main title to the plot.

```
qqnorm(mung.data$z[mung.data$Trtmt == 1],
       main = "Normal Probability Plot: Trtmt 1")
   qqline(mung.data$z)
```

5.9.2   Transforming the Data

If a variance-stabilizing transformation is needed, a plot of $\ln(s_i^2)$ against $\ln(\bar{y}_{i.})$ can be achieved as illustrated in the R program in Table 5.12 (shown for the mung bean experiment). First, we need to compute the statistics to be plotted. The R functions mean and var compute sample mean and variance, respectively, of a specified variable and, when coupled with the by function, can compute such statistics for a specified variable at each level of a factor. In our program, the by function in the code line

```
MeanLnth = by(mung.data$Length, mung.data$Trtmt, mean)
```

applies the function mean to the variable Length for (by) each level of Trtmt, saving the resulting sample means as elements of the column MeanLnth. The column VarLnth of sample variances is computed similarly, coupling the by and var functions. The function, log, is then used to compute the natural logarithm, or log base *e*, of the average length and the sample variance for each treatment, saving the results in the columns LnMean and LnVar, respectively. The levels 1–6 of Trtmt are assigned to the new column Trtmt for display purposes. The results are then displayed as columns using the cbind function. Note that these columns of data were created outside of the mung.data data set, since they have fewer entries.

Finally, the plot function generates the plot of $\ln(s_i^2)$ against $\ln(\bar{y}_{i.})$, shown in Fig. 5.20. The values do not fall along a straight line, so a variance-stabilizing transformation of the type given in Eq. (5.6.3) does not exist for this data set. However, since the ratio of the maximum to the minimum variance is less than 3.0, a transformation is not vital, according to our rule of thumb.

If an appropriate transformation is identified, then the transformed variable can be created from the untransformed variable by applying the appropriate R function. R functions useful for the variance-stabilizing transformations of Eq. (5.6.3) include:

Transformation	R Function
$h = \ln(y)$	`h = log(y)`
$h = \sin^{-1}(y)$	`h = asin(y)`
$h = y^p$	`h = y^p`

5.9.3   Implementing Satterthwaite's Method

In Example 5.6.3, given indications of unequal variances in the trout experiment, simultaneous approximate 95% confidence intervals for pairwise comparisons were computed using the Games–Howell method—namely, using Satterthwaite's approximation in conjunction with Tukey's method. This method is implemented in Table 5.13 by reading the author-defined R function `GamesHowell` from the file `GamesHowell.r` in the `funcs` subdirectory of the working directory, then calling this function via the following code line:

```
GamesHowell(y = trout.data$Hemo, T = trout.data$Sulfa, alpha = 0.05)
```

The function inputs are the column of observations `y`, the column of corresponding treatment levels `T`, and the joint significance level α with a default value of 0.05. The results, shown at the bottom of Table 5.13, match the corresponding information in Table 5.5 and Fig. 5.17.

This touches upon an important characteristic of the R software—namely, that one can create user-defined functions to implement methods and procedures that may not otherwise be available as R functions. For example, the code

```
GamesHowell = function(y, T, alpha = 0.05){function code}
```

Table 5.13 R program and output for multiple comparisons with unequal variances: trout experiment

```
trout.data = read.table("data/trout.txt", header = T)
head(trout.data, 3)

    Sulfa Hemo
  1     1  6.7
  2     1  7.8
  3     1  5.5

# Read user-defined function from file GamesHowell.r
source("funcs/GamesHowell.r")
# Call the function, which returns the results displayed below
GamesHowell(y = trout.data$Hemo, T = trout.data$Sulfa, alpha = 0.05)

[[1]]
[1] "Games-Howell method of MCP for tau_i-tau_s with alpha = 0.05"

[[2]]
  i s estimate    stde     df        t       p    msd     lcl      ucl
1 2 4     0.64 0.62831 14.482  1.01860 0.74148 1.8186 -1.1786  2.45860
2 3 4     0.34 0.47854 17.720  0.71050 0.89161 1.3546 -1.0146  1.69460
3 2 3     0.30 0.65083 15.609  0.46095 0.96645 1.8672 -1.5672  2.16720
4 1 4    -1.49 0.45153 17.994 -3.29990 0.01897 1.2762 -2.7662 -0.21381
5 1 3    -1.83 0.48237 17.793 -3.79380 0.00673 1.3649 -3.1949 -0.46514
6 1 2    -2.13 0.63123 14.640 -3.37430 0.01994 1.8246 -3.9546 -0.30540
```

Table 5.14 R function GamesHowell for multiple comparisons with unequal variances

```
# Contents of file GamesHowell.r:
GamesHowell = function(y, T, alpha=0.05){
# y is a data column, T the corresp column of trtmt levels.
# For the y-values corresponding to each level in T, compute:
r  = tapply(y, T, length)  # Column of reps r_i
ybar =  tapply(y, T, mean)  # Column of trtmt sample means ybar_i
s2 =  tapply(y, T, var)  # Column of trtmt sample variances s^2_i
v = length(r)  # v = number of treatments (length of column r)
combos = combn(v,2)  # 2 by v-choose-2, cols being combos (i,s)
i = combos[1,]  # Save row 1, i.e. the i's, as the column i
s = combos[2,]  # Save row 2, i.e. the s's, as the column s
# For each combo (i,s), compute est of tau_i - tau_s, stde, etc.
estimate = combn(v, 2, function(is) -diff(ybar[is]) )  # est's
stde = combn(v, 2, function(is) sqrt(sum(s2[is]/r[is])) )  # stde's
t = estimate/stde  # t-statistics
df = combn(v, 2, function(is)
     (sum(s2[is]/r[is]))^2/(sum((s2[is]/r[is])^2/(r[is]-1)))) )  # df's
p = ptukey(abs(t)*sqrt(2), v, df, lower.tail=F)  # p-values
p = round(p, digits=5)  # Keep at most 5 decimal places
w = qtukey(0.05,v,df,lower.tail=F)/sqrt(2)  # Critical coefficients
msd = w*stde  # msd's
lcl = estimate - msd  # Lower confidence limits
ucl = estimate + msd  # Upper confidence limits
results = cbind(i, s, estimate, stde, df, t, p, msd, lcl, ucl)
results = signif(results, digits=5)  # Keep 5 significant digits
results =  results[rev(order(estimate)),]  # Sort by estimates
rownames(results) = seq(1:nrow(results))  # Name rows 1,2,...,nrows
header=paste("Games-Howell method of MCP for tau_i-tau_s",
             "with alpha =",alpha)
return(list(header,results))
}  # end function
```

uses `function` to create and define a new function named `GamesHowell` in terms of three parameters `y`, `T`, and `alpha`, with 0.05 as the default value of `alpha`. Here "`function code`" would be replaced by R code defining what the function does given the input parameters and what information it returns when done. Such code defining a function can be saved in a separate file then read into a program using the `source` function, as illustrated in Table 5.13. This facilitates reuse of the function in other R programs. Alternatively, the code defining a function can simply be included directly in an R program, replacing the code line `source("GamesHowell.r")` in Table 5.13, for example. For the interested reader, the `GamesHowell` function code is provided and discussed in the following optional subsection.

The User-Defined R Function GamesHowell (Optional)

The author-defined function `GamesHowell` was used in Table 5.13 to implement the Games–Howell method of multiple comparisons. The R code defining the function is provided in Table 5.14 for the interested reader. R functions are defined via the R function `function`. In particular, the code `GamesHowell = function(y, T, alpha=0.05)` indicates that a new function named `GamesHowell` is being defined in terms of three parameters `y`, `T`, and `alpha`, and that the default value of `alpha` is 0.05. All of the subsequent code inside the brackets "`{}`" is the R code defining what the function does.

When calling the function `GamesHowell`, one can specify alpha or not; if not, the default value of 0.05 will be used. The other parameters `y` and `T` represent the column of response values and the corresponding column of treatment levels, respectively. In Table 5.13, the function call `GamesHowell(y=trout.data$Hemo, T=trout.data$Sulfa, alpha=0.05)` explicitly indicates that the column `trout.data$Hemo` contains the response values (`y` in the function) and the column `trout.data$Sulfa` contains the treatment levels (`T` in the function code). If one is explicit, using `y=`, `T=`, and `alpha=`, then the parameters may be entered in any order. Otherwise, they must be entered in the same order (`y,T,alpha`) as they are listed in the definition of the function. For example, the function call `GamesHowell(trout.data$Hemo, trout.data$Sulfa, 0.05)` also works, but not if the parameters were entered in any other order.

This code makes use of the R functions `tapply` and `combn`. Given data for a completely randomized design, the function `tapply(y,T,fn)` applies any specified R function `fn` separately to the subset of the observations `y` corresponding to each `trtmt` level. For example, given observations `y` and corresponding treatment levels `T` for a completely randomized design, the statement `ybar = tapply(y,T,mean)` applies the function `mean` to compute the mean $\bar{y}_{i.}$ of `y` for each level of `T`, saving these as $\text{ybar} = (\bar{y}_{1.}, \ldots, \bar{y}_{v.})$ but as a column. Similarly, `tapply` is used to compute the column `r` of replication numbers r_i and the column `s2` of treatment sample variances s_i^2.

Having v treatments, the function `combn(v,2)` returns the $\binom{v}{2} = v(v-1)/2$ combinations (i, s) of the integers $1, \ldots, v$ taken two at a time as the columns of a matrix, providing the treatment pairs for pairwise comparisons. For each combination or treatment pair (i, s), the function `combn(v,2,function(is),-diff(ybar[is]))` computes the negative difference of the ith and sth elements of the column `ybar`, yielding the column `estimate` of estimates $\bar{y}_{i.} - \bar{y}_{s.}$. The column `stde` of standard errors of the estimates is obtained similarly from the columns `r` and `s2`.

Other functions used include `ptukey` and `qtukey`, pertaining to the Studentized range distribution, (Table A.8). In particular, `ptukey(x,v,df,lower.tail=F)`, which provides the upper-tail probability $P(X > x)$ of the range X of v Studentized variates each involving df degrees of freedom, is used to compute p-values. Likewise, `qtukey(`α`,v,df,lower.tail=F)`, which provides the upper-α quantile of the same distribution, is used to obtain the critical coefficients for the simultaneous confidence intervals.

An R function can return one object, via the `return` function. In this case, the function returns one list consisting of two objects: (i) `header`, containing a description of the statistical procedure conducted; and (ii) `results`, an R data.frame containing the numerical results. This returned information, automatically displayed when the function finishes executing, is shown at the bottom of Table 5.13.

Exercises

1. **Meat cooking experiment, continued**

   Check the assumptions on the one-way analysis of variance model (3.3.1) for the meat cooking experiment, which was introduced in Exercise 14 of Chap. 3. The data were given in Table 3.14. (the order of collection of observations is not available).

2. **Soap experiment, continued**

   Check the assumptions on the one-way analysis of variance model (3.3.1) for the soap experiment, which was introduced in Sect. 2.5.1. The data are reproduced in Table 5.15 (the order of collection of observations is not available).

Table 5.15 Weight loss for the soap experiment

Soap		Weight loss			$\overline{y}_{i.}$	s_i^2
1	−0.30	−0.10	−0.14	0.40	−0.0350	0.09157
2	2.63	2.61	2.41	3.15	2.7000	0.09986
3	1.72	2.07	2.17	2.01	1.9925	0.03736

Table 5.16 Melting times for margarine in seconds

Brand	Times	$\overline{y}_{i.}$	s_i
1	167, 171, 178, 175, 184, 176, 185, 172, 178, 178	176.4	5.56
2	231, 233, 236, 252, 233, 225, 241, 248, 239, 248	238.6	8.66
3	176, 168, 171, 172, 178, 176, 169, 164, 169, 171	171.4	4.27
4	201, 199, 196, 211, 209, 223, 209, 219, 212, 210	208.9	8.45

3. **Margarine experiment** (Amy L. Phelps, 1987)

The data in Table 5.16 are the melting times in seconds for three different brands of margarine (coded 1–3) and one brand of butter (coded 4). The butter was used for comparison purposes. The sizes and shapes of the initial margarine/butter pats were as similar as possible, and these were melted one by one in a clean frying pan over a constant heat.

(a) Check the equal-variance assumption on model (3.3.1) for these data. If a transformation is required, choose the best transformation of the form (5.6.3), and recheck the assumptions.
(b) Using the transformed data, compute a 95% confidence interval comparing the average melting times for the margarines with the average melting time for the butter.
(c) Repeat part (b) using the untransformed data and Satterthwaite's approximation for unequal variances. Compare the results with those of part (b).
(d) For this set of data, which analysis do you prefer? Why?

4. **Reaction time experiment, continued**

The reaction time pilot experiment was described in Exercise 4 of Chap. 4. The experimenters were interested in the different effects on the reaction time of the aural and visual cues and also in the different effects of the elapsed time between the cue and the stimulus. There were six treatment combinations:

$$1 = \text{aural, 5 seconds} \quad 4 = \text{visual, 5 seconds}$$
$$2 = \text{aural, 10 seconds} \quad 5 = \text{visual, 10 seconds}$$
$$3 = \text{aural, 15 seconds} \quad 6 = \text{visual, 15 seconds}$$

The data are reproduced, together with their order of observation, in Table 5.17. The pilot experiment employed a single subject. Of concern to the experimenters was the possibility that the subject may show signs of fatigue. Consequently, fixed rest periods were enforced between every pair of observations.

(a) Check whether or not the assumptions on the one-way analysis of variance model (3.3.1) are approximately satisfied for these data. Pay particular attention to the experimenter's concerns about fatigue.

Table 5.17 Reaction times (in seconds) for the reaction time experiment

Time order	1	2	3	4	5	6
Coded treatment	6	6	2	6	2	5
Reaction time	0.256	0.281	0.167	0.258	0.182	0.283
Time order	7	8	9	10	11	12
Coded treatment	4	5	1	1	5	2
Reaction time	0.257	0.235	0.204	0.170	0.260	0.187
Time order	13	14	15	16	17	18
Coded treatment	3	4	4	3	3	1
Reaction time	0.202	0.279	0.269	0.198	0.236	0.181

(b) Suggest a way to design the experiment using more than one subject. (Hint: consider using subjects as blocks in the experiment).

5. **Catalyst experiment**

H. Smith, in the 1969 volume of *Journal of Quality Technology*, described an experiment that investigated the effect of four reagents and three catalysts on the production rate in a catalyst plant. He coded the reagents as A, B, C, and D, and the catalysts as X, Y, and Z, giving twelve treatment combinations, coded as AX, AY, ..., DZ. Two observations were taken on each treatment combination, and these are shown in Table 5.18, together with the order in which the observations were collected.

Are the assumptions on the one-way analysis of variance model (3.3.1) approximately satisfied for these data? If not, can you suggest what needs to be done in order to be able to analyze the experiment?

6. **Bicycle experiment** (Debra Schomer 1987)

The bicycle experiment was run to compare the crank rates required to keep a bicycle at certain speeds, when the bicycle was in twelfth gear on flat ground. The speeds chosen were 5, 10, 15, 20, and 25 mph, (coded 1–5). The data are given in Table 5.19. The experimenter fitted the one-way

Table 5.18 Production rates for the catalyst experiment

Time order	1	2	3	4	5	6	7	8
Treatment	CY	AZ	DX	AY	CX	DZ	AX	CZ
Yield	9	5	12	7	13	7	4	13
Time order	9	10	11	12	13	14	15	16
Treatment	BY	CZ	BZ	DX	BX	CX	DY	BZ
Yield	13	13	7	12	4	15	12	9
Time order	17	18	19	20	21	22	23	24
Treatment	BX	DY	AY	DZ	BY	AX	CY	AZ
Yield	6	14	11	9	15	6	15	9

Source: Smith (1969). Reprinted with Permission from Journal of Quality Technology © 1969 ASQ, www.asq.org

Table 5.19 Data for the bicycle experiment

Code	Treatment (mph)	Crank rates		
1	5	15	19	22
2	10	32	34	27
3	15	44	47	44
4	20	59	61	61
5	25	75	73	75

analysis of variance model (3.3.1) and plotted the standardized residuals. She commented in her report:

> Note the larger spread of the data at lower speeds. This is due to the fact that in such a high gear, to maintain such a low speed consistently for a long period of time is not only bad for the bike, it is rather difficult to do.

Thus the experimenter was not surprised to find a difference in the variances of the error variables at different levels of the treatment factor.

(a) Plot the standardized residuals against $\widehat{y}_{it}$, compare the sample variances, and evaluate equality of the error variances for the treatments.
(b) Choose the best transformation of the data of the form (5.6.3), and test the hypotheses that the linear and quadratic trends in crank rates due to the different speeds are negligible, using an overall significance level of 0.01.
(c) Repeat part (b), using the untransformed data and Satterthwaite's approximation for unequal variances,
(d) Discuss the relative merits of the methods applied in parts (b) and (c).

7. **Dessert experiment**

(P. Clingan, Y. Deng, M. Geil, J. Mesaros, and J. Whitmore, 1996)
The experimenters were interested in whether the melting rate of a frozen orange dessert would be affected (and, in particular, slowed down) by the addition of salt and/or sugar. At this point, they were not interested in taste testing. Six treatments were selected, as follows:

$$1 = 1/8 \text{ tsp salt, } 1/4 \text{ cup sugar} \quad 4 = 1/4 \text{ tsp salt, } 1/4 \text{ cup sugar}$$
$$2 = 1/8 \text{ tsp salt, } 1/2 \text{ cup sugar} \quad 5 = 1/4 \text{ tsp salt, } 1/2 \text{ cup sugar}$$
$$3 = 1/8 \text{ tsp salt, } 3/4 \text{ cup sugar} \quad 6 = 1/4 \text{ tsp salt, } 3/4 \text{ cup sugar}$$

For each observation of each treatment, the required amount of sugar and salt was added to the contents of a 12-ounce can of frozen orange juice together with 3 cups of water. The orange juice mixes were frozen in ice cube trays and allocated to random positions in a freezer. After 48 hours, the cubes were removed from the freezer, placed on half-inch mesh wire grid and allowed to melt into a container in the laboratory (which was held at 24.4 °C) for 30 minutes. The percentage melting (by weight) of the cubes are recorded in Table 5.20. The coded position on the table during melting is also recorded.

(a) Plot the data. Does it appear that the treatments have different effects on the melting of the frozen orange dessert?

Table 5.20 Percentage melting of frozen orange cubes for the dessert experiment

Position	1	2	3	4	5	6
Treatment	2	5	5	1	4	3
% melt	12.06	9.66	7.96	9.04	10.17	7.86
Position	7	8	9	10	11	12
Treatment	4	1	3	1	2	4
% melt	8.14	9.52	4.28	8.32	10.74	5.98
Position	13	14	15	16	17	18
Treatment	2	6	6	3	6	5
% melt	9.84	7.58	6.65	9.26	8.46	12.83

(b) Check whether the assumptions on the one-way analysis of variance model (3.3.1) are satisfied for these data. Pay particular attention to the equal-variance assumption.

(c) Use Satterthwaite's method to compare the pairs of treatments, using individual 99% confidence intervals. If doing the computations by hand, compute only the confidence intervals corresponding to the three most disparate pairs of treatment sample means.

(d) What conclusions can you draw about the effects of the treatments on the melting of the frozen orange dessert? If your concern was to produce frozen dessert with a long melting time, which treatment would you recommend? What other factors should be taken into account before production of such a dessert?

8. **Wildflower experiment** (Barbra Foderaro 1986)

An experiment was run to determine whether or not the germination rate of the endangered species of Ohio plant *Froelichia floridana* is affected by storage temperature or storage method. The two levels of the factor "temperature" were "spring temperature, 14–24 °C" and "summer temperature, 18–27 °C." The two levels of the factor "storage" were "stratified" and "unstratified." Thus, there were four treatment combinations in total. Seeds were divided randomly into sets of 20 and the sets assigned at random to the treatments. Each stratified set of seeds was placed in a mesh bag, spread out to avoid overlapping, buried in two inches of moist sand, and placed in a refrigeration unit for two weeks at 50 °F. The unstratified sets of seeds were kept in a paper envelope at room temperature. After the stratification period, each set of seeds was placed on a dish with 5 ml of distilled deionized water, and the dishes were put into one of two growth chambers for two weeks according to their assigned level of temperature. At the end of this period, each dish was scored for the number of germinated seeds. The resulting data are given in Table 5.21.

(a) For the original data, evaluate the constant-variance assumption on the one-way analysis of variance model (3.3.1) both graphically and by comparing sample variances.

(b) It was noted by the experimenter that since the data were the numbers of germinated seeds out of a total of 20 seeds, the observations Y_{it} should have a binomial distribution. Does the corresponding transformation help to stabilize the variances?

(c) Plot $\ln(s_t^2)$ against $\ln(\bar{y}_{i.})$ and discuss whether or not a power transformation of the form given in Eq. (5.6.3) might equalize the variances.

(d) Use Scheffé's method of multiple comparisons, in conjunction with Satterthwaite's approximation, to construct 95% confidence intervals for all pairwise comparisons and for the two contrasts

Table 5.21 Data for the wildflower experiment

Treatment combination	Number germinating					$\overline{y}_{i.}$	s_i
1: Spring/stratified	12	13	2	7	19	8.4	6.995
	0	0	3	17	11		
2: Spring/unstratified	6	2	0	2	4	2.5	3.308
	1	0	10	0	0		
3: Summer/stratified	6	4	5	7	6	5.0	1.633
	5	7	5	2	3		
4: Summer/unstratified	0	6	2	5	1	3.6	2.271
	5	2	3	6	6		

Table 5.22 Weights (in grams) for the spaghetti sauce experiment

Time order	1	2	3	4	5	6	7	8	9
Treatment	3	2	4	3	4	5	1	6	6
Weight	14	69	26	15	20	12	55	14	16

Time order	10	11	12	13	14	15	16	17	18
Treatment	5	1	2	4	6	3	5	2	1
Weight	16	66	64	23	17	22	18	64	53

$$\frac{1}{2}[1, 1, -1, -1] \quad \text{and} \quad \frac{1}{2}[1, -1, 1, -1],$$

which compare the effects of temperature and storage methods, respectively.

9. **Spaghetti sauce experiment**

(K. Brewster, E. Cesmeli, J, Kosa, M. Smith, and M. Soliman 1996)
The spaghetti sauce experiment was run to compare the thicknesses of three particular brands of spaghetti sauce, both when stirred and unstirred. The six treatments were:

> 1 = store brand, unstirred   2 = store brand, stirred
> 3 = national brand, unstirred   4 = national brand, stirred
> 5 = gourmet brand, unstirred 6 = gourmet brand, stirred

Part of the data collected is shown in Table 5.22. There are three observations per treatment, and the response variable is the weight (in grams) of sauce that flowed through a colander in a given period of time. A thicker sauce would give rise to smaller weights.

(a) Check the assumptions on the one-way analysis of variance model (3.3.1).
(b) Use Satterthwaite's method to obtain simultaneous confidence intervals for the six preplanned contrasts
$$\tau_1 - \tau_2, \quad \tau_3 - \tau_4, \quad \tau_5 - \tau_6, \quad \tau_1 - \tau_5, \quad \tau_1 - \tau_3, \quad \tau_3 - \tau_5,$$

Select an overall confidence level of at least 94%.

Experiments with Two Crossed Treatment Factors

6

6.1 Introduction

In this chapter, we discuss the use of completely randomized designs for experiments that involve two crossed treatment factors. We label the treatment factors as A and B, where factor A has a levels coded 1, 2, . . . , a, and factor B has b levels coded 1, 2, . . . , b. Factors are *crossed* if every combination of levels may be observed. For experiments considered in this chapter, every level of A is observed with every level of B, so the factors are crossed. In total, there are $v = ab$ treatments (treatment combinations), and these are coded as 11, 12, . . . , 1b, 21, 22, . . . , 2b, . . . , ab.

In the previous three chapters, we recoded the treatment combinations as 1, 2, . . . , v and used the one-way analysis of variance for comparing their effects. In this chapter, we investigate the contributions that each of the factors make individually to the response, and it is more convenient to retain the 2-digit code ij for a treatment combination in which factor A is at level i and factor B is at level j. In Sect. 6.2.1, we define the "interaction" of two treatment factors. Allowing for the possibility of interaction leads one to select a "two-way complete model" to model the data (Sect. 6.4). However, if it is known in advance that the factors do not interact, a "two-way main-effects model" would be selected (Sect. 6.5). Estimation of contrasts, confidence intervals, and analysis of variance techniques are described for these basic models. The calculation of sample sizes is also discussed (Sect. 6.6). The corresponding commands for SAS and R software are described in Sects. 6.8 and 6.9, respectively.

If each of the two factors has a large number of levels, the total number of treatment combinations could be quite large. When observations are costly, it may be necessary to limit the number of observations to one per treatment combination. Analysis for this situation is discussed in Sect. 6.7.

6.2 Models and Factorial Effects

6.2.1 The Meaning of Interaction

In order to understand the meaning of the interaction between two treatment factors, it is helpful to look at possible data sets from a hypothetical experiment. Universities have become increasingly interested in online courses and other nontraditional modes of instruction. While an online course may be offered for a group of students and involve interaction between the students in a common section, consider development of a course that students take independently of one another. Suppose that a hypothetical statistics department wishes to know to what extent student performance in an introductory

Fig. 6.1 Possible
configurations of effects
present for two factors,
presentation format (F) and
course structure (S) when
the significant interaction
effect is absent

(a) F=no, S=no, FS=no

(b) F=no, S=yes, FS=no

(c) F=yes, S=no, FS=no

(d) F=yes, S=yes, FS=no
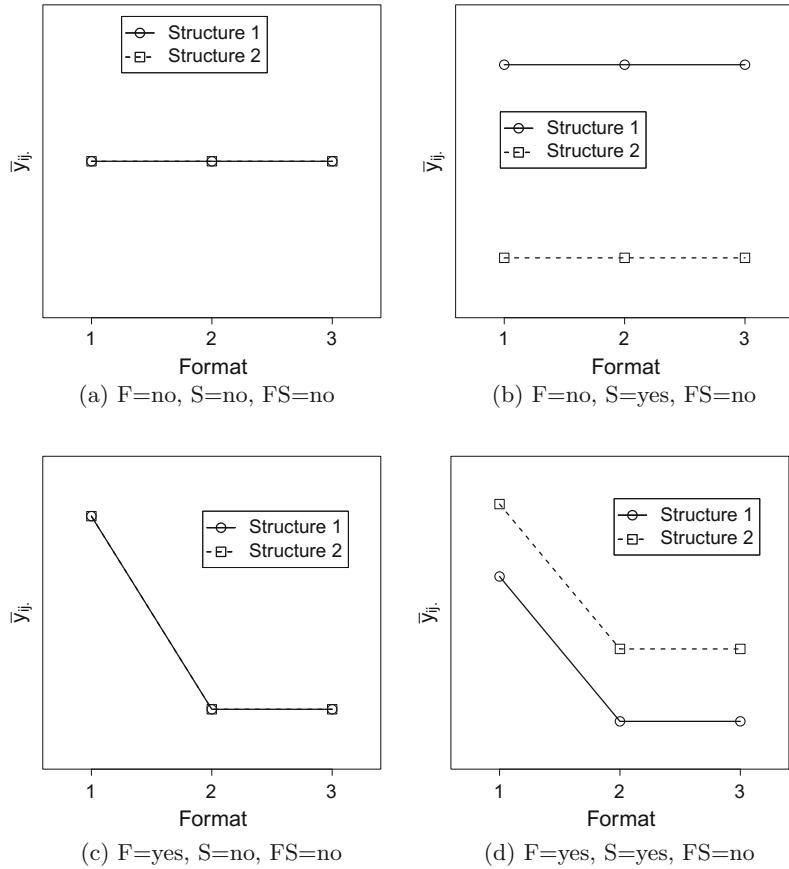
online course is affected by the primary presentation format (textbook reading assignments, videotaped lectures, or interactive software) and course structure (structured, with regular deadlines throughout the term; or unstructured, with only a deadline to finish by the end of the term).

There are two treatment factors of interest, namely "presentation format," which has three levels, coded 1, 2, and 3, and "course structure," which has two levels, coded 1 and 2. Both of the treatment factors have fixed effects, since their levels have been specifically chosen (see Sect. 2.2, p. 11, step (f)). The students who enroll in the introductory course are the experimental units and are allocated at random to one of the six treatment combinations in such a way that approximately equal numbers of students are assigned to each combination of presentation format and course structure. Student performance is to be measured by means of a computer-graded multiple-choice examination, and an average exam score $\bar{y}_{ij.}$ for each treatment combination will be obtained, averaging over students for each treatment combination.

There are eight different types of situations that could occur, and these are depicted in Figs. 6.1 and 6.2, where the plotted character indicates the course structure used. The plots are called *interaction plots* and give an indication of how the different format–structure combinations affect the average exam score.

In plots (a)–(d) of Fig. 6.1, the lines joining the average exam scores for the two course structures are parallel (and sometimes coincide). In plot (b), all the presentation formats have obtained higher exam scores with course structure 1 than with structure 2, but the presentation formats themselves look very similar in terms of the average exam scores obtained. Thus there is an effect on the average exam

Fig. 6.2 Possible configurations of effects present for two factors, presentation format (F) and course structure (S) when the significant interaction effect is present

(a) F=no?, S=yes, FS=yes

(b) F=yes, S=no?, FS=yes

(c) F=no?, S=no?, FS=yes

(d) F=yes, S=yes, FS=yes
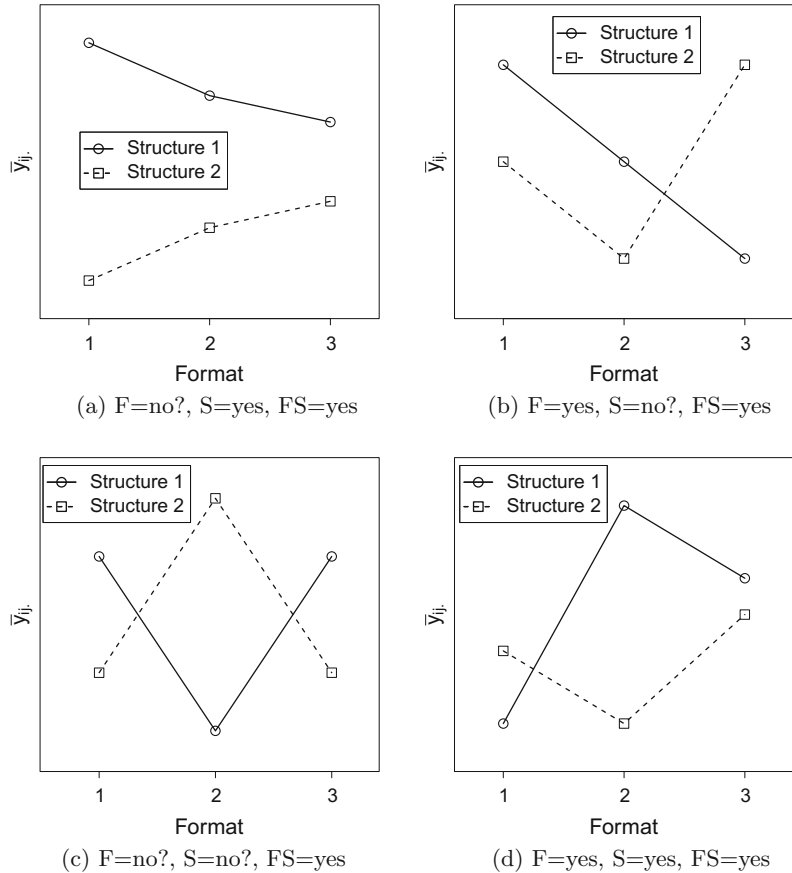
score of course structure (S) but no effect of presentation format (F). Below the plot this is highlighted by the notation "F = no, S = yes." The notation "FS = no" refers to the fact that the lines are parallel, indicating that there is no interaction (see below). In plot (c), no difference can be seen in the average scores obtained from the two course structures for any presentation format, although the presentation formats themselves appear to have achieved different average scores. Thus, the presentation formats have an effect on the average exam score, but the course structures do not (F = yes, S = no). Plot (d) shows the type of plot that might be obtained if there is both a presentation-format effect and a course-structure effect. The plot shows that all three presentation formats have obtained higher average exam scores using structure 1 than using structure 2. But also, presentation format 1 has obtained higher average scores than the other two presentation formats. The individual course-structure effects and presentation-format effects are known as *main effects*.

In plots (a)–(d) of Fig. 6.2, the lines are not parallel. This means that more is needed to explain the differences in exam scores than just course structure and presentation format effects. For example, in plot (a), all presentation formats have obtained higher exam scores using course structure 1 than using structure 2, but the difference is very small for presentation format 3 and very large for presentation format 1. In plot (d), presentation format 1 has obtained higher exam scores with structure 2, while the other two presentation formats have done better with structure 1. In all of plots (a)–(d) the presentation formats have performed differently with the different structures. This is called an effect of *interaction* between presentation format and course structure.

In plot (c), the presentation formats clearly differ. Two do better with structure 1 and one with structure 2. However, if we ignore course structures, the presentation formats appear to have achieved very similar average exam scores overall. So, averaged over the structures, there is little difference between them. In such a case, a standard computer analysis will declare that there is no difference between presentation formats, which is somewhat misleading. We use the notation "FS = yes" to denote an interaction between presentation format and Structure, and "F = no?" to highlight the fact that a conclusion of no difference between presentation formats should be interpreted with caution in the presence of interaction. In general, *if there is an interaction between two treatment factors, then it may not be sensible to examine either of the main effects separately*. Instead, it will often be preferable to compare the effects of the treatment combinations themselves.

While interaction plots are extremely helpful in interpreting the analysis of an experiment, they give no indication of the size of the experimental error. Sometimes a perceived interaction in the plot will not be distinguishable from error variability in the analysis of variance. On the other hand, if the error variability is very small, then an interaction effect may be statistically significant in the analysis, even if it appears negligible in the plot.

6.2.2   Models for Two Treatment Factors

If we use the two-digit codes ij for the treatment combinations in the one-way analysis of variance model (3.3.1), we obtain the model

$$
\begin{aligned}
Y_{ijt} &= \mu + \tau_{ij} + \epsilon_{ijt}\,, \\
\epsilon_{ijt} &\sim N(0, \sigma^2)\,, \\
\epsilon_{ijt}&'\text{s independent}\,, \\
t = 1, \ldots, r_{ij}; \quad i &= 1, \ldots, a; \quad j = 1, \ldots, b,
\end{aligned} \tag{6.2.1}
$$

where i and j are the levels of A and B, respectively. This model is known as the *cell-means model*. The "cell" refers to the cell of a table whose rows represent the levels of A and whose columns represent the levels of B.

Since the interaction plot arising from a two-factor experiment could be similar to any of the plots of Figs. 6.1 and 6.2, it is often useful to model the effect on the response of treatment combination ij to be the sum of the individual effects of the two factors, together with their interaction; that is,

$$
\tau_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij}.
$$

Here, α_i is the effect (positive or negative) on the response due to the fact that the ith level of factor A is observed, and β_j is the effect (positive or negative) on the response due to the fact that the jth level of factor B is observed, and $(\alpha\beta)_{ij}$ is the extra effect (positive or negative) on the response of observing levels i and j of factors A and B together. The corresponding model, which we call the *two-way complete model,* or the *two-way analysis of variance model,* is as follows:

$$
\begin{aligned}
Y_{ijt} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijt}\,, \\
\epsilon_{ijt} &\sim N(0, \sigma^2)\,, \\
\epsilon_{ijt}&'\text{s are mutually independent}\,, \\
t = 1, \ldots, r_{ij}; \quad i &= 1, \ldots, a; \quad j = 1, \ldots, b.
\end{aligned} \tag{6.2.2}
$$

The phrase "two-way" refers to the fact that there are two primary sources of variation, namely, the two treatment factors. Model (6.2.2) is equivalent to model (6.2.1), since all we have done is to express the effect of the treatment combination in terms of its constituent parts.

   Occasionally, an experimenter has sufficient knowledge about the two treatment factors being studied to state with reasonable certainty that the factors do not interact and that an interaction plot similar to one of the plots of Fig. 6.1 will occur. This knowledge may be gleaned from previous similar experiments or from scientific facts about the treatment factors. If this is so, then the interaction term can be dropped from model (6.2.2), which then becomes

$$
\begin{aligned}
Y_{ijt} &= \mu + \alpha_i + \beta_j + \epsilon_{ijt}\,, \\
\epsilon_{ijt} &\sim N(0, \sigma^2)\,, \\
\epsilon_{ijt}&\text{'s are mutually independent}\,, \\
t = 1, \ldots, r_{ij}; \quad i &= 1, \ldots, a; \quad j = 1, \ldots, b\,.
\end{aligned}
\tag{6.2.3}
$$

Model (6.2.3) is a "submodel" of the two-way complete model and is called a *two-way main-effects model,* or *two-way additive model*, since the effect on the response of treatment combination *ij* is modeled as the sum of the individual effects of the two factors. If an additive model is used when the factors really do interact, then inferences on main effects can be very misleading. Consequently, if the experimenter does not have reasonable knowledge about the interaction, then the two-way complete model (6.2.2) or the equivalent cell-means model (6.2.1) should be used.

6.2.3   Checking the Assumptions on the Model

The assumptions implicit in both the two-way complete model (6.2.2) and the two-way main-effects model (6.2.3) are that the error random variables have equal variances, are mutually independent, and are normally distributed. The strategy and methods for checking the error assumptions are the same as those in Chap. 5. The standardized residuals are calculated as

$$
z_{ijt} = (y_{ijt} - \hat{y}_{ijt})/\sqrt{ssE/(n-1)}
$$

with

$$
\hat{y}_{ijt} = \hat{\tau}_{ij} = \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij}
$$

or

$$
\hat{y}_{ijt} = \hat{\tau}_{ij} = \hat{\alpha}_i + \hat{\beta}_j\,,
$$

depending upon which model is selected, where the "hat" denotes a least squares estimate. The residuals are plotted against

  (i) the order of observation to check independence,
 (ii) the levels of each factor and $\hat{y}_{ijt}$ to check for outliers and for equality of variances,
(iii) the normal scores to check the normality assumption.

When the main-effects model is selected, interaction plots of the data, such as those in Figs. 6.2 and 6.1, can be used to check the assumption of no interaction. An alternative way to check for interaction is to plot the standardized residuals against the levels of one of the factors with the plotted labels being the levels of the second factor. An example of such a plot is shown in Fig. 6.3. (For details of the original

Fig. 6.3 Residual plot for
the temperature experiment
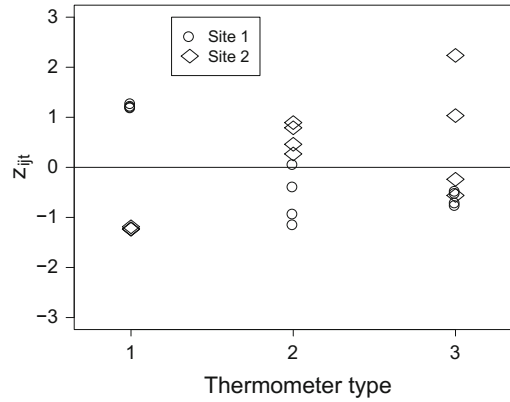
experiment, see Exercise 17.9.1, p. 650.) If the main-effects model had represented the data well, then the residuals would have been randomly scattered around zero. However, a pattern can be seen that is reminiscent of the interaction plot (b) of Fig. 6.2 suggesting that a two-way complete model would have been a much better description of the data. If the model is changed based on the data, subsequent stated confidence levels and significance levels will be inaccurate, and analyses must be interpreted with caution.

If there is some doubt about the equality of the variances, the rule of thumb $s^2_{max}/s^2_{min} < 3$ can be employed, where s^2_{max} is the maximum of the variances of the data values within the cells, and s^2_{min} is the minimum (see Sect. 5.6.1). In a two-way layout, however, there may not be sufficient observations per cell to allow this calculation to be made. Nevertheless, we can at least check that the error variances are the same for each level of any given factor by employing the rule of thumb for the variances of the nonstandardized residuals calculated at each level of the factor.

6.3   Contrasts

6.3.1   Contrasts for Main Effects and Interactions

Since the cell-means model (6.2.1) is equivalent to the one-way analysis of variance model, we know that all contrasts in the treatment effects τ_{ij} are estimable (cf. Sect. 3.4.1, p. 34). Contrasts of interest for a cell-means model are typically of three main types: treatment contrasts, interaction contrasts, and main-effect contrasts.

Treatment contrasts $\Sigma_i \Sigma_j d_{ij} \tau_{ij}$ are no different from the types of contrasts described in Chap. 4. For example, $\tau_{ij} - \tau_{sh}$ is a pairwise difference between treatment combinations ij and sh. All the confidence interval methods of Chap. 4 are directly applicable.

Interaction contrasts are the contrasts that we use in order to measure whether or not the lines on the interaction plots (cf. Figs. 6.1 and 6.2) are parallel. An example of an interaction contrast is

$$(\tau_{sh} - \tau_{(s+1)h}) - (\tau_{sq} - \tau_{(s+1)q}) \,. \tag{6.3.4}$$

We can verify that this is, indeed, an interaction contrast by using the equivalent two-way complete model notation with $\tau_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij}$. Substituting this into (6.3.4) gives the contrast

$$\left((\alpha\beta)_{sh} - (\alpha\beta)_{(s+1)h}\right) - \left((\alpha\beta)_{sq} - (\alpha\beta)_{(s+1)q}\right) \,, \tag{6.3.5}$$

which is a function of interaction parameters only. Interaction contrasts are always of the form

$$\sum_i \sum_j d_{ij} \tau_{ij} = \sum_i \sum_j d_{ij} (\alpha\beta)_{ij}, \tag{6.3.6}$$

where

$$\sum_i d_{ij} = 0 \text{ for each } j \quad \text{and} \quad \sum_j d_{ij} = 0 \text{ for each } i.$$

Some, but not all, interaction contrasts have coefficients $d_{ij} = c_i k_j$. For example, if we take $c_s = k_h = 1$ and $c_{s+1} = k_q = -1$ and all other c_i and k_j zero, then, setting $d_{ij} = c_i k_j$ in (6.3.6), we obtain the coefficients in contrast (6.3.5).

If the interaction effect is very small, then the lines on an interaction plot are almost parallel (as in plots (a)–(d) of Fig. 6.1). We can then compare the average effects of the different levels of A (averaging over the levels of B). Thus, contrasts of the form $\Sigma c_i \bar{\tau}_{i.}$, with $\Sigma c_i = 0$, would be of interest. However, if there is an interaction (as in plot (c) of Fig. 6.2), such an average may make little sense. This becomes obvious when we use the two-way complete model formulation, since a main effect contrast in A is

$$\sum_i c_i \bar{\tau}_{i.} = \sum_i c_i (\alpha_i + \overline{(\alpha\beta)}_{i.}) \tag{6.3.7}$$

where $\overline{(\alpha\beta)}_{i.} = \frac{1}{b} \sum_j (\alpha\beta)_{ij}$, and we can see clearly that we have averaged over any interaction effect that might be present. We will often write

$$\alpha_i^* = \alpha_i + \overline{(\alpha\beta)}_{i.} \text{ and } \beta_j^* = \beta_j + \overline{(\alpha\beta)}_{.j}$$

for convenience. A contrast in the main effect of A for the two-way complete model is then written as $\Sigma c_i \alpha_i^*$ ($\Sigma c_i = 0$), and a contrast in the main effect of B is

$$\sum_j k_j \bar{\tau}_{.j} = \sum_j k_j (\beta_j + \overline{(\alpha\beta)}_{.j}) = \sum_j k_j \beta_j^*, \tag{6.3.8}$$

where $\Sigma k_j = 0$ and $\overline{(\alpha\beta)}_{.j} = \frac{1}{a} \sum_i (\alpha\beta)_{ij}$.

Sometimes, it is of interest to compare the effects of the levels of one factor separately at each level of the other factor. Consider a variation on the hypothetical experiment in Sect. 6.2.1. Suppose the hypothetical statistics department also wishes to study the effects on student learning of two pedagogies (traditional lecture, and discovery-based learning) for three instructors teaching an introductory statistics course. Unless the department wants all instructors (factor A, say) to use the same pedagogy (factor B, say) in teaching the course, a natural objective might be to choose a best pedagogy for each instructor separately. If comparison of the effects of levels of factor B for each level of factor A is required, then contrasts of the form

$$\sum_j c_j \tau_{ij}, \quad \text{with} \quad \sum_j c_j = 0 \text{ for each } i = 1, 2, \dots, a,$$

are of interest. We call such contrasts *simple contrasts* in the levels of B. As a special case, we have the *simple pairwise differences* of factor B:

$$\tau_{ih} - \tau_{ij}, \quad \text{for each } i = 1, \ldots, a.$$

These are a subset of the pairwise comparison contrasts. Simple contrasts and simple pairwise differences of factor A are defined in an analogous way.

When it is known in advance of the experiment that factors A and B do not interact, the two-way main-effects model (6.2.3) would normally be used. In this model, there is no interaction term, so $\tau_{ij} = \alpha_i + \beta_j$. The main-effects contrasts for A and B are respectively of the form

$$\sum c_i \bar{\tau}_{i.} = \sum c_i \alpha_i \text{ and } \sum k_j \bar{\tau}_{.j} = \sum k_j \beta_j,$$

with $\sum c_i = 0$ and $\sum k_j = 0$.

6.3.2   Writing Contrasts as Coefficient Lists

Instead of writing out a contrast explicitly, it is sometimes sufficient, and more convenient, to list the contrast coefficients only. For the two-way complete model, we have a choice. We can refer to contrasts as either a list of coefficients of the parameters α_i^*, β_j^*, and $(\alpha\beta)_{ij}$ or as a list of coefficients of the τ_{ij}'s. This is illustrated in the following example.

Example 6.3.1 Battery experiment, continued

The four treatment combinations in the battery experiment of Sect. 2.5.2, p. 24, involved two treatment factors, "duty" and "brand," each having two levels (1 for alkaline and 2 for heavy duty; 1 for name brand and 2 for store brand), giving treatment combinations 11, 12, 21, and 22. (These were coded in previous examples as 1, 2, 3, and 4, respectively.) There were $r = 4$ observations on each treatment combination.

The interaction plot in Fig. 6.4 shows a possible interaction between the two factors, since the dotted lines on the plot are not close to parallel. However, we should remember that we cannot be certain whether the nonparallel lines are due to an interaction or to inherent variability in the data, and we will need to investigate the cause in more detail later.

The interaction is measured by the contrast

$$\tau_{11} - \tau_{12} - \tau_{21} + \tau_{22} = (\alpha\beta)_{11} - (\alpha\beta)_{12} - (\alpha\beta)_{21} + (\alpha\beta)_{22},$$

Fig. 6.4 Plot of average life per unit cost against "Duty" level i by "Brand" level j for the battery experiment
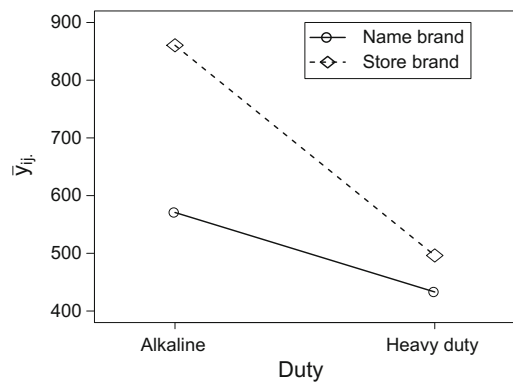
which can be written in terms of the coefficient list $[\,1,\,-1,\,-1,\,1\,]$.

The contrast that compares the average lifetimes of heavy duty and alkaline batteries (averaged across brands) is

$$\bar{\tau}_{2.} - \bar{\tau}_{1.} = \frac{1}{2}\left(\tau_{21} + \tau_{22}\right) - \frac{1}{2}\left(\tau_{11} + \tau_{12}\right) = \alpha_2^* - \alpha_1^*.$$

This has coefficient list $[-1,\,1\,]$ in terms of the effects α_1^*, α_2^* of the levels of duty, but coefficient list $\frac{1}{2}[-1,\,-1,\,1,\,1\,]$ in terms of the effects τ_{11}, τ_{12}, τ_{21}, τ_{22} of the treatment combinations. Similarly, the contrast that compares the average life of store brand with that of name brand (averaged over duty) has coefficient list $[-1,\,1\,]$ in terms of the effects β_j^* of brand, but coefficient list $\frac{1}{2}[-1,\,1,\,-1,\,1\,]$ in terms of the τ_{ij}'s.

Since the main-effect contrasts each have divisor 2, the interaction contrast is often divided by 2 also. This has the effect that the least squares estimators of all three contrasts have the same variances (see Example 6.4.1), and their magnitudes are more directly comparable. An alternative way to achieve equal variances is to normalize the contrasts (see Sect. 4.2), in which case all three contrasts would all be divided by $\sqrt{\Sigma c_i^2/r}$.   □

Contrast coefficients are often listed as columns in a table. For example, the contrast coefficients of the τ_{ij}'s for the main effect and interaction contrasts of Example 6.3.1 are written as below, with ± 1's in the body of the table, and the constants listed as divisors in the last row.

ij	A	B	AB
11	-1	-1	1
12	-1	1	-1
21	1	-1	-1
22	1	1	1
Divisor	2	2	2

The benefit of this representation is that we can see easily that each AB interaction coefficient can be obtained by multiplying the corresponding A and B main-effect coefficients. Most of the interaction contrasts that we shall use have this product form. We will mention the exceptions when they arise.

Example 6.3.2 `Trend contrasts`

Suppose that the two factors, A and B, have $a = 3$ and $b = 6$ equally spaced quantitative levels, respectively, and that the sample sizes are equal. From Table A.2, we see that A_{L}, the linear trend contrast for A, has contrast coefficient list $[-1,\,0,\,1\,]$ in terms of the α_i^*'s, and A_{Q}, the quadratic trend contrast for A, has contrast coefficient list $[\,1,\,-2,\,1\,]$; that is

$$A_{\mathrm{L}} = -\alpha_1^* + \alpha_3^*,$$
$$A_{\mathrm{Q}} = \alpha_1^* - 2\alpha_2^* + \alpha_3^*.$$

Similarly, in terms of the β_j^*'s, the coefficient lists for the linear and quadratic trends in the effects of the six levels of B are also obtained from Table A.2 as $[-5,\,-3,\,-1,\,1,\,3,\,5\,]$ and $[\,5,\,-1,\,-4,\,-4,\,-1,\,5\,]$, respectively; that is,

$$B_{\mathrm{L}} = -5\beta_1^* - 3\beta_2^* - \beta_3^* + \beta_4^* + 3\beta_5^* + 5\beta_6^*,$$
$$B_{\mathrm{Q}} = 5\beta_1^* - \beta_2^* - 4\beta_3^* - 4\beta_4^* - \beta_5^* + 5\beta_6^*.$$

Now,

$$\alpha_i^* = \overline{\tau}_{i.}, \text{ giving } \Sigma_i c_i \alpha_i^* = \Sigma_i c_i \left(\frac{1}{6} \Sigma_j \tau_{ij} \right) = \frac{1}{6} \Sigma_i \Sigma_j c_i \tau_{ij} ,$$

and

$$\beta_j^* = \overline{\tau}_{.j}, \text{ giving } \Sigma_j k_j \beta_j^* = \Sigma_j k_j \left(\frac{1}{3} \Sigma_i \tau_{ij} \right) = \frac{1}{3} \Sigma_i \Sigma_j k_j \tau_{ij} ,$$

and we can write all of the above trends in terms of contrasts in τ_{ij}, as shown in the columns of Table 6.1. Contrast coefficients are also listed for cubic, quartic, and quintic trends for B. If we wish to compare the A and B trends on the same scale, we can normalize the contrasts (see Sect. 4.2).

In order to model a three-dimensional surface, we need to know not only how the response is affected by the levels of each factor averaged over the levels of the other factor, but also how the response changes as the levels of A and B change together. The linear $A \times$ linear B trend ($A_L B_L$) measures whether or not the linear trend in A changes in a linear fashion as the levels of B are increased, and vice versa. This is an interaction contrast whose coefficients are of the form $d_{ij} = c_i k_j$, where c_i are the contrast coefficients for A, and k_j are the contrast coefficients for B. The $A_L B_L$ contrast coefficients are shown in Table 6.1, and it can be verified that they are obtained by multiplying together corresponding main-effect linear trend coefficients in the same row. Coefficients for the linear $A \times$ quintic B ($A_L B_{qn}$) contrast is also shown for use later in this chapter. □

Table 6.1 Trend contrasts when A and B have 3 and 6 equally spaced levels, respectively

ij	A_L	A_Q	B_L	B_Q	B_C	B_{qr}	B_{qn}	$A_L B_L$	$A_L B_{qn}$
11	−1	1	−5	5	−5	1	−1	5	1
12	−1	1	−3	−1	7	−3	5	3	−5
13	−1	1	−1	−4	4	2	−10	1	10
14	−1	1	1	−4	−4	2	10	−1	−10
15	−1	1	3	−1	−7	−3	−5	−3	5
16	−1	1	5	5	5	1	1	−5	−1
21	0	−2	−5	5	−5	1	−1	0	0
22	0	−2	−3	−1	7	−3	5	0	0
23	0	−2	−1	−4	4	2	−10	0	0
24	0	−2	1	−4	−4	2	10	0	0
25	0	−2	3	−1	−7	−3	−5	0	0
26	0	−2	5	5	5	1	1	0	0
31	1	1	−5	5	−5	1	−1	−5	−1
32	1	1	−3	−1	7	−3	5	−3	5
33	1	1	−1	−4	4	2	−10	−1	−10
34	1	1	1	−4	−4	2	10	1	10
35	1	1	3	−1	−7	−3	−5	3	−5
36	1	1	5	5	5	1	1	5	1
Divisor	6	6	3	3	3	3	3	1	1

6.4   Analysis of the Two-Way Complete Model

In the analysis of an experiment with two treatment factors that possibly interact, we may proceed with the analysis in two equivalent ways. We may use the cell-means model (6.2.1) together with all the analysis techniques of Chaps. 3 and 4, or we may use the two-way complete model (6.2.2) and isolate the contributions to the response made by each of the two factors and their interaction separately.

A sensible strategy is to start with the two-way complete model and test a hypothesis of no interaction. If the hypothesis is not rejected, we may then continue with the analysis by examining the main effects under the same two-way complete model. We would not change to the two-way main-effects model, since this is not an equivalent model. However, if the hypothesis of no interaction is rejected, then we would normally prefer to change to the equivalent cell-means model and examine differences in the effects of the treatment combinations. We would also use the cell-means model when the objective of the experiment is to find the best treatment combination.

6.4.1   Least Squares Estimators for the Two-Way Complete Model

As in Sect. 3.4.3, p. 35, the least squares estimator of $\mu + \tau_{ij}$ is $\overline{Y}_{ij.}$, so the least squares estimators of the parameters in the cell-means model (6.2.1) and the equivalent two-way complete model (6.2.2) are

$$\hat{\mu} + \hat{\tau}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij} = \overline{Y}_{ij.} \,,$$

and the corresponding variance is σ^2/r_{ij}. Any interaction contrast of the form $\Sigma\Sigma d_{ij}\tau_{ij}$ (with $\Sigma_i d_{ij} = 0$ and $\Sigma_j d_{ij} = 0$) has least squares estimator and associated variance equal to

$$\sum_i \sum_j d_{ij}\overline{Y}_{ij.} \text{ and } \sigma^2 \sum_i \sum_j \left(\frac{d_{ij}^2}{r_{ij}}\right).$$

In particular, the least squares estimator of the interaction contrast

$$(\tau_{sh} - \tau_{uh}) - (\tau_{sq} - \tau_{uq})$$

is

$$\overline{Y}_{sh.} - \overline{Y}_{uh.} - \overline{Y}_{sq.} + \overline{Y}_{uq.} \tag{6.4.9}$$

with variance

$$\sigma^2 \left(\frac{1}{r_{sh}} + \frac{1}{r_{uh}} + \frac{1}{r_{sq}} + \frac{1}{r_{uq}}\right). \tag{6.4.10}$$

The least squares estimators of main-effect contrasts $\Sigma c_i \alpha_i^*$ and $\Sigma k_j \beta_j^*$ are

$$\sum_i c_i\hat{\alpha}_i^* = \sum_i c_i \left(\frac{1}{b}\sum_j \overline{Y}_{ij.}\right) \text{ and } \sum_j k_j\hat{\beta}_j^* = \sum_j k_j \left(\frac{1}{a}\sum_i \overline{Y}_{ij.}\right) \tag{6.4.11}$$

with variances

$$\text{Var}(\Sigma c_i \hat{\alpha}_i^*) = \sigma^2 \left(\sum_i \sum_j \frac{c_i^2}{b^2 r_{ij}} \right) \text{ and } \text{Var}(\Sigma k_j \hat{\beta}_j^*) = \sigma^2 \left(\sum_i \sum_j \frac{k_j^2}{a^2 r_{ij}} \right), \qquad (6.4.12)$$

respectively. If the sample sizes are equal, the least squares estimators of $\sum c_i \alpha_i^*$ and $\sum k_j \beta_j^*$ reduce to

$$\sum_i c_i \hat{\alpha}_i^* = \sum_i c_i \overline{Y}_{i..} \text{ and } \sum_j k_j \hat{\beta}_j^* = \sum_j k_j \overline{Y}_{.j.}, \qquad (6.4.13)$$

where $\overline{Y}_{i..} = \sum_j \sum_t Y_{ijt}/br$ and $\overline{Y}_{.j.} = \sum_i \sum_t Y_{ijt}/ar$. Thus, for equal sample sizes,

$$\hat{\alpha}_i^* - \hat{\alpha}_s^* = \overline{Y}_{i..} - \overline{Y}_{s..} \text{ and } \hat{\beta}_j^* - \hat{\beta}_q^* = \overline{Y}_{.j.} - \overline{Y}_{.q.} \qquad (6.4.14)$$

with associated variances $2\sigma^2/(br)$ and $2\sigma^2/(ar)$, respectively.

Example 6.4.1 Battery experiment, continued

The four treatment combinations in the battery experiment of Sect. 2.5.2, p. 24, involved two treatment factors, "duty" and "brand," each having two levels (1 for alkaline and 2 for heavy duty; 1 for name brand and 2 for store brand), giving treatment combinations 11, 12, 21, and 22. There were $r = 4$ observations on each treatment combination. The observed average lifetimes per unit cost for the treatment combinations were

$$\overline{y}_{11.} = 570.75, \quad \overline{y}_{12.} = 860.50, \quad \overline{y}_{21.} = 433.00, \quad \overline{y}_{22.} = 496.25.$$

The interaction contrast

$$\frac{1}{2}(\tau_{11} - \tau_{12} - \tau_{21} + \tau_{22}) = \frac{1}{2}((\alpha\beta)_{11} - (\alpha\beta)_{12} - (\alpha\beta)_{21} + (\alpha\beta)_{22})$$

has least squares estimate

$$\frac{1}{2}(\overline{y}_{11.} - \overline{y}_{12.} - \overline{y}_{21.} + \overline{y}_{22.}) = -113.25,$$

with associated variance

$$\sigma^2 \left(\sum\sum d_{ij}^2/r \right) = \sigma^2 \left((\frac{1}{2})^2 + (-\frac{1}{2})^2 + (-\frac{1}{2})^2 + (\frac{1}{2})^2 \right)/4 = \sigma^2/4.$$

The duty contrast,

$$\alpha_1^* - \alpha_2^* = (\alpha_1 + \overline{(\alpha\beta)}_{1.}) - (\alpha_2 + \overline{(\alpha\beta)}_{2.}) = \frac{1}{2}(\tau_{11} + \tau_{12} - \tau_{21} - \tau_{22}),$$

has least squares estimate $\overline{y}_{1..} - \overline{y}_{2..} = 251.00$ and associated variance $\sigma^2/4$. The brand contrast,

$$\beta_1^* - \beta_2^* = (\beta_1 + \overline{(\alpha\beta)}_{.1}) - (\beta_2 + \overline{(\alpha\beta)}_{.2}) = \frac{1}{2}(\tau_{11} - \tau_{12} + \tau_{21} - \tau_{22}),$$

has least squares estimate $\overline{y}_{.1.} - \overline{y}_{.2.} = -176.50$ and associated variance $\sigma^2/4$.   □

6.4.2 Estimation of σ^2 for the Two-Way Complete Model

Since the two-way complete model (6.2.2) is equivalent to the cell-means model (6.2.1), an unbiased estimate of σ^2 is the same as that for the one-way analysis of variance model, apart from an extra subscript j. Thus, the error sum of squares ssE can be obtained from (3.4.4) or (3.4.5), p. 39, that is,

$$ssE = \sum_i \sum_j \sum_t (y_{ijt} - \bar{y}_{ij.})^2 \tag{6.4.15}$$

$$= \sum_i \sum_j \sum_t y_{ijt}^2 - \sum_i \sum_j r_{ij} \bar{y}_{ij.}^2 \tag{6.4.16}$$

An unbiased estimate for σ^2 is obtained as $msE = ssE/(n - v)$, with $v = ab$. An upper $100(1 - \alpha)\%$ confidence bound for σ^2 is given by (3.4.9), p. 40, that is,

$$\sigma^2 \leq \frac{ssE}{\chi^2_{n-ab,1-\alpha}} . \tag{6.4.17}$$

Example 6.4.2 `Reaction time experiment, continued`

The reaction time pilot experiment, run in 1996 by Liming Cai, Tong Li, Nishant, and Andre van der Kouwe, was described in Exercise 4 of Chap. 4. The experiment was run to compare the speed of response of a human subject to audio and visual stimuli. A personal computer was used to present a "stimulus" to a subject, and the time that the subject took to press a key in response was monitored. The subject was warned that the stimulus was forthcoming by means of an auditory or a visual cue. The two treatment factors were "Cue Stimulus" at two levels, "auditory" and "visual" (Factor A, coded 1, 2), and "Cue Time" at three levels, 5, 10, and 15 seconds between cue and stimulus (Factor B, coded 1, 2, 3), giving a total of $v = 6$ treatment combinations (coded 11, 12, 13, 21, 22, 23). Three observations were taken on each treatment combination for a single subject. The reaction times are shown in Table 6.2. It can be verified that $\sum \sum \sum y_{ijt}^2 = 0.96519$. Using (6.4.16) and the sums in Table 6.2, the sum of squares for error is

$$ssE = \sum_i \sum_j \sum_t y_{ijt}^2 - 3 \sum_i \sum_j \bar{y}_{ij.}^2$$
$$= 0.96519 - 3(0.32057) = 0.00347 ,$$

Table 6.2 Data (in seconds) for the reaction time experiment

A: Cue stimulus	B: Cue time	Treatment combination	Reaction time y_{ijt}			Sums $y_{ij.}$
1	1	11	0.204	0.170	0.181	0.555
1	2	12	0.167	0.182	0.187	0.536
1	3	13	0.202	0.198	0.236	0.636
2	1	21	0.257	0.279	0.269	0.805
2	2	22	0.283	0.235	0.260	0.778
2	3	23	0.256	0.281	0.258	0.795

and an unbiased estimate of σ^2 is $msE = ssE/(18-6) = 0.000289$ seconds^2. An upper 95% confidence bound for σ^2 is

$$\frac{ssE}{\chi^2_{12,.95}} = \frac{0.00347}{5.226} = 0.000664 \text{ seconds}^2 ,$$

and taking square roots, an upper 95% confidence bound for σ is 0.0257 seconds.   □

6.4.3   Multiple Comparisons for the Complete Model

In outlining the analysis at step (g) of the checklist of Chap. 2, the experimenter should specify which treatment contrasts are of interest, together with overall error rates for hypothesis tests and overall confidence levels for confidence intervals. If the two-way complete model has been selected, comparison of treatment combinations, comparison of main effects of A, and comparison of main effects of B may all be of interest. A possibility in outlining the analysis is to select error rates of α_1, α_2, and α_3 for the three sets of inferences. Then, by the Bonferroni method, the *experimentwise* simultaneous error rate is at most $\alpha_1 + \alpha_2 + \alpha_3$, and the experimentwise confidence level is at least $100(1 - \alpha_1 - \alpha_2 - \alpha_3)\%$. If interaction contrasts are also of interest, then the overall α-level can be divided into four parts instead of three.

Comparing Treatment Combinations

When comparison of treatment combinations is of most interest, the cell-means model (6.2.1) is used. The formulae for the Bonferroni, Scheffé, Tukey, and Dunnett methods can all be used in the same way as was done in Chap. 4, but with ssE given by (6.4.16) and with $v = ab$.

The best treatment combination can be found using Tukey's method of multiple comparisons. The best treatment combination may not coincide with the apparent best levels of A and B separately. For example, in Fig. 6.2(d), p. 141, the apparent best treatment combination occurs with presentation format 2 and structure 1, whereas the best presentation format, on average, appears to be number 3.

Comparing Main Effects

Main-effect contrasts compare the effects of the levels of one factor *averaging* over the levels of the other factor and may not be of interest if the two factors interact. If main-effect contrasts are to be examined, then the Bonferroni, Scheffé, Tukey, and Dunnett methods can be used for each factor separately. The general formula is equivalent to (4.4.20), p. 83. For factor A and *equal sample sizes* the formula is

$$\sum_i c_i \bar{\tau}_{i.} = \sum_i c_i \alpha_i^* \in \left(\sum_i c_i \bar{y}_{i..} \pm w \sqrt{msE \sum_i c_i^2 / br} \right) , \tag{6.4.18}$$

where the critical coefficient w for each of the four methods is, respectively,

$$w_B = t_{n-ab,\alpha/2m} \; ; \;\; w_S = \sqrt{(a-1)F_{a-1,n-ab,\alpha}} \; ;$$

$$w_T = q_{a,n-ab,\alpha}/\sqrt{2} \; ; \;\; w_{D2} = |t|^{(0.5)}_{a-1,n-ab,\alpha} \; .$$

The general formula for a confidence interval for a contrast in factor B is

$$\sum_j k_j \bar{\tau}_{.j} = \sum_j k_j \beta_j^* \in \left(\sum_j k_j \bar{y}_{.j.} \pm w \sqrt{msE \sum_j k_j^2 / (ar)} \right) \tag{6.4.19}$$

Fig. 6.5 Average times for the reaction time experiment
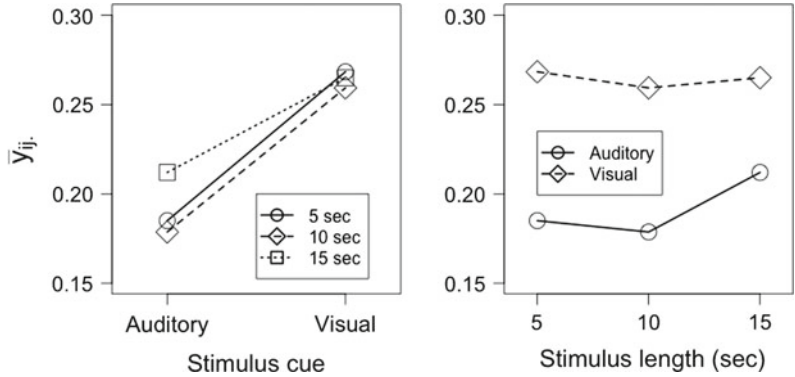
with critical coefficients as above but interchanging a and b. The error variance estimate is $msE = ssE/(n - ab)$, where ssE is obtained from (6.4.16).

For *unequal sample sizes*, the Bonferroni and Scheffé methods can be used, but the least squares estimates and variances must be replaced by (6.4.11) and (6.4.12), respectively. It has not yet been proved that the other two methods retain an overall confidence level of at least $100(1-\alpha)\%$ for unequal sample sizes, although this is widely believed to be the case for Tukey's method.

Example 6.4.3 `Reaction time experiment, continued`

Suppose the preplanned analysis for the reaction time experiment of Example 6.4.2 (p. 151) had been to use the two-way complete model and to test the null hypothesis of no interaction. If the hypothesis were to be rejected, then the plan was to use Tukey's method at level 99% for the pairwise comparisons of the treatment combinations. Otherwise, Tukey's method would be used at level 99% for the pairwise comparison of the levels of B (cue time), and a single 99% confidence interval would be obtained for comparing the two levels of A (cue stimulus). Then the experimentwise confidence level for the three sets of intervals would have been at least 97%.

After looking at the data plotted in Fig. 6.5, the experimenters might decide that comparison of the levels of cue stimulus (averaged over cue time) is actually the only comparison of interest. However, the experimentwise confidence level remains at least 97%, because two other sets of intervals were planned ahead of time and only became uninteresting after the data were examined.

The sample mean weights for the two cue stimuli (averaged over cue times) are

$$\bar{y}_{1..} = 0.1919, \quad \bar{y}_{2..} = 0.2642.$$

The mean square for error was calculated in Example 6.4.2 to be $msE = 0.000289$. The formula for a 99% confidence interval for the comparison of $a = 2$ treatments and $br = 9$ observations on each treatment is obtained from (6.4.18) with $w = w_B = t_{18-6, 0.005} = 3.055$, giving

$$\alpha_2^* - \alpha_1^* \in \left(\bar{y}_{2..} - \bar{y}_{1..} \pm w_B \sqrt{msE\,(1/br + 1/br)} \right)$$
$$= 0.0723 \pm (3.055)\sqrt{0.000289(2/9)} \; = \; (0.0478, 0.0968)\,.$$

Thus, at an experimentwise confidence level of at least 97%, we can conclude that the average reaction time with an auditory cue is between 0.0478 and 0.0968 seconds faster than with a visual cue.   □

Multiple Comparisons When Variances are Unequal

When the variances of the error variables are unequal, and no transformation can be found to remedy the problem, Satterthwaite's approximation, introduced in Sect. 5.6.3 (p. 115), can be used. This is illustrated in Example 6.4.4.

Example 6.4.4 `Bleach experiment`

The bleach experiment was run by Annie Autret in 1986 to study the effect of different bleach concentrations (factor A) and the effect of the type of stain (factor B) on the speed of stain removal from a piece of cloth. The bleach concentration was to be observed at levels 3, 5, and 7 teaspoonfuls of bleach per cup of water (coded 1, 2, 3), and three types of stain (blue ink, jam, tomato sauce; coded 1, 2, 3) were of interest, giving $v = 9$ treatment combinations in total. The experimenter calculated that she needed $r = 5$ observations per treatment combination in order to be able to detect, with probability 0.9, at significance level 0.05, a difference of 5 min in the time of stain removal between the levels of either treatment factor.

The data are shown in Table 6.3 together with the sample mean and standard deviation for each treatment combination. The maximum sample standard deviation is about 8.9 times the size of the minimum sample standard deviation, so the ratio of the maximum to the minimum variance is about 80, and a transformation of the data should be contemplated. The reader can verify, using the technique described in Sect. 5.6.2, that a plot of $\ln(s_{ij}^2)$ against $\ln(\bar{y}_{ij.})$ is not linear, so no transformation of the form $h(y_{ijt}) = y_{ijt}^{1-(q/2)}$ will adequately equalize the error variances.

An alternative is to apply Satterthwaite's approximation (Sect. 5.6.3, p. 115). The plan of the analysis was to use Tukey's method with an error rate of 0.01 for each of the main-effect comparisons and for the pairwise differences of the treatment combinations, giving an experimentwise confidence level of at least 97%. For the main effect of B, for example, a pairwise comparison of levels u and h of factor B is of the form

$$\beta_u^* - \beta_h^* = \overline{\tau}_{.u} - \overline{\tau}_{.h} = \frac{1}{3}\left(\tau_{1u} + \tau_{2u} + \tau_{3u} - \tau_{1h} - \tau_{2h} - \tau_{3h}\right),$$

which has least squares estimate

$$\widehat{\beta_u^*} - \widehat{\beta_h^*} = \bar{y}_{.u.} - \bar{y}_{.h.} = \frac{1}{3}\left(\bar{y}_{1u.} + \bar{y}_{2u.} + \bar{y}_{3u.} - \bar{y}_{1h.} - \bar{y}_{2h.} - \bar{y}_{3h.}\right).$$

Table 6.3   Data for the bleach experiment, with treatment factors "concentration" (*A*) and "stain type" (*B*)

ij	Time for stain removal (in seconds)					$\bar{y}_{ij.}$	s_{ij}
11	3600	3920	3340	3173	2452	3297.0	550.27
12	495	236	515	573	555	474.8	137.04
13	733	525	793	1026	510	717.4	212.85
21	2029	2271	2156	2493	2805	2350.8	305.94
22	428	432	335	288	376	371.8	61.60
23	880	759	1138	780	1625	1036.4	361.91
31	3660	4105	4545	3569	3342	3844.2	479.85
32	410	225	437	350	140	312.4	126.32
33	539	1354	347	584	781	721.0	386.02

If s_{ij}^2 denotes the sample variance of the data for treatment combination ij, the estimated variance of this estimator, as in (5.6.4), p. 115, is

$$\widehat{\text{Var}}\left(\widehat{\beta_u^*} - \widehat{\beta_h^*}\right) = \sum_i \sum_j d_{ij}^2 \frac{s_{ij}^2}{r_{ij}} = \frac{1}{9 \times 5}(s_{1u}^2 + s_{2u}^2 + s_{3u}^2 + s_{1h}^2 + s_{2h}^2 + s_{3h}^2),$$

and since $r = 5$, the approximate number of degrees of freedom for error is

$$df = \frac{(s_{1u}^2 + s_{2u}^2 + s_{3u}^2 + s_{1h}^2 + s_{2h}^2 + s_{3h}^2)^2}{(s_{1u}^4/4) + (s_{2u}^4/4) + (s_{3u}^4/4) + (s_{1h}^4/4) + (s_{2h}^4/4) + (s_{3h}^4/4)}$$

after canceling the factor $r^2 = 25$ in the numerator and denominator.

For Tukey's method of pairwise comparisons for factor B with $b = 3$ levels, the minimum significant difference is

$$msd = w_T \sqrt{\widehat{\text{Var}}\left(\widehat{\beta_u^*} - \widehat{\beta_h^*}\right)},$$

with $w_T = q_{3,df,.01}/\sqrt{2}$. For measurements in seconds, we have the following values:

(u, h)	df	$q_{3,df,0.01}$	$\widehat{\text{Var}}\left(\hat{\beta}_u^* - \hat{\beta}_h^*\right)$	msd	$\bar{y}_{.u.} - \bar{y}_{.h.}$
$(1, 2)$	11.5	5.09	14,780.6	437.57	2,777.67
$(1, 3)$	18.6	4.68	21,153.5	481.31	2,339.07
$(3, 2)$	12.6	4.99	8,084.7	317.26	438.60

The set of 99% simultaneous Tukey confidence intervals for pairwise differences is then

$$\beta_1^* - \beta_2^* \in (2777.67 \pm 437.57) = (2340.10, 3215.24),$$

$$\beta_1^* - \beta_3^* \in (1857.76, 2820.38), \qquad \beta_3^* - \beta_2^* \in (121.34, 755.86).$$

Since none of the intervals contains zero, we can state that all pairs of levels of B (stain types) have different effects on the speed of stain removal, averaged over the three concentrations of bleach. With experimentwise confidence level at least 97%, the mean time to remove blue ink (level 1) is between 1857 and 2820 seconds longer than that for tomato sauce (level 3), and the mean time to remove tomato sauce is between 121 and 755 seconds longer than that for jam (level 2).   □

6.4.4   Analysis of Variance for the Complete Model

There are three standard hypotheses that are usually examined when the two-way complete model is used. The first hypothesis is that the interaction between treatment factors A and B is negligible; that is,

$$H_0^{AB} : \{(\alpha\beta)_{ij} - (\alpha\beta)_{iq} - (\alpha\beta)_{sj} + (\alpha\beta)_{sq} = 0 \text{ for all } i \neq s, j \neq q\},$$

which occurs when the interaction plots show parallel lines. Notice that if all of the contrasts $(\alpha\beta)_{ij} - (\alpha\beta)_{iq} - (\alpha\beta)_{sj} + (\alpha\beta)_{sq}$ are zero, then their averages over s and q are also zero. This leads to an equivalent way to write H_0^{AB} as

$$H_0^{AB} : \{(\alpha\beta)_{ij} - \overline{(\alpha\beta)}_{i.} - \overline{(\alpha\beta)}_{.j} + \overline{(\alpha\beta)}_{..} = 0 \ \text{ for all } ij\}\,.$$

In this form, it appears that H_0^{AB} is based on ab estimable contrasts, but in fact, some of them are redundant, since the ab contrasts add to zero over the subscript $i = 1, 2, \ldots, a$ and also over the subscript $j = 1, 2, \ldots, b$. Consequently, H_0^{AB} is actually based on $(a-1)(b-1)$ estimable contrasts, and the test is based on $(a-1)(b-1)$ degrees of freedom.

The other two standard hypotheses are the main-effect hypotheses

$$H_0^{A} : \{\alpha_1^* = \alpha_2^* = \ldots = \alpha_a^*\} \ \text{ and } \ H_0^{B} : \{\beta_1^* = \beta_2^* = \ldots = \beta_b^*\}\,,$$

where $\alpha_i^* = \alpha_i + \overline{(\alpha\beta)}_{i.}$ and $\beta_j^* = \beta_j + \overline{(\alpha\beta)}_{.j}$. However, these main-effect hypotheses may not be of interest if there is a sizable interaction. Each of the main-effect hypotheses can be rephrased in terms of estimable contrasts in the parameters, and so can be tested. As in Chap. 3, the tests will be based on $(a-1)$ and $(b-1)$ degrees of freedom, respectively.

When the sample sizes are unequal, there are no neat algebraic formulae for the decision rules of the hypothesis tests. Therefore, we will obtain the tests for equal sample sizes and postpone discussion of the unequal sample size case to Sects. 6.8 and 6.9, where analysis will be done by computer.

Testing Interactions—Equal Sample Sizes

Since tests for main effects may not be relevant if the two factors interact, the hypothesis of negligible interaction should be tested first. As in Sect. 3.5.1, p. 41, in order to test

$$H_0^{AB} : \{(\alpha\beta)_{ij} - \overline{(\alpha\beta)}_{i.} - \overline{(\alpha\beta)}_{.j} + \overline{(\alpha\beta)}_{..} = 0 \ \text{ for all } ij\}$$

against the alternative hypothesis H_A^{AB}:{the interaction is not negligible}, we compare the sum of squares for error ssE under the two-way complete model (6.2.2) with the sum of squares for error ssE_0^{AB} under the reduced model obtained when H_0^{AB} is true. The difference

$$ssAB = ssE_0^{AB} - ssE$$

is called the *sum of squares for the interaction AB*, and the test rejects H_0^{AB} in favor of H_A^{AB} if $ssAB$ is large relative to ssE.

We can rewrite the two-way complete model as

$$
\begin{aligned}
y_{ijt} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijt} \\
&= \mu^* + \alpha_i^* + \beta_j^* + [(\alpha\beta)_{ij} - \overline{(\alpha\beta)}_{i.} - \overline{(\alpha\beta)}_{.j} + \overline{(\alpha\beta)}_{..}] + \epsilon_{ijt}\,,
\end{aligned}
$$

where μ^* is the constant $\mu - \overline{(\alpha\beta)}_{..}$. So, when H_0^{AB} is true, the reduced model is

$$y_{ijt} = \mu^* + \alpha_i^* + \beta_j^* + \epsilon_{ijt}\,,$$

which has the same form as the two-way main-effects model.

We will show in Sect. 6.5.1 that the least squares estimate of $\mu + \alpha_i + \beta_j$ for the two-way main-effects model is $\bar{y}_{i..} + \bar{y}_{.j.} - \bar{y}_{...}$, for equal sample sizes. Similarly, the least squares estimate of $\mu^* + \alpha_i^* + \beta_j^*$ in the above reduced model is also $\bar{y}_{i..} + \bar{y}_{.j.} - \bar{y}_{...}$. Hence, the sum of squares for error for the reduced model is

$$ssE_0^{AB} = \sum_i \sum_j \sum_t \left(y_{ijt} - \hat{\mu}^* - \hat{\alpha}_i^* - \hat{\beta}_j^* \right)^2$$

$$= \sum_i \sum_j \sum_t (y_{ijt} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 .$$

Adding and subtracting a term $\bar{y}_{ij.}$ to this expression, we have

$$ssE_0^{AB} = \sum_i \sum_j \sum_t \left((y_{ijt} - \bar{y}_{ij.}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) \right)^2$$

$$= \sum_i \sum_j \sum_t (y_{ijt} - \bar{y}_{ij.})^2 + \sum_i \sum_j \sum_t (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 .$$

But the first term is just *ssE* given in (6.4.15). So, for equal sample sizes,

$$\begin{aligned} ssAB &= ssE_0^{AB} - ssE \\ &= r \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \\ &= r \sum_i \sum_j \bar{y}_{ij.}^2 - br \sum_i \bar{y}_{i..}^2 - ar \sum_j \bar{y}_{.j.}^2 + abr\bar{y}_{...}^2 . \end{aligned} \qquad (6.4.20)$$

It can be shown that when H_0^{AB} is true, the corresponding random variable $SS(AB)/\sigma^2$ has a chi-squared distribution with $(a-1)(b-1)$ degrees of freedom. Also, $SSE/\sigma^2 \sim \chi^2_{n-ab}$ and SSE can be shown to be independent of $SS(AB)$. So, when H_0^{AB} is true,

$$\frac{SS(AB)/(a-1)(b-1)\sigma^2}{SSE/(n-ab)\sigma^2} = \frac{MS(AB)}{MSE} \sim F_{(a-1)(b-1),n-ab} .$$

We reject H_0^{AB} for large values of the ratio $msAB/msE$. Thus, the rule for testing the hypothesis H_0^{AB} against the alternative hypothesis that the interaction is not negligible is

$$\text{reject } H_0^{AB} \text{ if } \frac{msAB}{msE} > F_{(a-1)(b-1),n-ab,\alpha} , \qquad (6.4.21)$$

where $msAB = ssAB/(a-1)(b-1)$, $msE = ssE/(n-ab)$, $ssAB$ is given in (6.4.20), and ssE is

$$ssE = \sum_i \sum_j \sum_t y_{ijt}^2 - \sum_i \sum_j r\bar{y}_{ij.}^2 .$$

If H_0^{AB} is rejected, it is often preferable to use the equivalent cell-means model and look at contrasts in the treatment combinations. If H_0^{AB} is not rejected, then tests and contrasts for main effects are usually of interest, and the two-way complete model is retained. (We do not change to the inequivalent main-effects model.)

Testing Main Effects of *A*—Equal Sample Sizes

In testing the hypothesis that factor *A* has no effect on the response, one can either test the hypothesis that the levels of *A* (averaged over the levels of *B*) have the same average effect on the response, that is,

$$H_0^A : \{\alpha_1^* = \alpha_2^* = \cdots = \alpha_a^*\},$$

or one can test the hypothesis that the response depends only on the level of B, that is

$$H_0^{A+AB} : \{H_0^A \text{ and } H_0^{AB} \text{are both true}\}.$$

The traditional test, which is produced automatically by many computer packages, is a test of the former, and the sum of squares for error ssE under the two-way complete model is compared with the sum of squares for error ssE_0^A under the reduced model

$$Y_{ijt} = \mu^{**} + \beta_j^* + \left((\alpha\beta)_{ij} - \overline{(\alpha\beta)}_{i.} - \overline{(\alpha\beta)}_{.j} + \overline{(\alpha\beta)}_{..}\right) + \epsilon_{ijt}.$$

It is, perhaps, more intuitively appealing to test H_0^{A+AB} rather than H_0^A, since the corresponding reduced model is

$$Y_{ijt} = \mu^* + \beta_j^* + \epsilon_{ijt},$$

suggesting that A has no effect on the response whatsoever.

In this book, we take the view that the main effect of A would not be tested unless the hypothesis of no interaction were first accepted. If it is true that there is no interaction, then the two hypotheses and corresponding reduced models are the same, and the results of the two tests should be similar. Consequently, we will derive the test of the standard hypothesis H_0^A.

It can be shown that if the sample sizes are equal, the least squares estimate of $E[Y_{ijt}]$ for the reduced model under H_0^A is

$$\bar{y}_{ij.} - \bar{y}_{i..} + \bar{y}_{...},$$

and so the sum of squares for error for the reduced model is

$$ssE_0^A = \sum_i \sum_j \sum_t (y_{ijt} - \bar{y}_{ij.} + \bar{y}_{i..} - \bar{y}_{...})^2.$$

Taking the terms in pairs and expanding the terms in parentheses, we obtain

$$ssE_0^A = \sum_{i=1}^a \sum_{j=1}^b \sum_{t=1}^r (y_{ijt} - \bar{y}_{ij.})^2 - br \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2.$$

Since the first term is the formula (6.4.15) for ssE, the *sum of squares for treatment factor A* is

$$ssA = ssE_0^A - ssE = br \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2 = br \sum_{i=1}^a \bar{y}_{i..}^2 - abr\bar{y}_{...}^2. \qquad (6.4.22)$$

Notice that this formula for ssA is similar to the formula (3.5.12), p. 43, for ssT used to test the hypothesis $H_0:\{\tau_1 = \tau_2 = \cdots = \tau_a\}$ in the one-way analysis of variance.

We write SSA for the random variable corresponding to ssA. It can be shown that if H_0^A is true, SSA/σ^2 has a chi-squared distribution with $a - 1$ degrees of freedom, and that SSA and SSE are independent. So, writing $MSA = SSA/(a - 1)$, we have that MSA/MSE has an F-distribution when

Table 6.4 Two-way ANOVA, crossed fixed effects with interaction

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Ratio
Factor A	$a-1$	ssA	$\frac{ssA}{a-1}$	$\frac{msA}{msE}$
Factor B	$b-1$	ssB	$\frac{ssB}{b-1}$	$\frac{msB}{msE}$
AB	$(a-1)(b-1)$	$ssAB$	$\frac{ssAB}{(a-1)(b-1)}$	$\frac{msAB}{msE}$
Error	$n-ab$	ssE	$\frac{ssE}{n-ab}$	
Total	$n-1$	$sstot$		

Computational formulae for equal sample sizes	
$ssE = \sum_i \sum_j \sum_t y_{ijt}^2$ $\quad - r \sum_i \sum_j \bar{y}_{ij.}^2$	$ssA = br \sum_i \bar{y}_{i..}^2 - n\bar{y}_{...}^2$ $ssB = ar \sum_j \bar{y}_{.j.}^2 - n\bar{y}_{...}^2$
$sstot = \sum_i \sum_j \sum_t y_{ijt}^2 - n\bar{y}_{...}^2$	$ssAB = r \sum_i \sum_j \bar{y}_{ij.}^2 - br \sum_i \bar{y}_{i..}^2$
$n = abr$	$\quad - ar \sum_j \bar{y}_{.j.}^2 + n\bar{y}_{...}^2$

H_0^A is true, and the rule for testing $H_0^A : \{\alpha_1^* = \cdots = \alpha_a^*\}$ against H_A^A : {not all of the α_i^*'s are equal} is

$$\text{reject } H_0^A \text{ if } \frac{msA}{msE} > F_{a-1,n-ab,\alpha}, \tag{6.4.23}$$

where $msA = ssA/(a-1)$ and $msE = ssE/(n-ab)$.

Testing Main Effects of *B*—Equal Sample Sizes

Analogous to the test for main effects of A, we can show that the rule for testing $H_0^B : \{\beta_1^* = \beta_2^* = \cdots = \beta_b^*\}$ against H_A^B : {not all of the β_j^*'s are equal} is

$$\text{reject } H_0^B \text{ if } \frac{msB}{msE} > F_{b-1,n-ab,\alpha}, \tag{6.4.24}$$

where $msB = ssB/(b-1)$, $msE = ssE/(n-ab)$, and

$$ssB = ar \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2 = ar \sum_j \bar{y}_{.j.}^2 - abr\bar{y}_{...}^2. \tag{6.4.25}$$

Analysis of Variance Table

The tests of the three hypotheses are summarized in a two-way analysis of variance table, shown in Table 6.4. The computational formulae are given for equal sample sizes. The last line of the table is $sstot = \sum_i \sum_j \sum_t (y_{ijt} - \bar{y}_{...})^2$, which is the total sum of squares similar to (3.5.16). It can be verified that

$$ssA + ssB + ssAB + ssE = sstot.$$

When the sample sizes are not equal, the formulae for ssA, ssB, and $ssAB$ are more complicated, the corresponding random variables SSA, SSB, and $SS(AB)$ are not independent, and

$$ssA + ssB + ssAB + ssE \neq sstot.$$

The analysis of experiments with unequal sample sizes will be discussed in Sects. 6.8 and 6.9 using the software packages SAS and R, respectively.

Example 6.4.5 `Reaction time experiment, continued`

The reaction time experiment was described in Example 6.4.2, p. 151. There were $a = 2$ levels of cue stimulus and $b = 3$ levels of cue time, and $r = 3$ observations per treatment combination. Using the data in Table 6.2, we have

$$sstot = \sum_i \sum_j \sum_t y_{ijt}^2 - abr\bar{y}_{...}^2 = 0.96519 - 0.93617 = 0.02902,$$

$$ssA = br \sum_i \bar{y}_{i..}^2 - abr\bar{y}_{...}^2 = 9(0.1918^2 + 0.2642^2) - 0.93617 = 0.02354,$$

$$ssB = ar \sum_j \bar{y}_{.j.}^2 - abr\bar{y}_{...}^2 = 6(0.2267^2 + 0.2190^2 + 0.2385^2) - 0.93617 = 0.00116$$

$$ssAB = r \sum_i \sum_j \bar{y}_{ij.}^2 - br \sum_i \bar{y}_{i..}^2 - ar \sum_j \bar{y}_{.j.}^2 + abr\bar{y}_{...}^2$$
$$= 0.96172 - 0.95971 - 0.93733 + 0.93617 = 0.00085,$$

and in Example 6.4.2, ssE was calculated to be 0.00347. It can be seen that $sstot = ssA + ssB + ssAB + ssE$. The analysis of variance table is shown in Table 6.5. The mean squares are the sums of squares divided by their degrees of freedom.

There are three hypotheses to be tested. If the Type I error probability α is selected to be 0.01 for each test, then the probability of incorrectly rejecting at least one hypothesis when it is true is at most 0.03. The interaction plots in Fig. 6.5, p. 153, suggest that there is no interaction between cue stimulus (A) and cue time (B). To test this hypothesis, we obtain from the analysis of variance table

$$msAB/msE = 0.00043/0.00029 = 1.46.$$

which is less than $F_{2,12,.01} = 6.93$. Therefore, at individual significance level $\alpha = 0.01$, there is not sufficient evidence to reject the null hypothesis H_0^{AB} that the interaction is negligible. This agrees with the interaction plot.

Now consider the main effects. Looking at Fig. 6.5, if we average over cue stimulus, there does not appear to be much difference in the effect of cue time. If we average over cue time, then auditory cue stimulus (level 1) appears to produce a shorter reaction time than a visual cue stimulus (level 2). From the analysis of variance table, $msA/msE = 0.02354/0.00029 = 81.38$. This is larger than $F_{1,12,.01} = 9.33$, so we reject $H_0^A:\{\alpha_1^* = \alpha_2^*\}$, and we would conclude that there is a difference in cue stimulus averaged over the cue times. On the other hand, $msB/msE = 0.00058/0.00029 = 2.0$, which

Table 6.5 Two-way ANOVA for the reaction time experiment

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Ratio	*p*-value
Cue stimulus	1	0.02354	0.02354	81.38	0.0001
Cue time	2	0.00116	0.00058	2.00	0.1778
Interaction	2	0.00085	0.00043	1.46	0.2701
Error	12	0.00347	0.00029		
Total	17	0.02902			

is less than $F_{2,12,.01} = 6.93$. Consequently, we do not reject $H_0^B : \{\beta_1^* = \beta_2^* = \beta_3^*\}$ and conclude that there is no evidence for a difference in the effects of the cue times averaged over the two cue stimuli.

If the analysis were done by a computer program, the p-values in Table 6.5 would be printed. We would reject any hypothesis whose corresponding p-value is less than the selected individual α^* level. In this example, we selected $\alpha^* = 0.01$, and we would fail to reject H_0^{AB} and H_0^B, but we would reject H_0^A, as in the hand calculations.

This was a pilot experiment, and since the experimenters already believed that cue stimulus and cue time really do not interact, they selected the two-way main-effects model in planning the main experiment.   □

6.5   Analysis of the Two-Way Main-Effects Model

6.5.1   Least Squares Estimators for the Main-Effects Model

The two-way main-effects model (6.2.3) is

$$Y_{ijt} = \mu + \alpha_i + \beta_j + \epsilon_{ijt} \,,$$
$$\epsilon_{ijt} \sim N(0, \sigma^2) \,,$$
$$\epsilon_{ijt}\text{'s are mutually independent} \,,$$
$$t = 1, \ldots, r_{ij}; \quad i = 1, \ldots, a; \quad j = 1, \ldots, b.$$

This model is a submodel of the two-way complete model (6.2.2) in the sense that it can only describe situations similar to those depicted in plots (a)–(d) of Fig. 6.1 and cannot describe plots (a)–(d) of Fig. 6.2. When the sample sizes are unequal, the least squares estimators of the parameters in the main-effects model are not easy to obtain, and calculations are best left to a computer (see Sects. 6.8 and 6.9). In the optional subsection below, we show that when the sample sizes are all equal to r, the least squares estimator of $E[Y_{ijt}] = \mu + \alpha_i + \beta_j$ is

$$\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \overline{Y}_{i..} + \overline{Y}_{.j.} - \overline{Y}_{...} \,. \tag{6.5.26}$$

The least squares estimator for the estimable main-effect contrast $\sum_i c_i \alpha_i$ with $\sum_i c_i = 0$ is then

$$\sum_i c_i \hat{\alpha}_i = \sum_i c_i (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) \;\; = \;\; \sum_i c_i \left(\overline{Y}_{i..} + \overline{Y}_{.j.} - \overline{Y}_{...} \right)$$
$$= \sum_i c_i \overline{Y}_{i..} \,,$$

which has variance

$$\text{Var}\left(\sum_i c_i \hat{\alpha}_i \right) = \text{Var}\left(\sum_i c_i \overline{Y}_{i..} \right) = \frac{\sigma^2}{br} \sum_i c_i^2 \,. \tag{6.5.27}$$

For example, $\alpha_p - \alpha_s$, the pairwise comparison of levels p and s of A, has least squares estimator and associated variance

$$\hat{\alpha}_p - \hat{\alpha}_s = \overline{Y}_{p..} - \overline{Y}_{s..} \quad \text{with} \quad \text{Var}(\overline{Y}_{p..} - \overline{Y}_{s..}) = \frac{2\sigma^2}{br} \,.$$

These are exactly the same formulas as for the two-way complete model and similar to those for the one-way model. Likewise for B, a main-effect contrast $\sum k_j \beta_j$ with $\sum_j k_j = 0$ has least squares estimator and associated variance

$$\sum_j k_j \hat{\beta}_j = \sum_j k_j \overline{Y}_{.j.} \quad \text{and} \quad \text{Var}\left(\sum_j k_j \overline{Y}_{.j.}\right) = \frac{\sigma^2}{ar} \sum_j k_j^2, \quad (6.5.28)$$

and the least squares estimator and associated variance for the pairwise difference $\beta_h - \beta_q$ is

$$\hat{\beta}_h - \hat{\beta}_q = \overline{Y}_{.h.} - \overline{Y}_{.q.} \quad \text{with} \quad \text{Var}(\overline{Y}_{.h.} - \overline{Y}_{.q.}) = \frac{2\sigma^2}{ar}.$$

Example 6.5.1 `Nail varnish experiment`

An experiment on the efficacy of nail varnish solvent in removing nail varnish from cloth was run by Pascale Quester in 1986. Two different brands of solvent (factor A) and three different brands of nail varnish (factor B) were investigated. One drop of nail varnish was applied to a piece of cloth (dropped from the applicator 20 cm above the cloth). The cloth was immersed in a bowl of solvent and the time measured (in minutes) until the varnish completely dissolved. There were six treatment combinations 11, 12, 13, 21, 22, 23, where the first digit represents the brand of solvent and the second digit represents the brand of nail varnish used in the experiment. The design was a completely randomized design with $r = 5$ observations on each of the six treatment combinations. The data are listed in Table 6.6 in the order in which they were collected.

The experimenter had run a pilot experiment to estimate the error variance σ^2 and to check that the experimental procedure was satisfactory. The pilot experiment indicated that the interaction between nail varnish and solvent was negligible. The similarity of the chemical composition of the varnishes and solvents, and the verification from the pilot experiment, suggest that the main-effects model (6.2.3) will be a satisfactory model for the main experiment. The data from the main experiment give the interaction plots in Fig. 6.6. Although the lines are not quite parallel, the selected main-effects model would not be a severely incorrect representation of the data.

Using the data in Table 6.6, the average dissolving time (in minutes) for the two brands of solvent are

Table 6.6 Data (minutes) for the nail varnish experiment

Solvent	2	1	1	2	2	2	1	2
Varnish	3	3	3	3	2	2	2	2
Time	32.50	30.20	27.25	24.25	34.42	26.00	22.50	31.08
Solvent	1	2	1	1	2	1	2	2
Varnish	2	1	1	1	1	3	3	2
Time	25.17	29.17	27.58	28.75	31.75	29.75	30.75	29.17
Solvent	1	1	2	1	2	2	1	2
Varnish	2	1	2	2	1	3	3	1
Time	27.75	25.83	24.75	21.50	32.08	29.50	24.50	28.50
Solvent	2	1	1	2	1	1		
Varnish	3	3	1	1	1	2		
Time	28.75	22.75	29.25	31.25	22.08	25.00		

Fig. 6.6 Average dissolving times for the nail varnish experiment
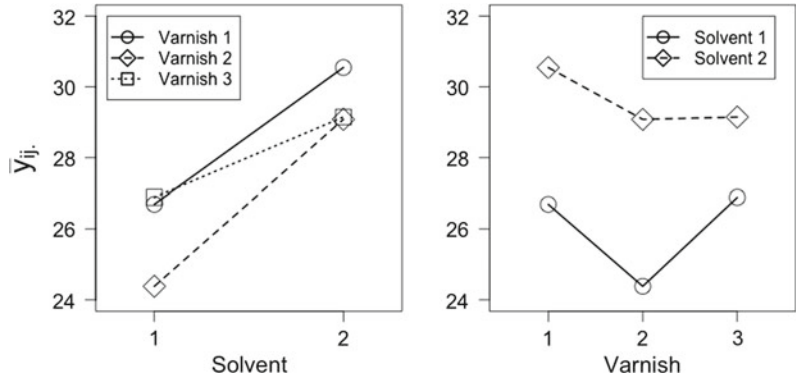
$$\bar{y}_{1..} = 25.9907 \quad \text{and} \quad \bar{y}_{2..} = 29.5947 \, .$$

So the least squares estimate of the difference in the dissolving times for the two solvents is

$$\hat{\alpha}_1 - \hat{\alpha}_2 = \bar{y}_{1..} - \bar{y}_{2..} = -3.6040 \, ,$$

and the variance of the estimator is $2\sigma^2/(rb) = 2\sigma^2/15$. A difference of 3.6 minutes seems quite substantial, but this needs to be compared with the experimental error via a confidence interval to see whether such a difference could have occurred by chance (see Examples 6.5.2 and 6.5.3).

The average dissolving times for the three brands of nail varnish are

$$\bar{y}_{.1.} = 28.624, \quad \bar{y}_{.2.} = 26.734, \text{ and } \bar{y}_{.3.} = 28.020 \, ,$$

and the least squares estimates of the pairwise comparisons are

$$\hat{\beta}_1 - \hat{\beta}_2 = 1.890 \, , \quad \hat{\beta}_1 - \hat{\beta}_3 = 0.604 \, , \text{ and } \hat{\beta}_2 - \hat{\beta}_3 = -1.286 \, ,$$

each with associated variance $2\sigma^2/10$. Since levels 1 and 2 of the nail varnish represented French brands, while level 3 represented an American brand, the difference of averages contrast

$$\frac{1}{2}(\beta_1 + \beta_2) - \beta_3$$

would also be of interest. The least squares estimate of this contrast is

$$\frac{1}{2}(\hat{\beta}_1 + \hat{\beta}_2) - \hat{\beta}_3 \;=\; \frac{1}{2}(\bar{y}_{.1.} + \bar{y}_{.2.}) - \bar{y}_{.3.} \;=\; -0.341 \, ,$$

with associated variance $6\sigma^2/40$.   □

Deriving Least Squares Estimators for Equal Sample Sizes (Optional)

We now sketch the derivation (using calculus) of the least squares estimators for the parameters of the two-way main-effects model (6.2.3), when the sample sizes are all equal to r. A reader without knowledge of calculus may jump to Sect. 6.5.2, p. 165.

As in Sect. 3.4.3, the least squares estimates of the parameters in a model are those estimates that give the minimum value of the sum of squares of the estimated errors. For the two-way main-effects model (6.2.3), the sum of squared errors is

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{t=1}^{r} e_{ijt}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{t=1}^{r} \left(y_{ijt} - (\mu + \alpha_i + \beta_j)\right)^2 .$$

The least squares estimates are obtained by differentiating the sum of squared errors with respect to each of the parameters μ, α_i $(i = 1, \ldots, a)$, and β_j $(j = 1, \ldots, b)$ in turn and setting the derivatives equal to zero. The resulting set of normal equations is as follows.

$$y_{...} - abr\hat{\mu} - br\sum_i \hat{\alpha}_i - ar\sum_j \hat{\beta}_j = 0, \tag{6.5.29}$$

$$y_{i..} - br\hat{\mu} - br\hat{\alpha}_i - r\sum_j \hat{\beta}_j = 0, \quad i = 1, \ldots, a, \tag{6.5.30}$$

$$y_{.j.} - ar\hat{\mu} - r\sum_i \hat{\alpha}_i - ar\hat{\beta}_j = 0, \quad j = 1, \ldots, b. \tag{6.5.31}$$

There are $1 + a + b$ normal equations in $1 + a + b$ unknowns. However, the equations are not all distinct (linearly independent), since the sum of the a equations listed in (6.5.30) is equal to the sum of the b equations listed in (6.5.31), which is equal to (6.5.29). Consequently, there are at most, and, in fact, exactly, $1 + a + b - 2$ distinct equations, and two extra equations are needed in order to obtain a solution. Many computer packages, including the SAS software, use the extra equations $\hat{\alpha}_a = 0$ and $\hat{\beta}_b = 0$, while the R package uses the extra equations $\hat{\alpha}_1 = 0$ and $\hat{\beta}_1 = 0$. However, when working by hand, it is easier to use the equations $\sum_i \hat{\alpha}_i = 0$ and $\sum_j \hat{\beta}_j = 0$, in which case (6.5.29)–(6.5.31) give the following least squares solutions:

$$\hat{\mu} = \bar{y}_{...}, \tag{6.5.32}$$
$$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \quad i = 1, \ldots, a,$$
$$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}, \quad j = 1, \ldots, b.$$

Then the least squares estimate of $\mu + \alpha_i + \beta_j$ is

$$\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{y}_{i..} + \bar{y}_{.j.} - \bar{y}_{...}, \quad i = 1, \ldots, a, \quad j = 1, \ldots, b.$$

Deriving Least Squares Estimators for Unequal Sample Sizes (Optional)

If the sample sizes are not equal, then the normal equations for the two-way main-effects model become

$$y_{...} - n\hat{\mu} - \sum_{p=1}^{a} r_{p.}\hat{\alpha}_p - \sum_{q=1}^{b} r_{.q}\hat{\beta}_q = 0, \tag{6.5.33}$$

$$y_{i..} - r_{i.}\hat{\mu} - r_{i.}\hat{\alpha}_i - \sum_{q=1}^{b} r_{iq}\hat{\beta}_q = 0, \quad i = 1, \ldots, a, \tag{6.5.34}$$

$$y_{.j.} - r_{.j}\hat{\mu} - \sum_{p=1}^{a} r_{pj}\hat{\alpha}_p - r_{.j}\hat{\beta}_j = 0, \quad j = 1, \ldots, b, \tag{6.5.35}$$

where $n = \sum_i \sum_j r_{ij}$, $r_{p.} = \sum_j r_{pj}$, and $r_{.q} = \sum_i r_{iq}$. As in the equal sample size case, the normal equations represent $a + b - 1$ distinct equations in $1 + a + b$ unknowns, and two extra equations are needed to obtain a particular solution. Looking at (6.5.33), a sensible choice might be $\sum_p r_{p.}\hat{\alpha}_p = 0$ and $\sum_q r_{.q}\hat{\beta}_q = 0$. Then $\hat{\mu} = \bar{y}_{...}$ as in the equal sample size case. However, obtaining solutions for the $\hat{\alpha}_i$'s and $\hat{\beta}_j$'s is not so easy. One can solve for $\hat{\beta}_j$ in (6.5.35) and substitute this into (6.5.34), which gives the following equations in the $\hat{\alpha}_i$'s:

$$\hat{\alpha}_i - \sum_{q=1}^{b} \frac{r_{iq}}{r_{.q} r_{i.}} \sum_{p=1}^{a} r_{pq}\hat{\alpha}_p = \bar{y}_{i..} - \sum_{q=1}^{b} \frac{r_{iq}}{r_{.q} r_{i.}} y_{.q}, \quad \text{for } i = 1, \ldots, a. \tag{6.5.36}$$

Equations in the $\hat{\beta}_j$'s can be obtained similarly. Algebraic expressions for the individual parameter estimates are generally complicated, and we will leave the unequal sample size case to a computer analysis (Sects. 6.8 and 6.9).

6.5.2   Estimation of σ^2 in the Main-Effects Model

The minimum value of the sum of squares of the estimated errors for the two-way main-effects model is

$$\begin{aligned}
ssE &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{t=1}^{r}(y_{ijt} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 \tag{6.5.37} \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{t=1}^{r}(y_{ijt} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \,.
\end{aligned}$$

Expanding the terms in parentheses in (6.5.37) yields the following formula useful for direct hand calculation of ssE:

$$ssE = \sum_i\sum_j\sum_t y_{ijt}^2 - br\sum_i \bar{y}_{i..}^2 - ar\sum_j \bar{y}_{.j.}^2 + abr\bar{y}_{...}^2 \tag{6.5.38}$$

Now, ssE is the observed value of

$$SSE = \sum_i\sum_j\sum_t (Y_{ijt} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 \,.$$

In Exercise 19, the reader will be asked to prove, for the equal sample size case, that

$$E[SSE] = (n - a - b + 1)\sigma^2 \,,$$

where $n = abr$, so an unbiased estimator of σ^2 is

$$MSE = SSE/(n - a - b + 1) \, .$$

It can be shown that SSE/σ^2 has a chi-squared distribution with $(n - a - b + 1)$ degrees of freedom. An upper $100(1 - \alpha)\%$ confidence bound for σ^2 is therefore given by

$$\sigma^2 \leq \frac{ssE}{\chi^2_{n-a-b+1,1-\alpha}} \, .$$

Example 6.5.2 `Nail varnish experiment, continued`

The data for the nail varnish experiment are given in Table 6.6 of Example 6.5.1, p. 162, and $a = 2$, $b = 3$, $r = 5$, $n = 30$. It can be verified that

$$\sum_i \sum_j \sum_t y_{ijt}^2 = 23,505.7976, \quad \bar{y}_{...} = 27.7927 \, ,$$

and

$$\bar{y}_{1..} = 25.9907, \qquad \bar{y}_{2..} = 29.5947,$$

$$\bar{y}_{.1.} = 28.624, \qquad \bar{y}_{.2.} = 26.734, \qquad \bar{y}_{.3.} = 28.020.$$

Thus, from (6.5.38),

$$ssE = 23,505.7976 - 23,270.3857 - 23,191.6053 + 23,172.9696$$
$$= 216.7762 \, ,$$

and an unbiased estimate of σ^2 is

$$msE = 216.7762/(30 - 2 - 3 + 1) = 8.3375 \text{ minutes}^2 \, .$$

A 95% upper confidence bound for σ^2 is

$$\frac{ssE}{\chi^2_{26,.95}} = \frac{216.7762}{15.3791} = 14.096 \text{ minutes}^2 \, ,$$

and taking square roots, a 95% upper confidence limit for σ is 3.7544 minutes.   □

6.5.3   Multiple Comparisons for the Main-Effects Model

When the sample sizes are equal, the Bonferroni, Scheffé, Tukey, and Dunnett methods described in Sect. 4.4 can all be used for obtaining simultaneous confidence intervals for sets of contrasts comparing the levels of A or of B. A set of $100(1-\alpha)\%$ simultaneous confidence intervals for contrasts comparing the levels of factor A is of the form (4.4.20), which for the two-way model becomes

$$\sum c_i \alpha_i \quad \in \quad \left(\sum c_i \bar{y}_{i..} \pm w\sqrt{msE \sum c_i^2/br} \right) , \qquad\qquad (6.5.39)$$

where the critical coefficients for the various methods are, respectively,

$$w_B = t_{n-a-b+1,\alpha/2m} \; ; \;\; w_S = \sqrt{(a-1)F_{a-1,n-a-b+1,\alpha}} \; ;$$

$$w_T = q_{a,n-a-b+1,\alpha}/\sqrt{2} \; ; \;\; w_{D2} = |t|^{(0.5)}_{a-1,n-a-b+1,\alpha} \; .$$

Similarly, a set of $100(1-\alpha)\%$ confidence intervals for contrasts comparing the levels of factor B is of the form

$$\sum k_j\beta_j \in \left(\sum k_j\bar{y}_{.j.} \pm w \sqrt{msE \sum k_j^2/ar} \right), \tag{6.5.40}$$

and the critical coefficients are as above after interchanging a and b.

We can also obtain confidence intervals for the treatment means $\mu + \alpha_i + \beta_j$ using the least squares estimators $\bar{Y}_{i..} + \bar{Y}_{.j.} - \bar{Y}_{...}$, each of which has a normal distribution and variance $\sigma^2(a+b-1)/(abr)$. We obtain a set of $100(1-\alpha)\%$ simultaneous confidence intervals for the ab treatment means as

$$\mu + \alpha_i + \beta_j \in \left\{ (\bar{y}_{i..} + \bar{y}_{.j.} - \bar{y}_{...}) \pm w\sqrt{msE\left(\frac{a+b-1}{abr}\right)} \right\}, \tag{6.5.41}$$

with critical coefficient

$$w_{BM} = t_{\alpha/(2ab),(n-a-b+1)} \quad \text{or} \quad w_{SM} = \sqrt{(a+b-1)F_{a+b-1,n-a-b+1,\alpha}}$$

for the Bonferroni and Scheffé methods, respectively.

When confidence intervals are calculated for treatment means and for contrasts in the main effects of factors A and B, an experimentwise confidence level should be calculated. For example, if intervals for contrasts for factor A have overall confidence level $100(1-\alpha_1)\%$, and intervals for B have overall confidence level $100(1-\alpha_2)\%$, and intervals for means have overall confidence level $100(1-\alpha_3)\%$, the experimentwise confidence level for all the intervals combined is at least $100(1-(\alpha_1+\alpha_2+\alpha_3))\%$. Alternatively, w_{SM} could be used in (6.5.39) and (6.5.41), and the overall level for all three sets of intervals together would be $100(1-\alpha)\%$.

Example 6.5.3 Nail varnish experiment, continued

The least squares estimates for the differences in the effects of the two nail varnish solvents and for the pairwise differences in the effects of the three nail varnishes were calculated in Example 6.5.1, p. 162. From Table 6.8, $msE = 8.3375$ with error degrees of freedom $n - a - b + 1 = 26$. There is only $m = 1$ contrast for factor A, and a simple 99% confidence interval of the form (6.5.39) can be used to give

$$\alpha_2 - \alpha_1 \in \left(\bar{y}_{2..} - \bar{y}_{1..} \pm t_{n-a-b+1,\alpha/2}\sqrt{msE(2/br)} \right)$$
$$= \left(3.6040 \pm t_{26,0.005}\sqrt{(8.3375/15)} \right).$$

From Table A.4, $t_{26,0.005} = 2.779$, so a 99% confidence interval for $\alpha_2 - \alpha_1$ is

$$1.5321 \le \alpha_2 - \alpha_1 \le 5.6759 \,.$$

The confidence interval indicates that solvent 2 takes between 1.5 and 5.7 minutes longer, on average, in dissolving the three nail varnishes than does solvent 1.

To compare the nail varnishes in terms of their speed of dissolving, confidence intervals are required for the three pairwise comparisons $\beta_1 - \beta_2$, $\beta_1 - \beta_3$, and $\beta_2 - \beta_3$. If an overall confidence level of 99% is required, Tukey's method gives confidence intervals of the form

$$\beta_j - \beta_p \in \left(\bar{y}_{.j.} - \bar{y}_{.p.} \pm (q_{b,df,0.01}/\sqrt{2}) \sqrt{msE(2/(ar))} \right).$$

From Table A.8, $q_{3,26,0.01} = 4.54$. Using the least squares estimates computed in Example 6.5.1, p. 162, and $msE = 8.3375$ with $n - a - b + 1 = 26$ as above, the minimum significant difference is $msd = (4.54/\sqrt{2}) \sqrt{8.3375(2/10)} = 4.145$. A set of 99% confidence intervals for the pairwise comparisons for factor B is

$$\beta_1 - \beta_2 \in (1.890 \pm 4.145) = (-2.255, 6.035),$$

$$\beta_1 - \beta_3 \in (-3.541, 4.749), \qquad \beta_2 - \beta_3 \in (-5.431, 2.859).$$

Each of these intervals includes zero, indicating insufficient evidence to conclude a difference in the speed at which the nail varnishes dissolve. The overall confidence level for the four intervals for factors A and B together is at least 98%. Bonferroni's method could have been used instead for all four intervals. To have obtained an overall level of at least 98%, we could have set $\alpha^* = \alpha/m = 0.02/4 = 0.005$ for each of the four intervals. The critical coefficient in (6.5.39) would then have been $w_B = t_{0.0025,26} = 3.067$. So the Bonferroni method would have given a longer interval for $\alpha_1 - \alpha_2$ but shorter intervals for $\beta_j - \beta_p$. □

6.5.4   Unequal Variances

When the variances of the error variables are unequal and no equalizing transformation can be found, Satterthwaite's approximation can be used. Since the approximation uses the sample variances of the observations for each treatment combination individually, and since the least squares estimates of the main-effect contrasts are the same whether or not interaction terms are included in the model, the procedure is exactly the same as that illustrated for the bleach experiment in Example 6.4.4, p. 154.

6.5.5   Analysis of Variance for Equal Sample Sizes

Testing Main Effects of B—Equal Sample Sizes

The hypothesis that the levels of B all have the same effect on the response is $H_0^B : \{\beta_1 = \beta_2 = \cdots = \beta_b\}$, which can be written in terms of estimable contrasts as $H_0^B:\{\beta_j - \bar{\beta}_. = 0, \text{ for all } j = 1, \ldots, b\}$. To obtain a test of H_0^B against the alternative hypothesis $H_A^B : \{$ at least two of the β_j's differ$\}$, the sum of squares for error for the two-way main-effects model is compared with the sum of squares for error for the reduced model

$$Y_{ijt} = \mu + \alpha_i + \epsilon_{ijt} . \tag{6.5.42}$$

This is identical to the one-way analysis of variance model (3.3.1) with μ replaced by $\mu* = \mu + \bar{\beta}_.$ and with br observations on the ith level of treatment factor A. Thus ssE_0^B is the same as the sum of squares

for error in a one-way analysis of variance, and can be obtained from (3.4.4), p. 39, by replacing the subscript t by the pair of subscripts jt, yielding

$$ssE_0^B = \sum_i \sum_j \sum_t (y_{ijt} - \bar{y}_{i..})^2 \,. \tag{6.5.43}$$

The sum of squares for testing H_0^B is $ssE_0^B - ssE$, where ssE was derived in (6.5.37), p. 165. So,

$$
\begin{aligned}
ssB &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{t=1}^{r} (y_{ijt} - \bar{y}_{i..})^2 - \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{t=1}^{r} \left((y_{ijt} - \bar{y}_{i..}) - (\bar{y}_{.j.} - \bar{y}_{...}) \right)^2 \\
&= ar \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2 \\
&= ar \sum_j \bar{y}_{.j.}^2 - abr\bar{y}_{...}^2 \,.
\end{aligned}
\tag{6.5.44}
$$

Notice that the formula for ssB is identical to the formula (6.4.25) for testing the equivalent main-effect hypothesis in the two-way complete model. It can be shown that when H_0^B is true, the corresponding random variable SSB/σ^2 has a chi-squared distribution with $(b-1)$ degrees of freedom, and SSB and SSE are independent. Therefore, when H_0^B is true,

$$\frac{SSB/(b-1)\sigma^2}{SSE/(n-a-b+1)\sigma^2} = \frac{MSB}{MSE} \sim F_{b-1,n-a-b+1} \,,$$

and the decision rule for testing H_0^B against H_A^B is

$$\text{reject } H_0^B \text{ if } \frac{msB}{msE} > F_{b-1,n-a-b+1,\alpha} \,. \tag{6.5.45}$$

Testing Main Effects of A—Equal Sample Sizes

A similar rule is obtained for testing $H_0^A : \{\alpha_1 = \alpha_2 = \cdots = \alpha_a\}$ against the alternative hypothesis $H_A^A : \{$at least two of the α_i's differ$\}$. The decision rule is

$$\text{reject } H_0^A \text{ if } \frac{msA}{msE} > F_{a-1,n-a-b+1,\alpha} \,, \tag{6.5.46}$$

where $msA = ssA/(a-1)$, and

$$ssA = br \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 = br \sum_i \bar{y}_{i..}^2 - abr\bar{y}_{...}^2 \,. \tag{6.5.47}$$

similar to the formula (6.4.22) for testing the equivalent hypothesis in the two-way complete model.

Analysis of Variance Table

The information for testing H_0^A and H_0^B is summarized in the analysis of variance table shown in Table 6.7. When sample sizes are equal, $ssE = sstot - ssA - ssB$. When the sample sizes are not equal, the formulae for the sums of squares are complicated, and the analysis should be done by computer (Sects. 6.8 and 6.9).

Table 6.7   Two-Way ANOVA, negligible interaction, equal sample sizes

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Ratio
Factor A	$a-1$	ssA	$\frac{ssA}{a-1}$	$\frac{msA}{msE}$
Factor B	$b-1$	ssB	$\frac{ssB}{b-1}$	$\frac{msB}{msE}$
Error	$n-a-b+1$	ssE	$\frac{ssE}{n-a-b+1}$	
Total	$n-1$	$sstot$		

<div align="center">Computational Formulae for Equal Sample Sizes</div>

$ssA = br \sum_i \bar{y}_{i..}^2 - n\bar{y}_{...}^2$ $ssB = ar \sum_j \bar{y}_{.j.}^2 - n\bar{y}_{...}^2$

$sstot = \sum_i \sum_j \sum_t y_{ijt}^2 - n\bar{y}_{...}^2$ $ssE = sstot - ssA - ssB$

$n = abr$

Table 6.8   Analysis of variance for the nail varnish experiment

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Ratio	p-value
Solvent	1	97.4161	97.4161	11.68	0.0021
Varnish	2	18.6357	9.3178	1.12	0.3423
Error	26	216.7761	8.3375		
Total	29	332.8279			

Example 6.5.4 `Nail varnish experiment, continued`

The analysis of variance table for the nail varnish experiment of Example 6.5.1, p. 162, is given in Table 6.8. The experimenter selected the Type I error probability as 0.05 for testing each of H_0^A and H_0^B, giving an overall error rate of at most 0.1. The ratio $msA/msE = 11.68$ is larger than $F_{1,26,0.05} \approx 4.0$, and therefore, the null hypothesis can be rejected. It can be concluded at individual significance level 0.05 that there is a difference in dissolving times for the two solvents.

The ratio $msB/msE = 1.12$ is smaller than $F_{2,26,0.05} \approx 3.15$. Therefore, the null hypothesis H_0^B cannot be rejected at individual significance level 0.05, and it is not possible to conclude that there is a difference in dissolving time among the three varnishes.   □

6.5.6   Model Building

In some experiments, the primary objective is to find a model that gives an adequate representation of the experimental data. Such experiments are called experiments for *model building*. If there are two crossed, fixed treatment factors, it is legitimate to use the two-way complete model (6.2.2) as a preliminary model. Then, if H_0^{AB} fails to be rejected, the two-way main effects model (6.2.3) can be accepted as a reasonable model to represent the same type of experimental data in *future* experiments.

Note that it is *not legitimate* to adopt the two-way main effects model and to use the corresponding analysis of variance table, Table 6.7, to test further hypotheses or calculate confidence intervals using the *same* set of data. If this is done, the model is changed based on the data, and the quoted significance levels and confidence levels associated with further inferences will not be correct. Model building should be regarded as a completely different exercise from confidence interval calculation. *They should be done using different experimental data.*

6.6   Calculating Sample Sizes

In Chaps. 3 and 4, we showed two methods of calculating sample sizes. The method of Sect. 3.6 aims to achieve a specified power of a hypothesis test, and the method of Sect. 4.5 aims to achieve a specified length of a confidence interval. Both of these techniques rely on knowledge of the largest likely value of σ^2 or msE and can also be used for the two-way complete model.

Alternatively, sample sizes can be calculated to ensure that confidence intervals for main-effect contrasts are no longer than a stated size, using the formulae (6.4.18) and (6.4.19) or, for the two-way main-effects model, the formulae (6.5.39) and (6.5.40).

Similarly, the method of Sect. 3.6 for choosing the sample size to achieve the required power of a hypothesis test can be used for each factor separately, with the modification that the sample size calculation is based on

$$r = 2a\sigma^2\phi^2/(b\Delta_A^2) \tag{6.6.48}$$

for factor A and

$$r = 2b\sigma^2\phi^2/(a\Delta_B^2)$$

for factor B, where Δ_A is the smallest difference in the α_i's (or α_i^*'s) and Δ_B is the smallest difference in the β_j's (or β_j^*'s) that are of interest. The calculation procedure is identical to that in Sect. 3.6, except that the error degrees of freedom are $\nu_2 = n - v$ for the complete model and $\nu_2 = n - a - b + 1$ for the main-effects model (with $n = abr$), and the numerator degrees of freedom are $\nu_1 = a - 1$ for factor A and $\nu_1 = b - 1$ for factor B.

If several different calculations are done and the calculated values of r differ, then the largest value should be selected.

6.7   Small Experiments

6.7.1   One Observation Per Cell

When observations are extremely time-consuming or expensive to collect, an experiment may be designed to have $r = 1$ observation on each treatment combination. Such experiments are called *experiments with one observation per cell* or *single replicate experiments*. Since the ability to choose the sample sizes is lost, it should be recognized that confidence intervals may be wide and hypothesis tests not very powerful.

If it is known in advance that the interaction between the two treatment factors is negligible, then the experiment can be analyzed using the two-way main-effects model (6.2.3). If this information is not available, then the two-way complete model (6.2.2) needs to be used. However, there is a problem. Under the two-way complete model, the number of degrees of freedom for error is $ab(r - 1)$. If $r = 1$, then this number is zero, and σ^2 cannot be estimated.

Thus, a single replicate experiment with a possible interaction between the two factors can be analyzed only if one of the following is true:

(i) σ^2 is known in advance.
(ii) The interaction is expected to be of a certain form that can be modeled with fewer than $(a-1)(b-1)$ degrees of freedom.

(iii) The number of treatment combinations is large, and only a few contrasts are likely to be nonnegligible (*effect sparsity*).

If σ^2 is known in advance, formulae for confidence intervals would be based on the normal distribution, and hypothesis tests would be based on the chi-squared distribution. However, this situation is unlikely to occur, and we will not pursue it. The third case tends to occur when the experiment involves a large number of treatment factors and will be discussed in detail in Chap. 7. Here, we look at the second situation and consider two methods of analysis, the first based on orthogonal contrasts, and the second known as Tukey's test for additivity.

6.7.2   Analysis Based on Orthogonal Contrasts

Two estimable contrasts are called *orthogonal contrasts* if and only if their least squares estimators are uncorrelated or, equivalently, have zero covariance. For the moment, we recode the treatment combinations to obtain a single-digit code, as we did in Chap. 3. Two contrasts $\Sigma c_i \tau_i$ and $\Sigma k_s \tau_s$ are orthogonal if and only if

$$
\begin{aligned}
0 = \mathrm{Cov}\left(\sum_{i=1}^{v} c_i \overline{Y}_{i.}, \sum_{s=1}^{v} k_s \overline{Y}_{s.}\right) &= \sum_{i=1}^{v}\sum_{s=1}^{v} c_i k_s \mathrm{Cov}(\overline{Y}_{i.}, \overline{Y}_{s.}) \\
&= \sum_i c_i k_i \mathrm{Cov}(\overline{Y}_{i.}, \overline{Y}_{i.}) + \sum_i \sum_{s \neq i} c_i k_s \mathrm{Cov}(\overline{Y}_{i.}, \overline{Y}_{s.}) \\
&= \sum_i c_i k_i \mathrm{Var}(\overline{Y}_{i.}) \ + \ 0 \\
&= \sigma^2 \sum_i c_i k_i / r_i \, .
\end{aligned}
$$

In the above calculation $\mathrm{Cov}(\overline{Y}_{i.}, \overline{Y}_{s.})$ is zero when $s \neq i$, because all the Y_{it}'s are independent of each other in the cell-means model. Thus, two contrasts $\Sigma c_i \tau_i$ and $\Sigma k_i \tau_i$ are orthogonal if and only if

$$
\sum_i c_i k_i / r_i = 0 \, . \tag{6.7.49}
$$

If the sample sizes are equal, then this reduces to

$$
\sum_i c_i k_i = 0 \, .
$$

Changing back to two subscripts, we have that two contrasts $\Sigma\Sigma d_{ij} \tau_{ij}$ and $\Sigma\Sigma h_{ij} \tau_{ij}$ are orthogonal if and only if

$$
\sum_{i=1}^{a}\sum_{j=1}^{b} d_{ij} h_{ij} / r_{ij} = 0 \, , \tag{6.7.50}
$$

or, for equal sample sizes, the contrasts are orthogonal if and only if

Table 6.9 Three orthogonal contrasts for the battery experiment

Contrast	Coefficients	$\sum c_i \bar{y}_{i.}$	$\sum c_i^2 / r_i$	ssc
Duty	$\frac{1}{2}[1, \quad 1, -1, -1]$	251.00	$\frac{1}{4}$	252,004.00
Brand	$\frac{1}{2}[1, -1, \quad 1, -1]$	−176.50	$\frac{1}{4}$	124,609.00
Interaction	$\frac{1}{2}[1, -1, -1, \quad 1]$	−113.25	$\frac{1}{4}$	51,302.25

$$\sum_{i=1}^{a} \sum_{j=1}^{b} d_{ij} h_{ij} = 0. \tag{6.7.51}$$

For equal sample sizes, the trend contrasts provide an illustration of orthogonal contrasts. For example, it can be verified that any pair of trend contrasts in Table 6.1, p. 148, satisfy (6.7.51). For the models considered in this book, the contrast estimators are normally distributed, so orthogonality of contrasts implies that their least squares estimators are independent.

For v treatments, or treatment combinations, a set of $v - 1$ orthogonal contrasts is called a *complete set of orthogonal contrasts*. It is not possible to find more than $v - 1$ contrasts that are mutually orthogonal. We write the sum of squares for the qth orthogonal contrast in a complete set as ssc_q, where

$$ssc_q = (\Sigma \Sigma c_{ij} \bar{y}_{ij.})^2 / (\Sigma \Sigma c_{ij}^2 / r_{ij})$$

is the square of the normalized contrast estimator (see Sect. 4.3.3, p. 77). The sum of squares for treatments, ssT, can be partitioned into the sums of squares for the $v - 1$ orthogonal contrasts in a complete set; that is,

$$ssT = ssc_1 + ssc_2 + \cdots + ssc_{v-1}. \tag{6.7.52}$$

Example 6.7.1 `Battery experiment, continued`

Main effect and interaction contrasts for the battery experiment were examined in Example 6.3.1, p. 146 and, following that example, were written as columns in a table. Since the sample sizes are all equal, we need only check that (6.7.51) holds by multiplying corresponding coefficients for any two contrasts and adding their products. The duty, brand, and interaction contrasts form a complete set of $v - 1 = 3$ orthogonal contrasts.

The sums of squares for the three contrasts are shown in Table 6.9. It can be verified that they add to the treatment sum of squares $ssT = 427,915.25$ that was calculated in Example 3.5.1, p. 44. □

We can use the same idea to split the interaction sum of squares $ssAB$ into independent pieces. For the two-way complete model (6.2.2) with $r = 1$ observation per cell, the sum of squares for testing the null hypothesis that a particular interaction contrast, say $\sum_i \sum_j d_{ij} (\alpha\beta)_{ij}$ (with $\sum_i d_{ij} = 0$ and $\sum_j d_{ij} = 0$), is negligible, against the alternative hypothesis that the contrast is not negligible, is

$$ssc = \frac{(\sum_i \sum_j d_{ij} y_{ij})^2}{\sum_i \sum_j d_{ij}^2}. \tag{6.7.53}$$

The interaction has $(a-1)(b-1)$ degrees of freedom. Consequently, there are $(a-1)(b-1)$ orthogonal interaction contrasts in a complete set, and their corresponding sums of squares add to $ssAB$, that is,

Table 6.10 Two-way ANOVA, one observation per cell, e negligible interaction contrasts, and $m = (a-1)(b-1) - e$ interaction degrees of freedom

Source of variation	Degrees of freedom	Sum of squares	Mean square
Factor A	$a-1$	ssA	msA
Factor B	$b-1$	ssB	msB
Interaction	m	$ssAB_m$	$msAB$
Error	e	ssE	msE
Total	$ab-1$	$sstot$	

$$ssAB = \sum_{h=1}^{(a-1)(b-1)} ssc_h ,$$

where ssc_h is the sum of squares for the hth such contrast.

Suppose it is known *in advance* that e specific orthogonal interaction contrasts are likely to be negligible. Then the sums of squares for these e negligible contrasts can be pooled together to obtain an estimate of error variance, based on e degrees of freedom,

$$ssE = \sum_{h=1}^{e} ssc_h \quad \text{and} \quad msE = ssE/e .$$

The sums of squares for the remaining interaction contrasts can be used to test the contrasts individually or added together to obtain an interaction sum of squares

$$ssAB_m = \sum_{h=e+1}^{(a-1)(b-1)} ssc_h .$$

Then the decision rule for testing the hypothesis H_0^{AB}:{the interaction AB is negligible} against the alternative hypothesis that the interaction is not negligible is

$$\text{reject } H_0^{AB} \text{ if } \frac{ssAB_m/m}{ssE/e} > F_{m,e,\alpha} ,$$

where $m = (a-1)(b-1) - e$. Likewise, the main effect test statistics have denominator ssE/e and error degrees of freedom $df = e$. The tests are summarized in Table 6.10, which shows a modified form of the analysis of variance table for the two-way complete model. A worked example is given in Sect. 6.7.4.

To save calculating the sums of squares for all of the contrasts, the error sum of squares is usually obtained by subtraction, that is,

$$ssE = sstot - ssA - ssB - ssAB_m .$$

The above technique is most often used when the factors are quantitative, since higher-order interaction trends are often likely to be negligible. The information about the interaction effects must be known prior to running the experiment. If this information is not available, then one of the techniques discussed in Sect. 7.5 must be used instead.

6.7.3 Tukey's Test for Additivity

Tukey's test for additivity uses only one degree of freedom to measure the interaction. It tests the null hypothesis $H_0^{\gamma} : \{(\alpha\beta)_{ij} = \gamma\alpha_i\beta_j \text{ for all } i, j\}$ against the alternative hypothesis that the interaction is not of this form. The test is appropriate only if the size of the interaction effect is expected to increase proportionally to each of the main effects, and it is not designed to measure any other form of interaction. The test requires that the normality assumption be well satisfied. The decision rule is

$$\text{reject } H_0^{\gamma} \text{ if } \frac{ssAB^*}{ssE/e} > F_{1,e,\alpha}, \tag{6.7.54}$$

where

$$ssAB^* = \frac{ab\left[\sum_i \sum_j y_{ij}\bar{y}_{i.}\bar{y}_{.j} - (ssA + ssB + ab\bar{y}_{..}^2)\bar{y}_{..}\right]^2}{(ssA)(ssB)}$$

and

$$ssE = sstot - ssA - ssB - ssAB^*.$$

The analysis of variance table is as in Table 6.10 with $m = 1$ and with $e = (a - 1)(b - 1) - 1$.

Table 6.11 Data for the air velocity experiment, with factors Rib Height (A) and Reynolds Number (B)

	i	1	2	3	4	5	6	$\bar{y}_{i.}$
				Reynolds Number, j				
Rib	1	−24	−23	1	8	29	23	2.333
Height	2	33	28	45	57	74	80	52.833
	3	37	79	79	95	101	111	83.667
	$\bar{y}_{.j}$	15.333	28.00	41.667	53.333	68.000	71.333	$46.278 = \bar{y}_{..}$

Source Willke (1962). Copyright ©1962 Blackwell Publishers. Reprinted with permission

Fig. 6.7 Data for the air velocity experiment
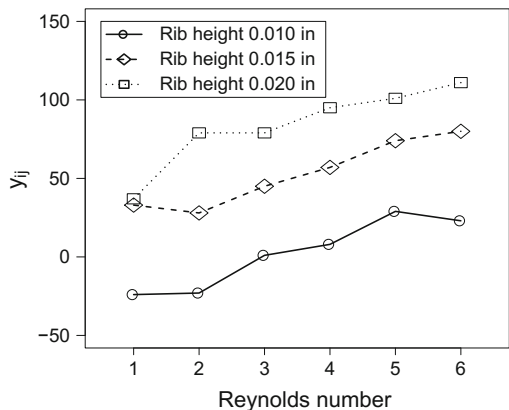
6.7.4   A Real Experiment—Air Velocity Experiment

The data given in Table 6.11, and plotted in Fig. 6.7, form part of an experiment described by D. Wilkie in the 1962 issue of *Applied Statistics* (volume 11, pages 184–195). The experiment was designed to examine the position of maximum velocity of air blown down the space between a roughened rod and a smooth pipe surrounding it. The treatment factors were the height of ribs on the roughened rod (factor A) at equally spaced heights 0.010, 0.015, and 0.020 inches (coded 1, 2, 3) and Reynolds number (factor B) at six levels (coded 1–6) equally spaced logarithmically over the range 4.8–5.3. The responses were measured as $y = (d - 1.4) \times 10^3$, where d is the distance in inches from the center of the rod.

Figure 6.7 shows very little interaction between the factors. However, prior to the experiment, the investigators had thought that the factors would interact to some extent. They wanted to use the set of orthogonal polynomial trend contrasts for the AB interaction and were reasonably sure that the contrasts $A_Q B_{qr}, A_L B_{qn}, A_Q B_{qn}$ would be negligible. Thus the sum of squares for these three contrasts could be used to estimate σ^2 with 3 degrees of freedom. We are using "L, Q, C, qr, qn" as shorthand notation for linear, quadratic, cubic, quartic, and quintic contrasts, respectively. The coefficients for these three orthogonal polynomial trend contrasts can be obtained by multiplying the corresponding main-effect coefficients shown in Table 6.1, p. 148. The coefficients for $A_L B_{qn}$ are shown in the table as an example. Also shown are the contrast coefficients for the linear$A \times$linear$B$ contrast, $A_L B_L$. These are

$$[\, 5, \ 3, \ 1, \ -1, \ -3, \ -5, \ 0, \ 0, \ 0, \ 0, \ 0, \ 0, \ -5, \ -3, \ -1, \ 1, \ 3, \ 5 \,].$$

The estimate of $A_L B_L$ is then

Table 6.12 Analysis of variance for the air velocity experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	*p*-value
Rib height (A)	2	20232.111			
A_L	1	19845.333	19845.333	338.77	0.0003
A_Q	1	386.778	386.778	6.60	0.0825
Reynolds number (B)	5	7386.944			
B_L	1	7262.976	7262.976	123.98	0.0016
B_Q	1	65.016	65.016	1.11	0.3695
B_C	1	36.296	36.296	0.62	0.4887
B_{qr}	1	13.762	13.762	0.23	0.6611
B_{qn}	1	8.894	8.894	0.15	0.7228
Interaction (AB)	7	616.817			
$A_L B_L$	1	20.829	20.829	0.36	0.5930
$A_L B_Q$	1	47.149	47.149	0.80	0.4358
$A_L B_C$	1	265.225	265.225	4.53	0.1233
$A_L B_{qr}$	1	33.018	33.018	0.56	0.5073
$A_Q B_L$	1	15.238	15.238	0.26	0.6452
$A_Q B_Q$	1	170.335	170.335	2.91	0.1867
$A_Q B_C$	1	65.023	65.023	1.11	0.3694
Error	3	175.739	58.580		
Total	17	28411.611			

$$\Sigma\Sigma d_{ij}y_{ij} = 5(-24) + 3(-23) + \cdots + 3(101) + 5(111) = 54\,.$$

Now,

$$\Sigma\Sigma d_{ij}^2 = (5^2 + 3^2 + \cdots + 3^2 + 5^2) = 140\,,$$

so the corresponding sum of squares is

$$ss(A_{\mathrm{L}}B_{\mathrm{L}}) = \frac{54^2}{140} = 20.829\,.$$

The sums of squares for the other contrasts are computed similarly, and the error sum of squares is calculated as the sum of the sums of squares of the three negligible contrasts. The analysis of variance table is given in Table 6.12.

The hypotheses that the individual contrasts are zero can be tested using Scheffé's procedure or Bonferroni's procedure. If Bonferroni's procedure is used, each of the 14 hypotheses should be tested at a very small α-level. Taking $\alpha = 0.005$, so that the overall level is at most 0.07, we have $F_{1,3,0.005} = 55.6$, and only the linear A and linear B contrasts appear to be significantly different from zero. The plot of the data shown in Fig. 6.7 supports this conclusion.

6.8   Using SAS Software

Table 6.13 contains a sample SAS program for analysis of the two-way complete model (6.2.2). For illustration, we use the data of the reaction time experiment shown in Table 4.4, p. 101, but with the last four observati ons missing, so that $r_{11} = r_{21} = 2, r_{12} = r_{22} = r_{23} = 3, r_{13} = 1$. In the data input lines, the levels of each of the two treatment factors A and B are shown together with the response, the order in which the observations were collected, and the treatment factor level TRTMT. A two-digit code for each treatment combination TC is easily generated by the statement TC = 10*A + B following the INPUT statement. This way of coding the treatment combinations works well for all applications except for drawing plots with TC on one axis. Such a plot would not show numeric codes 11, 12, …, 23 as equally spaced. In the statement TC = PUT(10* A + B, 2.), the function PUT converts the created variable from numeric to character.

6.8.1   Analysis of Variance

The GLM procedure in Table 6.13 is used to generate the analysis of variance table, to estimate and to test contrasts, and for multiple comparisons. As in the one-way analysis of variance, the treatment factors must be declared as class variables using a CLASS statement. The two-way complete model is represented as

```
MODEL Y = A B A*B;
```

with the main effects listed in either order, but before the interaction. The two-way main-effects model (6.2.3) would be represented as

```
MODEL Y = A B;
```

The program also shows the cell-means model (6.2.1) in a second GLM procedure, using

```
MODEL Y = TC;
```

Table 6.13 SAS program to illustrate aspects of analysis of a two-way complete model (reaction time experiment)

```
DATA RTIME;
 INPUT ORDER TRTMT A B Y;
 TC = PUT(10*A + B, 2.); * create TC as a character variable for plots;
 LINES;
 1 6 2 3 0.256
 2 6 2 3 0.281
 : : : :  :
 14 4 2 1 0.279
;
PROC GLM;
  CLASS A B;
  MODEL Y = A B A*B;
  LSMEANS A / PDIFF CL ALPHA=0.01;
  LSMEANS B / PDIFF = ALL CL ADJUST = TUKEY ALPHA = 0.01;
  CONTRAST '11-13-21+23'  A*B 1  0 -1 -1  0  1;
  CONTRAST 'B1-B2' B 1 -1  0;
  ESTIMATE 'B1-B2' B 1 -1  0;
  ESTIMATE 'B1-B3' B 1  0 -1;
  ESTIMATE 'B2-B3' B 0  1 -1;
PROC GLM;
  CLASS TC;
  MODEL Y = TC;
  LSMEANS TC / PDIFF = ALL CL ADJUST = TUKEY ALPHA = 0.01;
  LSMEANS TC / PDIFF = CONTROL CL ADJUST = DUNNETT ALPHA = 0.01;
  LSMEANS TC / PDIFF = CONTROLL CL ADJUST = DUNNETT ALPHA = 0.01;
  LSMEANS TC / PDIFF = CONTROLU CL ADJUST = DUNNETT ALPHA = 0.01;
  CONTRAST '11-13-21+23'  TC 1  0 -1  -1  0  1;
  CONTRAST 'B1-B2'        TC 1 -1  0   1 -1  0;
```

The output from the first GLM procedure is shown in Fig. 6.8. The analysis of variance table is organized differently from that in Table 6.4, p. 159. The five "model" degrees of freedom are the treatment degrees of freedom corresponding to the six treatment combinations. Information concerning main effects and interactions is provided underneath the table under the heading "Type I" and "Type III" sums of squares.

The *Type III sums of squares* are the values *ssA*, *ssB*, and *ssAB* and are used for hypothesis testing whether or not the sample sizes are equal. They are calculated by comparing the sum of squares for error in the full and reduced models as in Sect. 6.4.4. The sums of squares listed in the output are always in the same order as the effects in the MODEL statement, but the hypothesis of no interaction should be tested first.

The *Type I sum of squares* for an effect is the additional variation in the data that is explained by adding that effect to a model containing the previously listed sources of variation. For example, in the program output, the Type I sum of squares for *A* is the reduction in the error sum of squares that is achieved by adding the effect of factor *A* to a model containing only an intercept term. The reduction in the error sum of squares is equivalent to the extra variation in the data that is explained by adding *A* to the model. Here, the "full model" contains *A* and the intercept, while the "reduced model" contains only

Fig. 6.8 Some output for the SAS program for a two-way complete model with unequal sample sizes, using data from the reaction time experiment Table 4.4, p. 101, omitting the last 4 observations

Results Viewer - SAS Output

The GLM Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.02153160	0.00430632	13.38	0.0010
Error	8	0.00257533	0.00032192		
Corrected Total	13	0.02410693			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	1	0.02101572	0.02101572	65.28	<.0001
B	2	0.00033302	0.00016651	0.52	0.6148
A*B	2	0.00018286	0.00009143	0.28	0.7600

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	0.01682504	0.01682504	52.27	<.0001
B	2	0.00045773	0.00022887	0.71	0.5198
A*B	2	0.00018286	0.00009143	0.28	0.7600

Results Viewer - SAS Output

The GLM Procedure

Dependent Variable: Y

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
11-13-21+23	1	0.00013886	0.00013886	0.43	0.5298
B1-B2	1	0.00017340	0.00017340	0.54	0.4839

Parameter	Estimate	Standard Error	t Value	Pr > |t|
B1-B2	0.00850000	0.01158153	0.73	0.4839
B1-B3	-0.00600000	0.01370346	-0.44	0.6731
B2-B3	-0.01450000	0.01268694	-1.14	0.2861

the intercept. The Type I sum of squares for B is the additional variation in the data that is explained by adding the effect of factor B to a model that already contains the intercept and the effect of A (so that the "full model" contains A, B and the intercept, while the "reduced model" contains only the A and the intercept). The Type I sums of squares (also known as *sequential sums of squares*) depend upon the order in which the effects are listed in the MODEL statement. Type I sums of squares are used for model building, not for hypothesis testing under an assumed model. Consequently, we will use only the Type III sums of squares.

The Type I and Type III sums of squares are identical when the sample sizes are equal, since the factorial effects are then estimated independently of one another. But when the sample sizes are

unequal, as in the illustrated data set, the Type I and Type III sums of squares differ. In the absence of a sophisticated computer package, each Type I and Type III sum of squares can be calculated as the difference of the error sums of squares obtained from two analysis of variance tables, one for the full model and one for the reduced model.

6.8.2   Contrasts and Multiple Comparisons

In the first GLM procedure in Table 6.13, the two-way complete model is used, and the coefficient lists are entered for each factor separately, rather than for the treatment combinations. The first CONTRAST statement is used to test the hypothesis that the interaction contrast $(\alpha\beta)_{11}-(\alpha\beta)_{13}-(\alpha\beta)_{21}+(\alpha\beta)_{23}$ is negligible, and the second CONTRAST statement is used to test the hypothesis that $\beta_1^*-\beta_2^*$ is negligible. These same contrasts are entered as coefficient lists for the treatment combinations in the second GLM procedure. In either case, the contrast sum of squares is as shown under Contrast SS in Fig. 6.8, and the p-value for the test is as shown under Pr > F.

The statement

```
LSMEANS A / PDIFF CL ALPHA = 0.01;
```

of the first GLM procedure causes generation of a 99% confidence interval for the main effect of A pairwise comparison, $\alpha_1^* - \alpha_2^*$, comparing the effects of A averaged over the levels of B, as well as an individual 99% confidence interval for each of the A means $\mu + \alpha_i + \overline{\beta}_{.} + \overline{(\alpha\beta)}_{i.}$.

The statement

```
LSMEANS B / PDIFF = ALL CL ADJUST = TUKEY ALPHA = 0.01;
```

of the first GLM procedure causes generation of Tukey's simultaneous 99% confidence intervals, comparing pairwise the main effects of the three levels of B, each averaged over the levels of A. The option PDIFF = ALL requests p-values for all pairwise comparisons, the option CL asks for the comparisons to be displayed as confidence intervals, and the option ADJUST = TUKEY when coupled with PDIFF = ALL requests Tukey's method for the pairwise comparisons. The output for the reaction time experiment, shown in Fig. 6.9, includes not only the confidence intervals for pairwise comparisons, but also p-values for simultaneous hypothesis tests using the Tukey method. Also given are individual 99% confidence intervals for the B means $\mu + \overline{\alpha}_{.} + \beta_j + \overline{(\alpha\beta)}_{.j}$. Because sample sizes are unequal, these least squares means are not simply the corresponding treatment sample means. If CL is omitted, then only the simultaneous tests and intervals for means are printed. The request TUKEY can be replaced by BON or SCHEFFE as appropriate.

In the second GLM procedure in Table 6.13, the cell-means model is used, with a treatment effect τ_{ij} associated with each treatment combination ij. The corresponding LSMEANS statements illustrate multiple comparisons of the effects τ_{ij} of the six treatment combinations, the first LSMEANS statement generating Tukey's method for all pairwise comparisons, and the remaining LSMEANS statement generating Dunnett's method for comparing all treatments with a control. To generate Dunnett's method, the option PDIFF = ALL is replaced by the option PDIFF = CONTROL for two-sided confidence intervals, and by the option PDIFF = CONTROLL or PDIFF = CONTROLU for upper or lower confidence bounds on the treatment-versus-control differences, as follows:

```
LSMEANS TC / PDIFF = CONTROL   CL ADJUST = DUNNETT ALPHA = 0.01;
LSMEANS TC / PDIFF = CONTROLL CL ADJUST = DUNNETT ALPHA = 0.01;
LSMEANS TC / PDIFF = CONTROLU CL ADJUST = DUNNETT ALPHA = 0.01;
```

Fig. 6.9 LSMEANS output for a two-way complete model with unequal sample sizes (reaction time experiment)
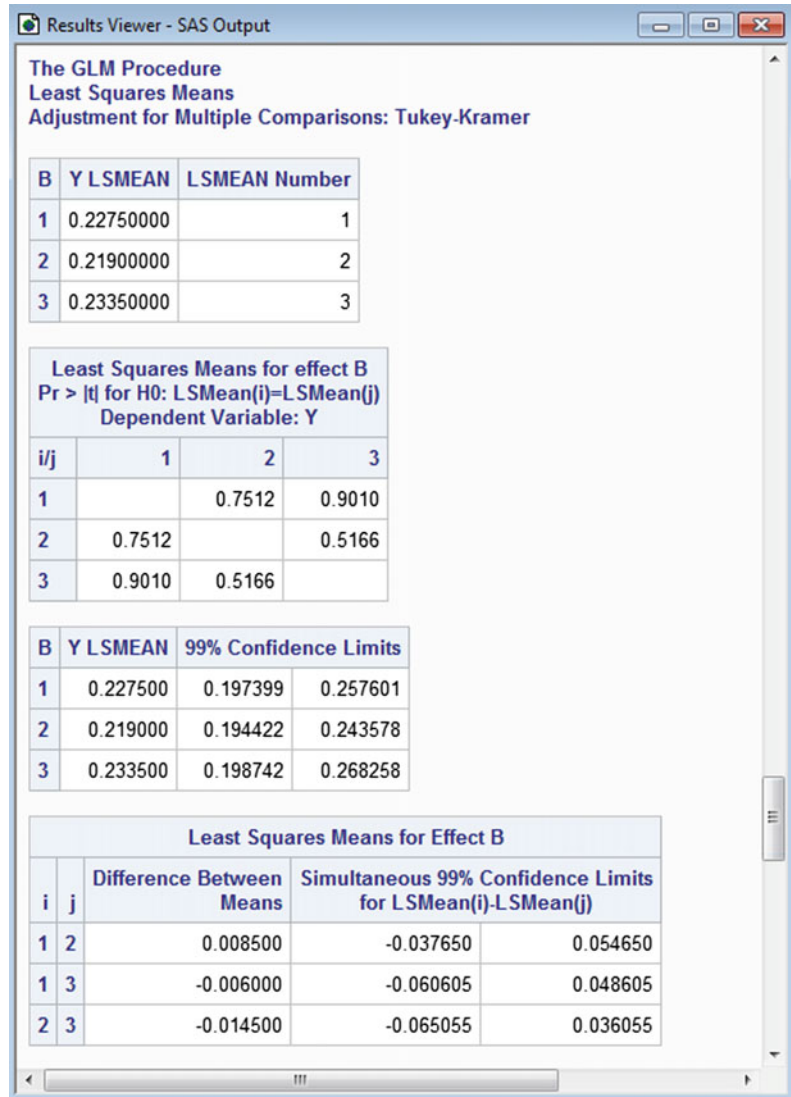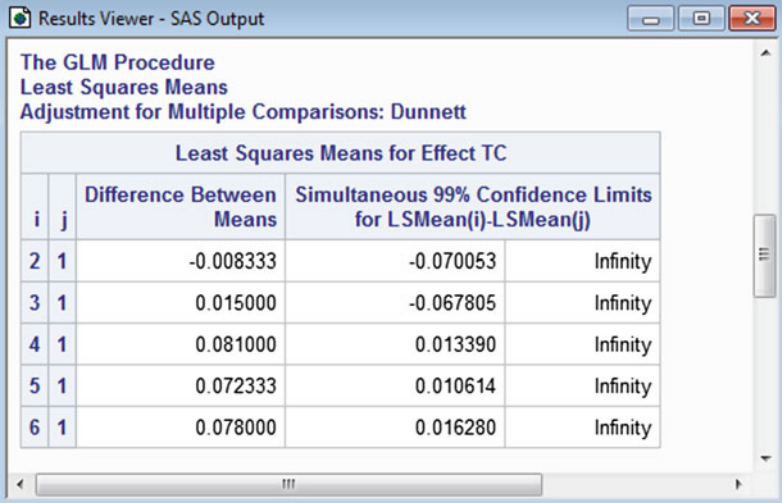
Figure 6.10 contains the output for the third set of simultaneous confidence intervals—namely, corresponding to the CONTROLU option. This set gives lower bounds for the treatment-minus-control comparisons, corresponding to upper-tailed inferences. The treatments are renumbered by SAS in numerical order. In our program, in Table 6.13, we have requested the treatment-versus-control contrasts be done for the treatment combinations 11, 12, 13, 21, 22, 23. SAS recodes these as 1–6, and treatment 1 (our treatment combination 11), as the lowest treatment combination, is taken as the control by default. One could specify treatment combination 23 as the control, for example, via the option PDIFF = CONTROL('23'). We have shown only the simultaneous confidence intervals, but simultaneous tests are also given by SAS software.

We remind the reader that for unequal sample sizes, it has not yet been proved that the overall confidence levels achieved by the Tukey and Dunnett methods are at least as great as those stated, except in some special cases such as the one-way layout.

An alternative method of obtaining simultaneous confidence intervals for pairwise comparisons can be obtained from the output of the ESTIMATE statement for each contrast. The corresponding

Fig. 6.10 Dunnett's lower bound output for a two-way complete model with unequal sample sizes (reaction time experiment)

confidence intervals are of the form

$$\text{Estimate} \pm w \, (\text{Std Error of Estimate}),$$

where w is the critical coefficient given in (6.4.18) for the complete model and in (6.5.39) for the main-effects model.

6.8.3 Plots

Residual plots for checking the error assumptions on the model are generated in the same way as shown in Chap. 5. If the two-way main-effects model (6.2.3) is used, the assumption of additivity should also be checked. For this purpose it is useful to plot the standardized residuals against the level of one factor, using the levels of the other factor for plotting labels (see, for example, Fig. 6.3, p. 144, for the temperature experiment). A plot of the standardized residuals z against the levels of factor A using the labels of factor B can be generated using the following SAS program lines:

```
PROC SGPLOT;
  SCATTER X = A Y = Z / GROUP = B;
```

An interaction plot can be obtained by adding the following statements to the end of the program in Table 6.13:

```
PROC SORT DATA = RTIME; BY A B;
PROC MEANS DATA = RTIME NOPRINT MEAN VAR; BY A B;
  VAR Y;
  OUTPUT OUT = RTIME2 MEAN = AV_Y VAR = VAR_Y;
PROC PRINT;
  VAR A B AV_Y VAR_Y;
PROC SGPLOT;
  SERIES X = A Y = AV_Y / GROUP = B;
```

The `PROC PRINT` statement following `PROC MEANS` also gives the information about the variances that would be needed to check the rule of thumb that $s_{max}^2/s_{min}^2 \leq 3$.

In order to check for equal error variances, the residuals or the observations may be plotted against the treatment combinations using the following SAS code:

```
PROC SGPLOT;
   SCATTER X = TC Y = Z;   *or Y = Y for observations;
```

If the treatment combination codes were created as `TC = 10*A + B`, they will not be equally spaced along the axis, since the codes 11, 12, 13, 21, 22, 23 when regarded as 2-digit numbers are not equally spaced. A simple solution to this problem, as shown in Table 6.13, is to convert the variable `TC` from numeric to character via the statement

```
TC = PUT(10*A + B, 2.);
```

A plot of the residuals or the observations against the character variable `TC` will show the character variable codes evenly spaced along the axis.

When there are not sufficient observations to be able to check equality of error variances for all the cells, the standardized residuals should be plotted against the levels of each factor. The rule of thumb may be checked for the levels of each factor by comparing the maximum and minimum variances of the (nonstandardized) residuals. This is done for factor A by the following lines after creation of RTIME data set as in Table 6.13.

```
PROC GLM;
   CLASS TC;
   MODEL Y = TC;
   OUTPUT OUT = RESIDS RESIDUAL = E;
PROC SORT DATA = RESIDS; BY A;
PROC MEANS DATA = RESIDS NOPRINT VAR; BY A;
   VAR E;
   OUTPUT OUT = RTIME2 VAR = VAR_E;
PROC PRINT;
   VAR A VAR_E;
```

6.8.4   One Observation Per Cell

In order to split the interaction sum of squares into parts corresponding to negligible and nonnegligible orthogonal contrasts, we can enter the data in the usual manner and obtain the sums of squares for all of the contrasts via CONTRAST statements in the procedure PROC GLM. The analysis of variance table can then be constructed with the error sum of squares being the sum of the contrast sums of squares for the negligible contrasts. It is possible, however, to achieve this in a more direct way, as follows.

First, enter the contrast coefficients as part of the input data as shown in Table 6.14 for the air velocity experiment. In the air velocity experiment, factor A had $a = 3$ levels and factor B had $b = 6$ levels. The main-effect trend contrast coefficients are entered via the INPUT statement line by line directly from Table 6.1, p. 148, and the interaction trend contrast coefficients are obtained by multiplication following the INPUT statement. In the PROC GLM statement, the CLASS designation is omitted. If it were included, then `Aln`, for example, would be interpreted as one factor with three coded levels -1, 0, 1, and `Aqd` as a second factor with two coded levels 1, -2, and so on. The model is fitted using those contrasts that have not been declared to be negligible. The error sum of squares will be based on the three omitted contrasts `AlnBqn`, `AqdBqr`, and `AqdBqn`, and the resulting analysis of variance table will be equivalent to that in Table 6.12, p. 176.

It is not necessary to input the levels of A and B separately as we have done in columns 2 and 3 of the data, but these would be needed if plots of the data were required.

Table 6.14 Fitting a model in terms of contrasts (air velocity experiment)

```
DATA AIR;
  INPUT Y A B Aln Aqd Bln Bqd Bcb Bqr Bqn;
  AlnBln = Aln*Bln;
  AlnBqd = Aln*Bqd;
  AlnBcb = Aln*Bcb;
  AlnBqr = Aln*Bqr;
  AqdBln = Aqd*Bln;
  AqdBqd = Aqd*Bqd;
  AqdBcb = Aqd*Bcb;
  LINES;
  -24      1   1       -1    1       -5    5   -5     1   -1
  -23      1   2       -1    1       -3   -1    7    -3    5
    1      1   3       -1    1       -1   -4    4     2  -10
    8      1   4       -1    1        1   -4   -4     2   10
   29      1   5       -1    1        3   -1   -7    -3   -5
   23      1   6       -1    1        5    5    5     1    1
   33      2   1        0   -2       -5    5   -5     1   -1
   28      2   2        0   -2       -3   -1    7    -3    5
   45      2   3        0   -2       -1   -4    4     2  -10
   57      2   4        0   -2        1   -4   -4     2   10
   74      2   5        0   -2        3   -1   -7    -3   -5
   80      2   6        0   -2        5    5    5     1    1
   37      3   1        1    1       -5    5   -5     1   -1
   79      3   2        1    1       -3   -1    7    -3    5
   79      3   3        1    1       -1   -4    4     2  -10
   95      3   4        1    1        1   -4   -4     2   10
  101      3   5        1    1        3   -1   -7    -3   -5
  111      3   6        1    1        5    5    5     1    1
;
PROC PRINT;
PROC GLM; * omit the class statement;
  MODEL Y = Aln Aqd Bln Bqd Bcb Bqr Bqn AlnBln AlnBqd
                  AlnBcb AlnBqr AqdBln AqdBqd  AqdBcb;
```

6.9   Using R Software

Table 6.15 contains a sample R program for analysis of the two-way complete model (6.2.2). For illustration, we use the data of the reaction time experiment shown in Table 4.4, p. 101, but with the last four observations missing, so that $r_{11} = r_{21} = 2, r_{12} = r_{22} = r_{23} = 3, r_{13} = 1$. The data file includes the levels of each of the two treatment factors A and B, as well as the response, the order in which the observations were collected, and treatment factor levels 1–6. A more descriptive two-digit code for each treatment combination TC = 10*A + B is easily generated and added to the data set react.data, along with factor variables fTC, fA, and fB, by the statement

```
react.data = within(react.data,
  {TC = 10*A + B; fTC = factor(TC); fA = factor(A); fB = factor(B)})
```

Table 6.15   R program to illustrate aspects of analysis of a two-way complete model (reaction time experiment)

```
react.data = read.table("data/reaction.time.txt", header=T)
react.data = head(react.data, 14) # Keep first 14 observations
head(react.data, 3)

    Order Trtmt A B     y
  1     1     6 2 3 0.256
  2     2     6 2 3 0.281
  3     3     2 1 2 0.167

# Create trtmt combo vbl TC and factors fTC, fA, and fB within data set
react.data = within(react.data,
  {TC = 10*A + B; fTC = factor(TC); fA = factor(A); fB = factor(B)})
summary(react.data[,c("fA","fB","fTC","y")])

# ANOVA
options(contrasts = c("contr.sum", "contr.poly"))
modelAB = aov(y ~ fA + fB + fA:fB, data = react.data)
anova(modelAB) # Type I ANOVA
drop1(modelAB, ~., test = "F") # Type III ANOVA
modelTC = aov(y ~ fTC, data = react.data)
anova(modelTC) # Model F-test

# Contrasts: estimates, CIs, tests
library(lsmeans)
# Main-effect-of-B contrast: B1-B2
lsmB = lsmeans(modelAB, ~ fB)
summary(contrast(lsmB, list(B12=c( 1,-1, 0))), infer=c(T,T))
# AB-interaction contrast: AB11-AB13-AB21+AB23
lsmAB = lsmeans(modelAB, ~ fB:fA) # Using "fB:fA" yields AB lex order
lsmAB # Display to see order of AB combos for contrast coefficients
summary(contrast(lsmAB, list(AB=c( 1 ,0,-1,-1, 0, 1))), infer=c(T,T))

# Multiple comparisons: B
confint(lsmB, level=0.99) # lsmeans for B and 99% CIs
# Tukey's method
summary(contrast(lsmB, method="pairwise", adjust="tukey"),
        infer=c(T,T), level=0.99)
# Dunnett's method
summary(contrast(lsmB, method="trt.vs.ctrl", adj="mvt", ref=1),
        infer=c(T,T), level=0.99)
```

This way of coding the treatment combinations works well for all applications except for drawing plots with TC on one axis. Such a plot would not show codes 11, 12, and 21 as equally spaced. An alternative way of creating the treatment combinations axis for plots will be given in Sect. 6.9.3.

Table 6.16 Analysis of variance output for the R program for a two-way complete model with unequal sample sizes (reaction time experiment)

```
> # ANOVA
> options(contrasts = c("contr.sum", "contr.poly"))
> modelAB = aov(y ~ fA + fB + fA:fB, data = react.data)
> anova(modelAB) # Type I ANOVA

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value   Pr(>F)
fA         1 0.02102 0.02102   65.28 0.000041
fB         2 0.00033 0.00017    0.52     0.61
fA:fB      2 0.00018 0.00009    0.28     0.76
Residuals  8 0.00258 0.00032

> drop1(modelAB, ~., test = "F") # Type III ANOVA

Single term deletions

Model:
y ~ fA + fB + fA:fB
       Df Sum of Sq      RSS     AIC F value  Pr(>F)
<none>               0.00258 -108.4
fA      1   0.01683 0.01940   -82.1   52.27 0.00009
fB      2   0.00046 0.00303  -110.1    0.71    0.52
fA:fB   2   0.00018 0.00276  -111.5    0.28    0.76

> modelTC = aov(y ~ fTC, data = react.data)
> anova(modelTC) # Model F-test

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value Pr(>F)
fTC        5 0.02153 0.00431    13.4  0.001
Residuals  8 0.00258 0.00032
```

6.9.1 Analysis of Variance

Referring to the R program in Table 6.15, the block of code under the comment "ANOVA" generates the analysis of variance output as shown in Table 6.16. As in the one-way analysis of variance, the treatment factors must be factor variables to be modeled as qualitative variables. The statements

```
modelAB = aov(y ~ fA + fB + fA:fB, data = react.data)
modelTC = aov(y ~ fTC, data = react.data)
```

fit the two-way complete model (6.2.2) and the cell-means model (6.2.1), respectively, saving the results as modelAB and modelTC. In the first model, the main effects may be listed in either order,

but before the interaction effects `fA:fB`. Equivalently, the model could be specified as `y ~ fA*fB`, since `fA*fB` indicates inclusion of all main effects and interactions involving the factors `fA` and `fB`. The two-way main-effects model (6.2.3) would be represented as `y ~ fA + fB`.

Using the saved information from the fitted models, the three statements

```
anova(modelAB)
drop1(modelAB, ~., test ="F")
anova(modelTC)
```

respectively generate the "`Type I ANOVA`" shown at the top of Table 6.16, the "`Type III ANOVA`" shown in the middle of the table, and the "`Model F-test`" shown at the bottom of the table. In the first and last portion of the table, "`Residuals`" is synonymous with error. Totals for the degrees of freedom and sum of squares are not provided. For technical reasons, the statement

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

must be executed prior to fitting the two-way complete model for all Type III sum of squares to be correct. (This option imposes common sum-to-zero constraints on the least squares estimates of the treatment factors effects, so each $\hat{\alpha}_i$ estimates a main effect of A contrast, each $\hat{\beta}_j$ estimates a main effect of B contrast, and each $\widehat{(\alpha_i\beta_j)}$ estimates an AB-interaction contrast.)

In the "`Type III ANOVA`" given in the middle of Table 6.16, the (Type III) sums of squares are the values *ssA*, *ssB*, and *ssAB* and are used for hypothesis testing whether or not the sample sizes are equal. Each is calculated by comparing the sum of squares for error in the full and reduced models as in Sect. 6.4.4, where each reduced model is obtained by dropping the corresponding term from the full model—the two-way complete model in this case. The hypothesis of no interaction should be tested first, even though its sum of squares is listed last in the output. For each effect, its listed `F value` may be computed from its listed sum of squares and degrees of freedom, and from the mean square for residuals listed in the analysis of variance table at either the top and bottom of Table 6.16. The corresponding *p*-value is listed under "`Pr(>F)`". (The reader may ignore the `RSS` and `AIC` columns.)

The sums of squares listed in the "`Type I ANOVA`" shown at the top of Table 6.16 are *Type I sum of squares*, or sequential sum of squares. For each effect, this is the additional variation in the data that is explained by adding that effect to a model containing the previously listed sources of variation. *Type I sum of squares* are discussed in more detail in Sect. 6.8.1. They are used for model building, not for hypothesis testing under an assumed model. Consequently, we will use only the Type III sums of squares. That said, the Type I and Type III sums of squares are identical when the sample sizes are equal, since the factorial effects are then estimated independently of one another. When the Type I and III analyses are the same, it seems preferable to use the cleaner Type I analysis of variance table as shown at the top of Table 6.16.

Finally, under "`Model F-test`" at the bottom of Table 6.16, an analysis of variance table is provided for testing model significance.

6.9.2   Contrasts and Multiple Comparisons

Information on individual contrasts is generated by coupling the `summary` and `contrast` functions of the `lsmeans` package, as was illustrated in Chap. 4. After loading the `lsmeans` package, the statements

```
lsmB = lsmeans(modelAB, ~ fB)
summary(contrast(lsmB, list(B12=c( 1,-1, 0))), infer=c(T,T))
```

Table 6.17 Contrasts output for the R program for a two-way complete model with unequal sample sizes (reaction time experiment)

```
> # Main-effect-of-B contrast: B1-B2
> lsmB = lsmeans(modelAB, ~ fB)

> summary(contrast(lsmB, list(B12=c( 1,-1, 0))), infer=c(T,T))

 contrast estimate        SE df  lower.CL upper.CL t.ratio p.value
 B12         0.0085 0.011582  8 -0.018207 0.035207   0.734  0.4839

Results are averaged over the levels of: fA
Confidence level used: 0.95

> # AB-interaction contrast: AB11-AB13-AB21+AB23
> lsmAB = lsmeans(modelAB, ~ fB:fA) # Using "fB:fA" yields AB lex order
> lsmAB # Display to see order of AB combos for contrast coefficients

 fB fA  lsmean       SE df lower.CL upper.CL
 1  1  0.18700 0.012687  8  0.15774  0.21626
 2  1  0.17867 0.010359  8  0.15478  0.20255
 3  1  0.20200 0.017942  8  0.16063  0.24337
 1  2  0.26800 0.012687  8  0.23874  0.29726
 2  2  0.25933 0.010359  8  0.23545  0.28322
 3  2  0.26500 0.010359  8  0.24111  0.28889

Confidence level used: 0.95

> summary(contrast(lsmAB, list(AB=c( 1 ,0,-1,-1, 0, 1))), infer=c(T,T))

 contrast estimate        SE df lower.CL upper.CL t.ratio p.value
 AB         -0.018 0.027407  8  -0.0812   0.0452  -0.657  0.5298

Confidence level used: 0.95
```

in Table 6.15 generate information for the main effect of B contrast $\beta_1^* - \beta_2^*$, comparing the effects of B averaged over the levels of A, using the two-way complete model previously fit and saved as modelAB. The following statements generate analogous information for the interaction contrast $(\alpha\beta)_{11} - (\alpha\beta)_{13} - (\alpha\beta)_{21} + (\alpha\beta)_{23}$.

```
lsmAB = lsmeans(modelAB, ~ fB:fA)
summary(contrast(lsmAB, list(AB=c( 1 ,0,-1,-1, 0, 1))), infer=c(T,T))
```

Using fB:fA rather than fA:fB yields least squares means in the standard lexicographical order, as can be seen by displaying lsmAB, and the contrast coefficients must be in the corresponding order. These statements and their output are shown in Table 6.17.

Multiple comparisons procedures are also implemented using functions of the lsmeans package, as illustrated for levels of B by sample code at the bottom of Table 6.15. The statements

```
lsmB = lsmeans(modelAB, ~ fB)
confint(lsmB, level = 0.99)
```

Table 6.18 Multiple comparisons output for a two-way complete model with unequal sample sizes (reaction time experiment)

```
> # Multiple comparisons: B
> confint(lsmB, level=0.99) # lsmeans for B and 99% CIs

 fB lsmean        SE df lower.CL upper.CL
 1  0.2275 0.0089710  8  0.19740  0.25760
 2  0.2190 0.0073248  8  0.19442  0.24358
 3  0.2335 0.0103588  8  0.19874  0.26826

Results are averaged over the levels of: fA
Confidence level used: 0.99

> # Tukey's method
> summary(contrast(lsmB, method="pairwise", adjust="tukey"),
+         infer=c(T,T), level=0.99)

 contrast estimate       SE df  lower.CL upper.CL t.ratio p.value
 1 - 2       0.0085 0.011582  8 -0.037650 0.054650   0.734  0.7512
 1 - 3      -0.0060 0.013703  8 -0.060606 0.048606  -0.438  0.9010
 2 - 3      -0.0145 0.012687  8 -0.065055 0.036055  -1.143  0.5166

Results are averaged over the levels of: fA
Confidence level used: 0.99
Confidence-level adjustment: tukey method for a family of 3 estimates
P value adjustment: tukey method for a family of 3 tests
```

in turn (i) compute least squares estimates of the means $\mu + \beta_j + \overline{\alpha}_. + \overline{(\alpha\beta)}_{.j}$ and related information, saving this information as lsmB; and (ii) display the least squares estimates, standard errors, degrees of freedom, and individual 99% confidence intervals shown at the top of Table 6.18. Because sample sizes are unequal, these least squares means are not simply the corresponding treatment sample means.

The statement

```
summary(contrast(lsmB, method = "pairwise", adjust = "tukey"),
        infer = c(T, T), level = 0.99)
```

applies Tukey's method, comparing pairwise the main effects of the three levels of *B*, each averaged over the levels of *A*. The contrast function coupled with the options method="pairwise" and adjust="tukey" requests tests including *p*-values for all pairwise comparisons via Tukey's method. Other adjustment options for pairwise comparisons include "Scheffe" for Scheffé's method, "Bonf" for the Bonferroni method, and "none" for individual inferences. The summary function with its infer=c(T,T) and level=0.99 options requests Tukey's 99% confidence intervals. Specifically, the option infer=c(T,T) indicates "true" for confidence intervals and tests, respectively. The above statement and the corresponding output are shown in the bottom of Table 6.18.

Implementation of Dunnett's method for all treatment-versus-control comparisons is similar and illustrated by the following statement

```
summary(contrast(lsmB, method="trt.vs.ctrl", adj="mvt", ref=1),
        infer=c(T,T), level=0.99)
```

Table 6.19 Dunnett's lower band for a two-way complete model with unequal sample sizes (reaction time experiment)

```
> # Dunnett's method comparing cell means
> lsmTC = lsmeans(modelTC, ~ fTC)
> summary(contrast(lsmTC, method="trt.vs.ctrl", adj="mvt"),
+          infer=c(T,T), level=0.99, side=">")

 contrast    estimate         SE df  lower.CL upper.CL t.ratio p.value
 12 - 11  -0.0083333 0.016379  8 -0.070333      Inf  -0.509  0.9351
 13 - 11   0.0150000 0.021974  8 -0.068181      Inf   0.683  0.5629
 21 - 11   0.0810000 0.017942  8  0.013083      Inf   4.515  0.0037
 22 - 11   0.0723333 0.016379  8  0.010334      Inf   4.416  0.0042
 23 - 11   0.0780000 0.016379  8  0.016000      Inf   4.762  0.0025

Confidence level used: 0.99
Confidence-level adjustment: mvt method for 5 estimates
P value adjustment: mvt method for 5 tests
P values are right-tailed
```

from the bottom of Table 6.15. The option `method="trt.vs.ctrl"` yields all treatment-versus-control comparisons. Here, other levels of B are compared to the first level, which happens to be `"1"`, averaging over levels of A. Available options include the following, as discussed in Sect. 4.7.2. Dunnett's method uses (simulation based) critical values from the multivariate t-distribution, corresponding to `adjust="mvt"`, but the default option `adjust="dunnettx"` provides an approximation of Dunnett's method for two-sided confidence intervals that runs faster and dependably (but is appropriate only applicable when the contrast estimates have pairwise correlations of 0.5). The first level of a factor is the control by default, corresponding to reference level 1 (`ref=1`), but one could, for example, specify the second level as the control by the syntax `ref=2`. Also, `"two-sided"` is the default for confidence intervals and tests, but one can specify `side="<"` for the one-sided alternative $H_A : \tau_i < \tau_1$ and the corresponding upper confidence bound for $\tau_i - \tau_1$, or `side=">"` for the alternative $H_A : \tau_i > \tau_1$ and the corresponding lower confidence bound for $\tau_i - \tau_1$.

Multiple comparisons of all treatments may be obtained using the cell-means model, as illustrated for Dunnett's method by the following code, reproduced with corresponding output in Table 6.19.

```
lsmTC = lsmeans(modelTC, ~ fTC)
summary(contrast(lsmTC, method="trt.vs.ctrl", adj="mvt"),
        infer=c(T,T), level=0.99, side=">")
```

The same could be accomplished using `lsmAB` instead of `lsmTC`. Note that the default control here is `"11"`, which is the first level of `fTC`.

We remind the reader that for unequal sample sizes, it has not yet been proved that the overall confidence levels achieved by the Tukey and Dunnett methods are at least as great as those stated, except in some special cases such as the one-way layout.

6.9.3   Plots

Residual plots for checking the error assumptions on the model are generated in the same way as shown in Chap. 5. If the two-way main-effects model (6.2.3) is used, the assumption of additivity should also be checked. For this purpose it is useful to plot the standardized residuals against the level of one factor, using the levels of the other factor for plotting labels (see, for example, Fig. 6.3, p. 144, for the temperature experiment). A plot of the standardized residuals z against the levels of factor A using the labels of factor B can be generated using the following R program lines:

```
plot(z ~ A, data=react.data, xaxt="n", type="n") # Suppress x-axis, pts
  axis(1, at=seq(1,2,1)) # Add x-axis with tick marks from 1 to 2 by 1
  text(z ~ A, B, cex=0.75, data=react.data) # Plot z vs A using B label
  mtext("B=1,2,3", side=3, adj=1, line=1) # Margin text, top-rt, line 1
  abline(h=0) # Horizontal line at zero
```

An interaction plot of treatment means $\bar{y}_{ij.}$ versus levels of factor A using the labels of factor B can be generated by adding the following statement to the end of the program in Table 6.15.

```
interaction.plot(x.factor = react.data$fA, trace.factor = react.data$fB,
                 response = react.data$y, type ="b",
                 xlab ="A", trace.label ="B", ylab ="Mean of y")
```

The option `type="b"` plots both points and lines.

In order to check for equal error variances, the residuals or observations could be plotted against the treatment combinations using the following R code:

```
plot(modelAB$res ~ react.data$TC, xlab ="AB", ylab ="Residual")
plot(react.data$y ~ react.data$TC, xlab ="AB", ylab ="y")
```

However, since the treatment combination codes `TC = 10*A + B` are numeric as originally created in Table 6.13, they will not be equally spaced along the axis, since the codes 11, 12, 13, 21, 22, 23 when regarded as 2-digit numbers are not equally spaced. One solution to this problem is to plot the residuals or the observations against the treatment variable `Trtmt`, since its levels 1–6 are equally spaced, but replace each of the labels 1–6 with the corresponding treatment combination label. This is accomplished by the following code.

```
plot(modelAB$res ~ react.data$Trtmt, xaxt="n", xlab="AB", ylab="Residual")
  axis(1, at = react.data$Trtmt, labels = react.data$fTC)
plot(react.data$y ~ react.data$Trtmt, xaxt="n", xlab="AB", ylab="y")
  axis(1, at = react.data$Trtmt, labels = react.data$fTC)
```

In the first `plot` statement, for example, the residuals are plotted against `Trtmt`, which has equally spaced levels 1–6, but the x-axis is suppressed by the option `xaxt="n"`. The `axis` statement is then used to create an x-axis, still with tick marks at the equally-spaced `Trtmt` levels 1–6, but using the treatment combination labels 11, 12,…, 23 of `fTC`.

When there are not sufficient observations to be able to check equality of error variances for all the cells, the standardized residuals should be plotted against the levels of each factor. The rule of thumb may be checked for the levels of each factor by comparing the maximum and minimum variances of the (nonstandardized) residuals. The sample variance of the residuals may be computed by level of A, for example, by augmenting the statements in Table 6.15 with the following `by` command.

```
by(modelAB$res, react.data$A, var)
```

Table 6.20 Fitting a model in terms of contrasts (air velocity experiment)

```
air.data = read.table("data/air.velocity.contrasts.txt", header=T)
air.data

        y A B Aln Aqd Bln Bqd Bcb Bqr Bqn
   1  -24 1 1  -1   1  -5   5  -5   1  -1
   2  -23 1 2  -1   1  -3  -1   7  -3   5
   3    1 1 3  -1   1  -1  -4   4   2 -10
   4    8 1 4  -1   1   1  -4  -4   2  10
   5   29 1 5  -1   1   3  -1  -7  -3  -5
   6   23 1 6  -1   1   5   5   5   1   1
   7   33 2 1   0  -2  -5   5  -5   1  -1
   8   28 2 2   0  -2  -3  -1   7  -3   5
   9   45 2 3   0  -2  -1  -4   4   2 -10
  10   57 2 4   0  -2   1  -4  -4   2  10
  11   74 2 5   0  -2   3  -1  -7  -3  -5
  12   80 2 6   0  -2   5   5   5   1   1
  13   37 3 1   1   1  -5   5  -5   1  -1
  14   79 3 2   1   1  -3  -1   7  -3   5
  15   79 3 3   1   1  -1  -4   4   2 -10
  16   95 3 4   1   1   1  -4  -4   2  10
  17  101 3 5   1   1   3  -1  -7  -3  -5
  18  111 3 6   1   1   5   5   5   1   1

# Fit linear regression model, save as model1
model1 = lm(y ~ Aln + Aqd + Bln + Bqd + Bcb + Bqr + Bqn
               + Aln:Bln + Aln:Bqd + Aln:Bcb + Aln:Bqr
               + Aqd:Bln + Aqd:Bqd + Aqd:Bcb, data=air.data)
# ANOVA
anova(model1)
```

6.9.4 One Observation Per Cell

In this section we present a direct way to split the interaction sum of squares into parts corresponding to negligible and nonnegligible orthogonal contrasts. The R program in Table 6.20 illustrates the method, using the data of the air velocity experiment.

First, the main effect contrast coefficients from Table 6.1, p. 148, are entered as part of the data, as is evident from the displayed data in Table 6.20. In the air velocity experiment, factor *A* had $a = 3$ levels and factor *B* had $b = 6$ levels.

The lm statement in Table 6.20 fits a linear model, including as predictors those contrasts that have not been declared to be negligible. For example, Aln and Bln terms contain the coefficients of the *A*-linear and *B*-linear contrasts, respectively, and the term Aln:Bln contains the coefficients of the *A*-linear-by-*B*-linear interaction trend contrast, which R obtains by multiplication of the corresponding Aln and Bln contrast coefficients. The results are saved as model1.

Intentionally, none of the predictor variables in the model are factor variables. If they were factor variables, then Aln would be interpreted as one factor with three coded levels −1, 0, 1, and Aqd as a second factor with two coded levels 1, −2, and so on. The error sum of squares will be based on the

Table 6.21 Data (beats per 15 seconds) for the weight lifting experiment

A	2	1	1	1	2	2	1	2	1	2	1
B	2	1	3	1	2	3	3	2	2	2	1
Rate	31	27	37	28	32	32	35	30	32	31	27
A	2	1	1	1	2	1	1	2	2	2	2
B	2	3	3	2	1	1	3	1	3	3	3
Rate	34	33	34	31	26	25	35	24	33	31	36
A	1	1	1	1	1	1	1	1	2	1	2
B	3	1	1	2	1	2	2	3	3	2	1
Rate	36	27	30	33	29	32	34	37	32	34	27
A	2	1	1	2	2	1	1	2	2	1	2
B	2	1	3	1	2	1	2	1	2	1	3
Rate	31	27	38	27	30	29	34	25	34	28	34
A	2	1	2	1	1	2	1	1	2	2	2
B	2	1	3	2	2	1	3	2	1	1	1
Rate	31	30	34	35	34	24	35	31	27	26	25
A	2	2	2	1	2	2	2	2	2	1	1
B	2	3	1	2	1	2	3	3	3	3	3
Rate	32	35	24	33	23	30	34	32	33	37	38

three omitted contrasts `Aln:Bqn`, `Aqd:Bqr`, and `Aqd:Bqn`, and the resulting analysis of variance table generated by the `anova(model1)` statement will be equivalent to that in Table 6.12, p. 176.

It is not necessary to input the levels of A and B separately as we have done in columns 2 and 3 of the data, but these would be needed if plots of the data were required.

Exercises

1. Under what circumstances should the two-way main effects model (6.2.3) be used rather than the two-way complete model (6.2.2)? Discuss the interpretation of main effects in each model.

2. Verify that $(\tau_{ij} - \overline{\tau}_{i.} - \overline{\tau}_{.j} + \overline{\tau}_{..})$ is an interaction contrast for the two-way complete model. Write down the list of contrast coefficients in terms of the τ_{ij}'s when factor A has $a = 3$ levels and factor B has $b = 4$ levels.

3. Consider the functions $\{\alpha_1^* - \alpha_2^*\}$ and $\{(\alpha\beta)_{11} - (\alpha\beta)_{21} - (\alpha\beta)_{12} + (\alpha\beta)_{22}\}$ under the two-way complete model (6.2.2).

   (a) Verify that the functions are estimable contrasts.
   (b) Discuss the meaning of each of these contrasts for plot (d) of Fig. 6.1, p. 140, and for plot (g) of Fig. 6.2, p. 141.
   (c) If $a = b = 3$, give the list of contrast coefficients for each contrast, first for the parameters involved in the contrast, and then in terms of the parameters τ_{ij} of the equivalent cell-means model.

4. Show that when the parentheses is expanded in formula (6.4.15) for *ssE* on p. 151, the computational formula (6.4.16) is obtained.

5. **Weight Lifting Experiment** (Gary Mirka 1986)
   The experimenter was interested in the effect on pulse rate (heart rate) of lifting different weights with legs either straight or bent (factor *A*, coded 1, 2). The selected weights were 50 lb, 75 lb, 100 lb (factor *B*, coded 1, 2, 3). He expected to see a higher pulse rate when heavier weights were lifted. He also expected that lifting with legs bent would result in a higher pulse rate than lifting with legs straight.

   (a) Write out a detailed checklist for running an experiment similar to this. In the calculation of the number of observations needed, either run your own pilot experiment or use the information that for a single subject in the above study, the error sum of squares was $ssE = 130.909$ bpfs^2 based on $df = 60$ error degrees of freedom (where bpfs is beats per 15 seconds).
   (b) The data collected for a single subject by the above experimenter are shown in Table 6.21 in the order collected. The experimenter wanted to use a two-way complete model. Check the assumptions on this model, paying particular attention to the facts that (i) these are count data and may not be approximately normally distributed, and (ii) the measurements were made in groups of ten at a time in order to reduce the fatigue of the subject.
   (c) Taking account of your answer to part (a), analyze the experiment, especially noting any trends in the response.

6. **Battery experiment, continued**

   Consider the battery experiment introduced in Sect. 2.5.2, p. 24, for which $a = b = 2$ and $r = 4$. Suppose it is of interest to calculate confidence intervals for the four simple effects $\tau_{11} - \tau_{12}$, $\tau_{21} - \tau_{22}$, $\tau_{11} - \tau_{21}$, $\tau_{12} - \tau_{22}$, with an overall confidence level of 95%.

   (a) Determine whether the Tukey or Bonferroni method of multiple comparisons would provide shorter confidence intervals.
   (b) Apply the better method from part (a) and comment on the results. (The data give $\bar{y}_{11.} = 570.75$, $\bar{y}_{12.} = 860.50$, $\bar{y}_{21.} = 433.00$, and $\bar{y}_{22.} = 496.25$ minutes per unit cost and $msE = 2{,}367.71$.)
   (c) Discuss the practical meaning of the contrasts estimated in (b) and explain what you have learned from the confidence intervals.

7. **Weld strength experiment**

   The data shown in Table 6.22 are a subset of the data given by Anderson and McLean (1974) and show the strength of a weld in a steel bar. Two factors of interest were gage bar setting (the distance the weld die travels during the automatic weld cycle) and time of welding (total time of the automatic weld cycle). Assume that the levels of both factors were selected to be equally spaced.

   (a) Using the cell-means model (6.2.1) for these data, test the hypothesis that there is no difference in the effects of the treatment combinations on weld strength against the alternative hypothesis that at least two treatment combinations have different effects.
   (b) Suppose the experimenters had planned to calculate confidence intervals for all pairwise comparisons between the treatment combinations, and also to look at the confidence interval for the difference between gage bar setting 3 and the average of the other two. Write down the contrasts

Table 6.22 Strength of weld

	i	Time of welding (j)				
		1	2	3	4	5
Gage	1	10, 12	13, 17	21, 30	18, 16	17, 21
bar	2	15, 19	14, 12	30, 38	15, 11	14, 12
setting	3	10, 8	12, 9	10, 5	14, 15	19, 11

Source Reprinted from Anderson and McLean (1974), pp. 62–63, by courtesy of Marcel Dekker, Inc

in terms of the parameters τ_{ij} of the cell-means model, and suggest a strategy for calculating all intervals at overall level "at least 98%."

(c) Consider the intervals in part (b). Give the formulae and calculate the actual interval for $\tau_{13} - \tau_{15}$ (the difference in the true mean strengths at the 3rd and 5th times of welding for the first gage bar setting), and explain what this interval tells you. Also calculate the actual interval for the difference between gage bar setting 3 and the average of the other two, and explain what this interval tells you.

(d) Calculate an upper 90% confidence limit for σ^2.

(e) If the experimenters were to repeat this experiment and needed the pairwise comparison intervals in (b) to be of width at most 8, how many observations should they take on each treatment combination? How many observations is this in total?

8. **Weld strength experiment, continued**

   For the experiment described in Exercise 7, use the two-way complete model instead of the equivalent cell means model.

   (a) Test the hypothesis of no interaction between gage bar setting and time of weld and state your conclusion.

   (b) Draw an interaction plot for the two factors Gage bar setting and Time of welding. Does your interaction plot support the conclusion of your hypothesis test? Explain.

   (c) In view of your answer to part (b), is it sensible to investigate the differences between the effects of gage bar setting? Why or why not? Indicate on your plot what would be compared.

   (d) Regardless of your answer to (c), suppose the experimenters had decided to look at the linear trend in the effect of gage bar settings. Test the hypothesis that the linear trend in gage setting is negligible (against the alternative hypothesis that it is not negligible).

9. **Sample size calculation**

   An experiment is to be run to examine three levels of factor A and four levels of factor B, using the two-way complete model (6.2.2). Determine the required sample size if the error variance σ^2 is expected to be less than 15 and simultaneous 99% confidence intervals for pairwise comparisons between treatment combinations should have length at most 10 to be useful.

10. **Bleach experiment, continued**

    Use the data of the bleach experiment of Example 6.4.4, on p. 154.

    (a) Evaluate the effectiveness of a variance-equalizing transformation.

    (b) Apply Satterthwaite's approximation to obtain 99% confidence intervals for the pairwise comparisons of the main effects of factor A using Tukey's method of multiple comparisons.

11. **Bleach experiment, continued**

The experimenter calculated that she needed $r = 5$ observations per treatment combination in order to be able to detect a difference in the effect of the levels of either treatment factor of 5 minutes (300 seconds) with probability 0.9 at significance level 0.05. Verify that her calculations were correct. She obtained a mean squared error of 43220.8 in her pilot experiment.

12. **Memory experiment** (James Bost 1987)

The memory experiment was planned in order to examine the effects of external distractions on short-term memory and also to examine whether some types of words were easier to memorize than others. Consequently, the experiment involved two treatment factors, "word type" and "type of distraction." The experimenter selected three levels for each factor. The levels of "word type" were

Level 1   (fruit): words representing fruits and vegetables commonly consumed;
Level 2   (nouns): words selected at random from Webster's pocket dictionary, representing tangible (i.e., visualizable) items;
Level 3   (mixed): words of any description selected at random from Webster's pocket dictionary.

A list of 30 words was prepared for each level of the treatment factor, and the list was not altered throughout the experiment.
The levels of "type of distraction" were

Level 1   : No distraction other than usual background noise;
Level 2   : Constant distraction, supplied by a regular banging of a metal spoon on a metal pan;
Level 3   : Changing distraction, which included vocal, music, banging and motor noise, and varying lighting.

The response variable was the number of words remembered (by a randomly selected subject) for a given treatment combination. The response variable is likely to have approximately a binomial distribution, with variance $30_p(1 - p)$ where p is the probability that a subject remembers a given word and 30 is the number of words on the list. It is unlikely that p is constant for all treatment combinations or for all subjects. However, since $np(1 - p)$ is less than $30(0.5)(0.5) = 7.5$, a reasonable guess for the variance σ^2 is that it is less than 7.5.
The experimenter wanted to reject each of the main-effect hypotheses H_0^A:{the memorization rate for the three word lists is the same} and H_0^B:{the three types of distraction have the same effect on memorization} with probability 0.9 if there was a difference of four words in memorization rates between any two word lists or any two distractions (that is $\Delta_A = \Delta_B = 4$), using a significance level of $\alpha = 0.05$. Calculate the number of subjects that are needed if each subject is to be assigned to just one treatment combination and measured just once.

13. **Memory experiment, continued**

(a) Write out a checklist for the memory experiment of Exercise 12. Discuss how you would obtain the subjects and how applicable the experiment would be to the general population.
(b) Consider the possibility of using each subject more than once (i.e., consider the use of a blocking factor). Discuss whether or not an assumption of independent observations is likely to be valid.

14. **Memory experiment, continued**

The data for the memory experiment of Exercise 12 are shown in Table 6.23 with three observations per treatment combination.

Table 6.23 Data and standardized residuals for the memory experiment

Word list	Distraction								
	None			Constant			Changing		
Fruit	20	14	24	15	22	17	17	13	12
	0.27	−2.16	1.89	−1.21	1.62	−0.40	1.21	−0.40	−0.81
Nouns	19	14	19	12	11	14	12	15	8
	0.67	−1.35	0.67	−0.13	−0.54	0.67	0.13	1.35	−1.48
Mixed	11	12	15	8	8	9	12	7	10
	−0.67	−0.27	0.94	−0.13	−0.13	0.27	0.94	−1.08	0.13

(a) The experimenter intended to use the two-way complete model. Check the assumptions on the model for the given data, especially the equal-variance assumption.

(b) Analyze the experiment. A transformation of the data or use of the Satterthwaite approximation may be necessary.

15. **Ink experiment**

Teaching associates who give classes in computer labs at the Ohio State University are required to write on white boards with "dry markers" rather than on chalk boards with chalk. The ink from these dry markers can stain rather badly, and an experiment was planned by M. Chambers, Y.-W. Chen, E. Kurali and R. Vengurlekar in 1996 to determine which type of cloth (factor A, 1 = cotton/polyester, 2 = rayon, 3 = polyester) was most difficult to clean, and whether a detergent plus stain remover was better than a detergent without stain remover (factor B, levels 1, 2) for washing out such a stain.

Pieces of cloth were to be stained with 0.1 ml of dry marker ink and allowed to air dry for 24 hours. The cloth pieces were then to be washed in a random order in the detergent to which they were allocated. The stain remaining on a piece of cloth after washing and drying was to be compared with a 19 point scale and scored accordingly, where 1 = black and 19 = white.

(a) Make a list of the difficulties that might be encountered in running and analyzing an experiment of this type. Give suggestions on how these difficulties might be overcome or their effects reduced.

(b) Why should each piece of cloth be washed separately? (Hint: think about the error variability.)

(c) The results of a small pilot study run by the four experimenters are shown in Table 6.24. Plot the data against the levels of the two treatment factors. Can you learn anything from this plot? Which model would you select for the main experiment? Why?

(d) Calculate the number of observations that you would need to take on each treatment combination in order to try to ensure that the lengths of confidence intervals for pairwise differences in the effects of the levels of each of the factors were no more than 2 points (on the 19-point scale).

Table 6.24 Data for the ink experiment in the order of collection

Cloth type	3	1	3	1	2	1	2	2	2	3	3	1
Stain remover	2	2	2	2	1	1	1	2	2	1	1	1
Stain score	1	6	1	5	11	9	9	8	6	3	4	8

Table 6.25 Data for the survival experiment (units of 10 hours)

Poison	Treatment			
	1	2	3	4
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

Source Box and Cox (1964). Copyright 1964 Blackwell Publishers. Reprinted with permission

16. **Survival experiment** (G.E.P. Box and D.R. Cox, 1964)

    The data in Table 6.25 show survival times of animals to whom a poison and a treatment have been administered. The data were presented by G. E. P. Box and D. R. Cox in an article in the *Journal of the Royal Statistical Society* in 1964. There were three poisons (factor A at $a = 3$ levels), four treatments (factor B at $b = 4$ levels), and $r = 4$ animals (experimental units) assigned at random to each treatment combination.

    (a) Check the assumptions on a two-way complete model for these data. If the assumptions are satisfied, then analyze the data and discuss your conclusions.
    (b) Take a reciprocal transformation (y^{-1}) of the data. The transformed data values then represent "rates of dying." Check the assumptions on the model again. If the assumptions are satisfied, then analyze the data and discuss your conclusions.
    (c) Draw an interaction plot for both the original and the transformed data. Discuss the interaction between the two factors in each of the measurement scales.

17. Use the two-way main-effects model (6.2.3) with $a = b = 3$.

    (a) Which of the following are estimable?

       (i) $\mu + \alpha_1 + \beta_2$.
       (ii) $\mu + \alpha_1 + \frac{1}{2}(\beta_1 + \beta_2)$.
       (iii) $\beta_1 - \frac{1}{3}(\beta_2 + \beta_3)$.

    (b) Show that $\overline{Y}_{i..} + \overline{Y}_{.j.} - \overline{Y}_{...}$ is an unbiased estimator of $\mu + \alpha_i + \beta_j$ with variance, $\sigma^2(a+b-1)/(abr)$.
    (c) Show that $\sum_i c_i \overline{Y}_{i..}$ is an unbiased estimator of the contrast $\sum_i c_i \alpha_i$.

18. **Meat cooking experiment, continued**

    The meat cooking experiment was introduced in Exercise 14 of Chap. 3, with the data given in Table 3.14, p. 68.

Table 6.26 Data for the water boiling experiment, in minutes. (Order of observation is in parentheses.)

| Burner | Salt (teaspoons) | | | |
	0	2	4	6
Right back	7(7)	4(13)	7(24)	5(15)
	8(21)	7(25)	7(34)	7(33)
	7(30)	7(26)	7(41)	7(37)
Right front	4(6)	4(36)	4(1)	4(28)
	4(20)	5(44)	4(14)	4(31)
	4(27)	4(45)	5(18)	4(38)
Left back	6(9)	6(46)	7(8)	5(35)
	7(16)	6(47)	6(12)	6(39)
	6(22)	5(48)	7(43)	6(40)
Left front	9(29)	8(5)	8(3)	8(2)
	9(32)	8(10)	9(19)	8(4)
	9(42)	8(11)	10(23)	7(17)

(a) Using the two-way complete model, conduct an analysis of variance, testing each hypothesis using a 1% significance level. State your conclusions.

(b) Draw an interaction plot for the two treatment factors. Does your interaction plot support the conclusion of your hypothesis test concerning interactions? Explain.

(c) Compare the effects of the three levels of fat content pairwise, averaging over cooking methods, using Scheffé's method for all treatment contrasts with a 95% confidence level. Interpret the results.

(d) Give a confidence interval for the average difference in weight after cooking between frying and grilling 110 g hamburgers, using Scheffé's method for all treatment contrasts with a 95% confidence level. Interpret the results.

(e) Obtain a 95% confidence interval for comparing the effect on post-cooked weight of the low fat content versus the average of the two higher fat contents (averaged over cooking method), using Scheffé's method.

(f) What is the overall confidence level of the intervals in parts (c), (d) and (e) taken together?

(g) If the contrast in part (e) had been the *only* contrast of interest, would your answer to part (e) have been different? If so, show the new calculation. If not, explain why not.

19. For the two-way main-effects model (6.2.3) with equal sample sizes,

(a) verify the computational formulae for ssE given in (6.5.38),

(b) and, if SSE is the corresponding random variable, show that $E[SSE]$ is $(n - a - b + 1)\sigma^2$. [Hint: $E[X^2] = \text{Var}(X) + E[X]^2$.]

20. An experiment is to be run to compare the two levels of factor A and to examine the pairwise differences between the four levels of factor B, with a simultaneous confidence level of 90%. The experimenter is confident that the two factors will not interact. Find the required sample size if the error variance will be at most 25 and the confidence intervals should have length at most 10 to be useful.

21. **Water boiling experiment** (Kate Ellis 1986)

The experiment was run in order to examine the amount of time taken to boil a given amount of water on the four different burners of her stove, and with 0, 2, 4, or 6 teaspoons of salt added to the water. Thus the experiment had two treatment factors with four levels each. The experimenter ran the experiment as a completely randomized design by taking $r = 3$ observations on each of the 16 treatment combinations in a random order. The data are shown in Table 6.26. The experimenter believed that there would be no interaction between the two factors.

(a) Check the assumptions on the two-way main-effects model.
(b) Calculate a 99% set of Tukey confidence intervals for pairwise differences between the levels of salt, and calculate separately a 99% set of intervals for pairwise differences between the levels of burner.
(c) Test a hypothesis that there is no linear trend in the time to boil water due to the level of salt. Do a similar test for a quadratic trend.
(d) The experimenter believed that observation number 13 was an outlier, since it has a large standardized residual and it was an observation taken late on a Friday evening. Using statistical software, repeat the analysis in (b) removing this observation. (Tukey's method is approximate for nearly balanced data.) Also repeat the test in part (c) but for the linear contrast only. (The formula for the linear contrast coefficients is given in (4.2.4) on p. 73.) Do you prefer the analysis that uses all the data, or that which removes observation 13? Explain your choice.

22. For $v = 5$ and $r = 4$, show that the first three "orthogonal polynomial contrasts" listed in Table A.2 are mutually orthogonal. (In fact all four are.) Find a pair of orthogonal contrasts that are not orthogonal polynomial contrasts. Can you find a third contrast that is orthogonal to each of these? How about a fourth? (This gets progressively harder!)

23. **Air velocity experiment, continued**

(a) For the air velocity experiment introduced in Sect. 6.7.4 (p. 176), calculate the sum of squares for each of the three interaction contrasts assumed to be negligible, and verify that these add to the value $ssE = 175.739$, as in Table 6.12.
(b) Check the assumptions on the model by plotting the standardized residuals against the predicted responses, the treatment factor levels, and the normal scores. State your conclusions.

Several Crossed Treatment Factors 7

7.1 Introduction

Experiments that involve more than two treatment factors are designed and analyzed using many of the same principles that were discussed in Chap. 6 for two-factor experiments. We continue to label the factors with uppercase Latin letters and their numbers of levels with the corresponding lowercase letters. An experiment that involves four factors, A, B, C, and D, having a, b, c, and d levels, respectively, for example, is known as an "$a \times b \times c \times d$ *factorial experiment*" (read "a by b by c by d") and has a total of $v = abcd$ treatment combinations.

There are several different models that may be appropriate for analyzing a factorial experiment with several treatment factors, depending on which interactions are believed to be negligible. These models, together with definitions of interaction between three or more factors, and estimation of contrasts, form the topic of Sect. 7.2. General rules are given in Sect. 7.3 for writing down confidence intervals and hypothesis tests when there are equal numbers of observations on all treatment combinations. In Sect. 7.5, methods are investigated for analyzing small experiments where there is only one observation per treatment combination. Finally, SAS and R commands for analyzing experiments with several treatment factors are given in Sects. 7.6 and 7.7, respectively, and can be used for unequal sample sizes. Problems caused by empty cells are also investigated.

7.2 Models and Factorial Effects

7.2.1 Models

One of a number of different models may be appropriate for describing the data from an experiment with several treatment factors. The selection of a suitable model prior to the experiment depends upon available knowledge about which factors do and do not interact. We take as an example an experiment with three factors. Our first option is to use the *cell-means model*, which is similar to the one-way analysis of variance model (3.3.1), p. 33. For example, the cell-means model for three treatment factors is

$$
\begin{aligned}
Y_{ijkt} &= \mu + \tau_{ijk} + \epsilon_{ijkt}\,, \\
\epsilon_{ijkt} &\sim N(0, \sigma^2)\,, \\
\epsilon_{ijkt} &\text{'s mutually independent}\,, \\
t = 1, \ldots, r_{ijk}; \quad i = 1, \ldots, a; \quad & j = 1, \ldots, b; \quad k = 1, \ldots, c\,.
\end{aligned}
\tag{7.2.1}
$$

If there are more than three factors, the cell-means model has correspondingly more subscripts. As in Chap. 6, use of this model allows all of the formulae presented in Chaps. 3 and 4 for one treatment factor to be used to compare the effects of the treatment combinations.

Alternatively, we can model the effect on the response of treatment combination ijk to be

$$\tau_{ijk} = \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk},$$

where α_i, β_j, γ_k are the effects (positive or negative) on the response of factors A, B, C at levels i, j, k, respectively, $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, and $(\beta\gamma)_{jk}$ are the additional effects of the pairs of factors together at the specified levels, and $(\alpha\beta\gamma)_{ijk}$ is the additional effect of all three factors together at levels i, j, k. The three sets of factorial effects are called the main-effect parameters, the two-factor interaction parameters, and the three-factor interaction parameter, respectively. The interpretation of a three-factor interaction is discussed in the next section. If we replace τ_{ijk} in model (7.2.1) by the main-effect and interaction parameters, we obtain the equivalent *three-way complete model*; that is,

$$
\begin{aligned}
Y_{ijkt} &= \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkt}, \\
&\quad \epsilon_{ijkt} \sim N(0, \sigma^2), \\
&\quad \epsilon_{ijkt}\text{'s mutually independent}, \\
&t = 1, \ldots, r_{ijk}; \quad i = 1, \ldots, a; \quad j = 1, \ldots, b; \quad k = 1, \ldots, c.
\end{aligned}
\tag{7.2.2}
$$

This form of the model extends in an obvious way to more than three factors by including a main-effect parameter for every factor and an interaction effect parameter for every combination of two factors, three factors, etc.

If prior to the experiment certain interaction effects are known to be negligible, the corresponding parameters can be removed from the complete model to give a submodel. For example, if the factors A and B are known not to interact in a three-factor experiment, then the AB and ABC interaction effects are negligible, so the terms $(\alpha\beta)_{ij}$ and $(\alpha\beta\gamma)_{ijk}$ are excluded from model (7.2.2). In the extreme case, if no factors are expected to interact, then a *main-effects model* (which includes no interaction terms) can be used.

When a model includes an interaction between a specific set of m factors, then all interaction terms involving subsets of those m factors should be included in the model. For example, a model that includes the effect of the three-factor interaction ABC would also include the effects of the AB, AC, and BC interactions as well as the main effects A, B, and C.

Use of a submodel, when appropriate, is advantageous, because simpler models generally yield tighter confidence intervals and more powerful tests of hypotheses. However, if interaction terms are removed from the model when the factors do, in fact, interact, then the resulting analysis and conclusions may be totally incorrect.

7.2.2   The Meaning of Interaction

The same type of interaction plot as that used in Sect. 6.2.1, p. 139, can be used to evaluate interactions between pairs of factors in an experiment involving three or more factors. The graphical evaluation of three-factor interactions can be done by comparing separate interaction plots at the different levels of a third factor. Such plots will be illustrated for experiments that involve only three factors, but the methods are similar for experiments with four or more factors, except that the sample means being plotted would be averages over the levels of all the other factors.

Fig. 7.1 *AB*-interaction plot (averaged over levels of *C*)
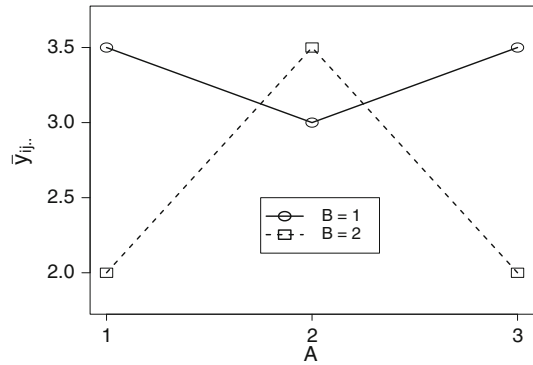
The following sample means are for a hypothetical $3 \times 2 \times 2$ experiment involving the factors A, B, and C at 3, 2, and 2 levels, respectively.

ijk	:	111	112	121	122	211	212	221	222	311	312	321	322
$\overline{y}_{ijk.}$	:	3.0	4.0	1.5	2.5	2.5	3.5	3.0	4.0	3.0	4.0	1.5	2.5

An *AB*-interaction plot for these hypothetical data is shown in Fig. 7.1. As in the previous chapter, we must remember that interaction plots give no indication of the size of the experimental error and must be interpreted with a little caution. The lines of the plot in Fig. 7.1 are not parallel, indicating that the factors possibly interact. For factor A, level 2 appears to be the best (highest response) on average, but not consistently the best at each level of B. Likewise, level 1 of factor B appears to be better on average, but not consistently better at each level of A. The perceived AB interaction is averaged over the levels of C and may have no practical meaning if there is an ABC interaction. Consequently, the three-factor interaction should be investigated first.

A three-factor interaction would be indicated if the interaction effect between any pair of factors were to change as the level of the third factor changes. In Fig. 7.2, a separate *AB*-interaction plot is shown for each level of factor C. Each of the two plots suggests the presence of an *AB*-interaction effect, but the patterns in the two plots are the same. In other words, the factors A and B apparently interact in the same way at each level of factor C. This indicates a negligible *ABC*-interaction effect. The shift in the interaction plot as the level of C changes from one plot to the other indicates a possible main effect of factor C. The *AB* interaction plot in Fig. 7.1 is the average of the two plots in Fig. 7.2, showing the *AB* interaction averaged over the two levels of C.

Other three-factor interaction plots can be obtained by interchanging the roles of the factors. For example, Fig. 7.3 contains plots of $\overline{y}_{ijk.}$ against the levels i of A for each level j of factor B, using the levels k of C as labels and the same hypothetical data. Lines are parallel in each plot, indicating no *AC*-interaction at either level of B. Although the patterns differ from plot to plot, if there is no *AC*-interaction at either of the levels of B, there is no change in the *AC*-interaction from one level of B to the other. So, again the *ABC*-interaction effect appears to be negligible. An *AC* interaction plot would show the average of the two plots in Fig. 7.3, and although the plot would again look different, the lines would still be parallel.

To see what the plots might look like when there is an *ABC*-interaction present, we look at the following second set of hypothetical data.

ijk	:	111	112	121	122	211	212	221	222	311	312	321	322
$\overline{y}_{ijk.}$	:	3.0	2.0	1.5	4.0	2.5	3.5	3.0	4.0	3.0	5.0	3.5	6.0
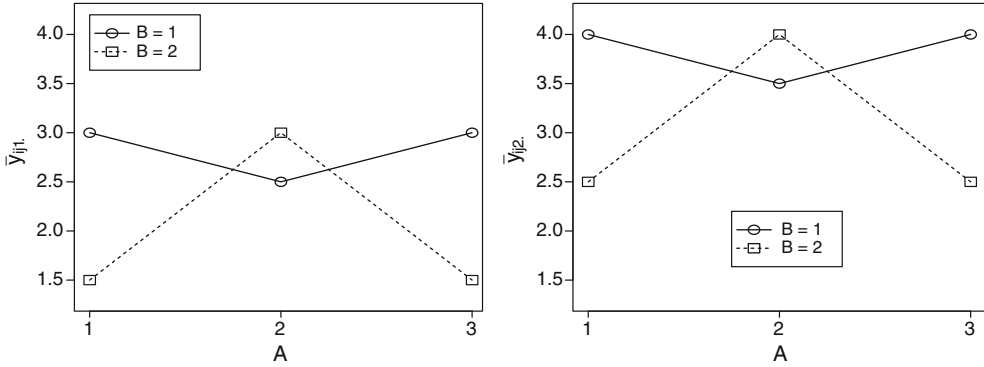
Fig. 7.2 *ABC*-interaction plots with *C* as the third factor

(a) Level 1 of *B*   (b) Level 2 of *B*
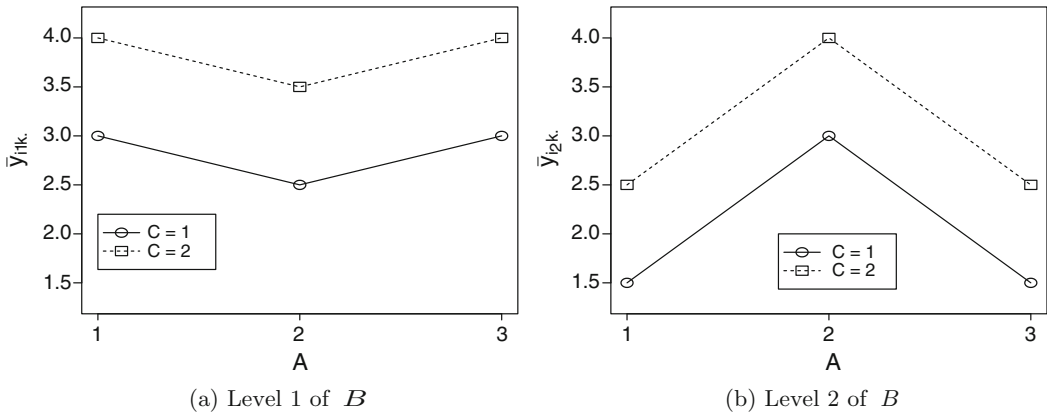
Fig. 7.3 *ABC*-interaction plots with *B* as the third factor

Figure 7.4 shows plots of $\bar{y}_{ijk.}$ against the level i of factor A for each level k of factor C, using the level j of factor B as the plotting label. In each plot, corresponding lines are not all parallel, *and* the pattern changes from one plot to the next. In other words, the interaction effect between factors A and B apparently changes with the level of C, so there appears to be an *ABC*-interaction effect.

Four-factor interactions can be evaluated graphically by comparing the pairs of plots representing a three-factor interaction for the different levels of a fourth factor. Clearly, higher-order interactions are harder to envisage than those of lower order, and we would usually rely solely on the analysis of variance table for evaluating the higher-order interactions. In general, one should examine the higher-order interactions first, and work downwards. In many experiments, high-order interactions do tend to be small, and lower-order interactions can then be interpreted more easily.

(a) Level 1 of C (b) Level 2 of C
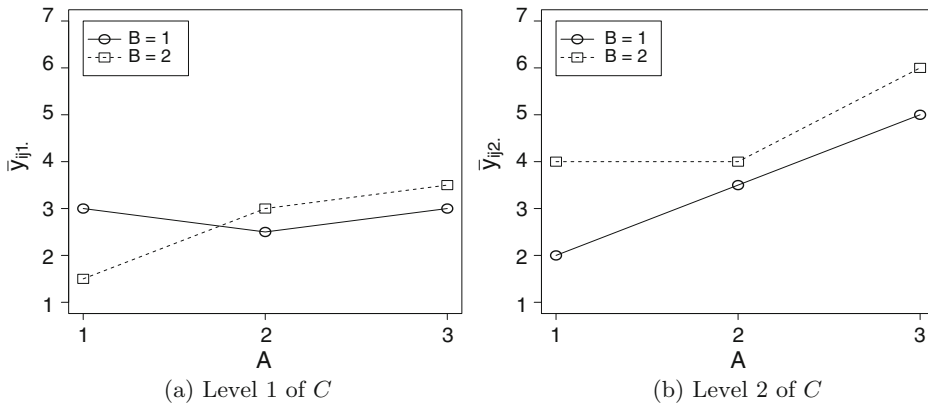
Fig. 7.4 *ABC*-interaction plots showing an *ABC*-interaction effect

7.2.3   Separability of Factorial Effects

In an experiment involving three factors, A, B, and C, for which it is known in advance that factor C will not interact with factor A or B, the AC, BC, and ABC interaction effects can be excluded from the model. Interpretation of the results of the experiment is simplified, because a specific change in the level of factor C has the same effect on the mean response for every combination ij of levels of A and B. Likewise, a specific change in the combination of levels of A and B has the same effect on the mean response regardless of the level of C. If the objective is to find the best treatment combination ijk, then the task is reduced to two smaller problems involving fewer comparisons, namely choice of the best combination ij of levels of A and B and, separately, choice of the best level k of C.

When there is such separability of effects, the experimenter should generally avoid the temptation to run separate experiments, one to determine the best combination ij of levels of factors A and B and another to determine the best level k of C. A single factorial experiment involving n observations provides the same amount of information on the A, B, C, and AB effects as would two separate experiments—a factorial experiment for factors A and B and another experiment for factor C—*each* involving n observations!

One way to determine an appropriate model for an experiment is as follows. Suppose that the experiment involves p factors. Draw p points, labeling one point for each factor (see, for example, Fig. 7.5 for four factors A–D). Connect each pair of factors that might conceivably interact with a line to give a *line graph*. For every pair of factors that are joined by a line in the line graph, a two-factor interaction should be included in the model. If three factors are joined by a triangle, then it may be appropriate to include the corresponding three-factor interaction in the model as well as the three two-factor interactions. Similarly, if four factors are joined by all six possible lines, it may be appropriate to include the corresponding four-factor interaction as well as the three-factor and two-factor interactions.

The line graphs in Fig. 7.5 fall into two pieces. Line graph (a) represents the situation where A and B are thought to interact, as are C and D. The model would include the AB and CD interaction effects, in addition to all main effects. Line graph (b) represents an experiment in which it is believed that A and B interact and also A and C and also B and C. An appropriate model would include all main effects and the AC, AB, and BC interactions. The three-factor ABC interaction effect might also

Fig. 7.5 Separability plots
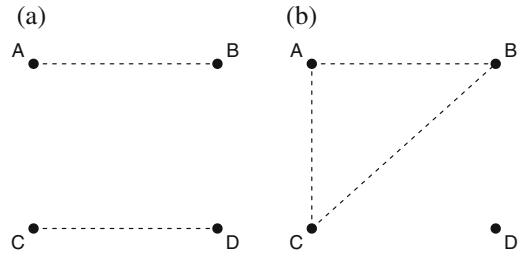
be included in the model depending upon the type of interaction plots expected by the experimenter. Thus, a possible model would be

$$Y_{ijklt} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij}$$
$$+ (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijklt} \, .$$

7.2.4   Estimation of Factorial Contrasts

For an $a \times b \times c$ factorial experiment and the three-way complete model, all treatment contrasts are of the form

$$\sum_i \sum_j \sum_k h_{ijk} \tau_{ijk} \quad \text{with} \quad \sum_i \sum_j \sum_k h_{ijk} = 0 \, ,$$

and are estimable when there is at least one observation per treatment combination.

A contrast in the main effect of A is any treatment contrast for which the coefficients h_{ijk} depend only on the level i of A. For example, if we set h_{ijk} equal to $p_i/(bc)$, with $\Sigma p_i = 0$, then the contrast $\Sigma\Sigma\Sigma h_{ijk}\tau_{ijk}$ becomes

$$\sum_i p_i \bar{\tau}_{i..} = \sum_i p_i[\alpha_i + \overline{(\alpha\beta)}_{i.} + \overline{(\alpha\gamma)}_{i.} + \overline{(\alpha\beta\gamma)}_{i..}] = \sum_i p_i \alpha_i^* \, .$$

We notice that a main-effect contrast for factor A can be interpreted only as an average over all of the interaction effects involving A in the model and, consequently, may not be of interest if any of these interactions are nonnegligible. The B and C main-effect contrasts are defined in similar ways.

An AB interaction contrast is any treatment contrast for which the coefficients h_{ijk} depend only on the combination ij of levels of A and B, say $h_{ijk} = d_{ij}/c$, and for which $\sum_i d_{ij} = 0$ for all j and $\sum_j d_{ij} = 0$ for all i. An AB interaction contrast can be expressed as

$$\sum_i \sum_j d_{ij} \bar{\tau}_{ij.} = \sum_i \sum_j d_{ij}[(\alpha\beta)_{ij} + \overline{(\alpha\beta\gamma)}_{ij.}] = \sum_i \sum_j d_{ij}(\alpha\beta)_{ij}^* \, .$$

Thus, the AB interaction contrast can be interpreted only as an average over the ABC interaction and may not be of interest if the ABC interaction is nonnegligible. The AC and BC interaction contrasts are defined in similar ways.

An ABC interaction contrast is any contrast of the form

$$\sum_i \sum_j \sum_k h_{ijk} \tau_{ijk} = \sum_i \sum_j \sum_k h_{ijk} (\alpha\beta\gamma)_{ijk}$$

for which $\sum_i h_{ijk} = 0$ for all jk, $\sum_j h_{ijk} = 0$ for all ik, and $\sum_k h_{ijk} = 0$ for all ij. When we investigated the interaction plot for ABC using Figs. 7.2–7.4, we compared the AB interaction at two different levels of factor C. In other words, we looked at contrasts of the type

$$(\tau_{112} - \tau_{122} - \tau_{212} + \tau_{222}) - (\tau_{111} - \tau_{121} - \tau_{211} + \tau_{221}).$$

If the levels 1 and 2 of A and B interact in the same way at each level of C, then this ABC interaction contrast is zero. If all interaction contrasts of this type (for all levels of A, B, and C) are zero, then the ABC interaction is negligible.

When a sub-model, rather than a complete model, is used, parameters for the negligible interactions are removed from the above expressions. If the experiment involves more than three factors, then the above definitions can be generalized by including the additional subscripts on $h_{ijk}\tau_{ijk}$ and averaging over the appropriate higher-order interactions.

As in Chap. 6, all contrasts can be represented by coefficient lists in terms of the main-effect and interaction parameters or in terms of the treatment combination parameters. This is illustrated in the next example.

Example 7.2.1 Coefficient lists for contrasts

Suppose that we have an experiment that involves four factors, A, B, C, and D, each to be examined at two levels (so that $a = b = c = d = 2$ and $v = 16$). Suppose that a model that includes AB, BC, BD, CD, and BCD interactions is expected to provide a good description of the data; that is,

$$\begin{aligned} Y_{ijklt} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} \\ + (\beta\gamma)_{jk} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} + (\beta\gamma\delta)_{jkl} + \epsilon_{ijklt}, \\ \epsilon_{ijklt} \sim N(0, \sigma^2), \end{aligned}$$

ϵ_{ijklt}'s are mutually independent,

$$t = 1, \ldots, r_{ijkl}, \quad i = 1, 2, \quad j = 1, 2, \quad k = 1, 2, \quad l = 1, 2.$$

The contrast that compares the two levels of C is

$$\overline{\tau}_{..2.} - \overline{\tau}_{..1.} = \gamma_2^* - \gamma_1^*,$$

where $\gamma_k^* = \gamma_k + \overline{(\beta\gamma)}_{.k} + \overline{(\gamma\delta)}_{k.} + \overline{(\beta\gamma\delta)}_{.k.}$. This contrast can be represented as a coefficient list $[-1, \ 1\ ]$ in terms of the parameters γ_1^* and γ_2^* or as

$$\frac{1}{8}[-1, -1, \ 1, \ 1, -1, -1, \ 1, \ 1, -1, -1, \ 1, \ 1, -1, -1, \ 1, \ 1\ ]$$

in terms of the τ_{ijkl}. These are listed under the heading C in Table 7.1.Coefficient lists for the other main-effect contrasts in terms of the τ_{ijkl} are also shown in Table 7.1. The treatment combinations in the table are listed in ascending order when regarded as 4-digit numbers. The main-effect contrast coefficients are -1 when the corresponding factor is at level 1, and the coefficients are $+1$ when the corresponding factor is at level 2, although these can be interchanged if contrasts such as $\gamma_1^* - \gamma_2^*$ are

Table 7.1 Contrast coefficient lists in terms of treatment combination parameters

Treatment combination	A	B	C	D	AB	BC	BD	CD	BCD
1111	-1	-1	-1	-1	1	1	1	1	-1
1112	-1	-1	-1	1	1	1	-1	-1	1
1121	-1	-1	1	-1	1	-1	1	-1	1
1122	-1	-1	1	1	1	-1	-1	1	-1
1211	-1	1	-1	-1	-1	-1	-1	1	1
1212	-1	1	-1	1	-1	-1	1	-1	-1
1221	-1	1	1	-1	-1	1	-1	-1	-1
1222	-1	1	1	1	-1	1	1	1	1
2111	1	-1	-1	-1	-1	1	1	1	-1
2112	1	-1	-1	1	-1	1	-1	-1	1
2121	1	-1	1	-1	-1	-1	1	-1	1
2122	1	-1	1	1	-1	-1	-1	1	-1
2211	1	1	-1	-1	1	-1	-1	1	1
2212	1	1	-1	1	1	-1	1	-1	-1
2221	1	1	1	-1	1	1	-1	-1	-1
2222	1	1	1	1	1	1	1	1	1
Divisor	8	8	8	8	4	4	4	4	2

required rather than $\gamma_2^* - \gamma_1^*$. The divisor shown in the table is the number of observations taken on each level of the factor.

The two-factor interaction contrast for CD is

$$(\gamma\delta)_{11}^* - (\gamma\delta)_{12}^* - (\gamma\delta)_{21}^* + (\gamma\delta)_{22}^*,$$

where $(\gamma\delta)_{kl}^* = (\gamma\delta)_{kl} + \overline{(\beta\gamma\delta)}_{.kl}$. This has coefficient list $[\,1, -1, -1, \ 1\,]$ in terms of the interaction parameters $(\gamma\delta)_{kl}^*$ but has coefficient list

$$\frac{1}{4}[\,1, -1, -1, \ 1, \ 1, -1, -1, \ 1, \ 1, -1, -1, \ 1, \ 1, -1, -1, \ 1\,]$$

in terms of the treatment combination parameters τ_{ijkl}. The coefficients are $+1$ when C and D are at the same level and -1 when they are at different levels. Notice that these coefficients can easily be obtained by multiplying together the C and D coefficients in the same rows of Table 7.1. The coefficient lists for some of the other interaction contrasts are also shown in the table, and it can be verified that their coefficients are also products of the corresponding main-effect coefficients. The divisors are the numbers of observations on each pair of levels of C and D. To obtain the same precision (estimator variance) as a main effect contrast, the divisor would need to be changed to 8 (or all contrasts would need to be normalized).

Contrast coefficients are also shown for the BCD interaction. These are the products of the main-effect coefficients for B, C, and D. This contrast compares the CD interaction at the two levels of B (or, equivalently, the BC interaction at the two levels of D, or the BD interaction at the two levels of C). □

7.3 Analysis—Equal Sample Sizes

For an experiment involving p factors, we can select a cell-means model or the equivalent p-way complete model, or any of the possible submodels. When the sample sizes are equal, the formulae for the degrees of freedom, least squares estimates, and sums of squares for testing hypotheses follow well-defined patterns. We saw in Chap. 6 that for an experiment with two factors, we obtain similar formulae for the least squares estimates of the contrasts $\sum c_i \alpha_i$ in the two-way main-effects model and $\sum c_i \alpha_i^*$ in the two-way complete model. Similarly, the sum of squares for A was of the same form in both cases. This is also true for experiments with more than two factors.

We now give a series of rules that can be applied to any complete model with equal sample sizes. The rules are illustrated for the ABD interaction in an experiment involving four treatment factors A, B, C, and D, with corresponding symbols α, β, γ, and δ and subscripts i, j, k, and l to represent their effects on the response in a four-way complete model with r observations per treatment combination. The corresponding rules for submodels are obtained by dropping the relevant interaction terms from the complete model. When the sample sizes are not equal, the formulae are more complicated, and we will analyze such experiments only via a computer package (see Sects. 7.6 and 7.7).

Rules for Estimation and Hypothesis Testing—Equal Sample Sizes

1. Write down the name of the main effect or interaction of interest and the corresponding numbers of levels and subscripts.

    Example: ABD; numbers of levels a, b, and d; subscripts i, j, and l.

2. The number of degrees of freedom ν for a factorial effect is the product of the "number of levels minus one" for each of the factors included in the effect.

    Example: For ABD, $\nu = (a-1)(b-1)(d-1)$.

3. Multiply out the number of degrees of freedom and replace each letter with the corresponding subscript.

    Example: For ABD, $df = abd - ab - ad - bd + a + b + d - 1$, which gives $ijl - ij - il - jl + i + j + l - 1$.

4. The sum of squares for testing the hypothesis that a main effect or an interaction is negligible is obtained as follows. Use each group of subscripts in rule 3 as the subscripts of a term $\overline{y}$, averaging over all subscripts not present and keeping the same signs. Put the resulting estimate in brackets, square it, and sum over all possible subscripts. To expand the parentheses, square each term in the parentheses, keep the same signs, and sum over all possible subscripts.

Example:

$$ss(ABD) = rc \sum_i \sum_j \sum_l (\bar{y}_{ij.l.} - \bar{y}_{ij...} - \bar{y}_{i..l.} - \bar{y}_{.j.l.}$$

$$+ \bar{y}_{i....} + \bar{y}_{.j...} + \bar{y}_{...l.} - \bar{y}_{.....})^2$$

$$= rc \sum_i \sum_j \sum_l \bar{y}_{ij.l.}^2 - rcd \sum_i \sum_j \bar{y}_{ij...}^2 - rbc \sum_i \sum_l \bar{y}_{i..l.}^2$$

$$- rac \sum_j \sum_l \bar{y}_{.j.l.}^2 + rbcd \sum_i \bar{y}_{i....}^2 + racd \sum_j \bar{y}_{.j...}^2$$

$$+ rabc \sum_l \bar{y}_{...l.}^2 - rabcd \bar{y}_{.....}^2 .$$

5. The total sum of squares *sstot* is the sum of all the squared deviations of the data values from their overall mean. The total degrees of freedom is $n - 1$, where n is the total number of observations.

$$\text{Example:} \quad sstot = \sum_i \sum_j \sum_k \sum_l \sum_t (y_{ijklt} - \bar{y}_{.....})^2$$

$$= \sum_i \sum_j \sum_k \sum_l \sum_t y_{ijklt}^2 - n\bar{y}_{.....}^2 ,$$

$$n - 1 = abcdr - 1 .$$

6. The error sum of squares *ssE* is *sstot* minus the sums of squares for all other effects in the model. The error degrees of freedom *df* is $n - 1$ minus the degrees of freedom for all of the effects in the model. For a complete model, $df = n - v$, where v is the total number of treatment combinations.

Example:

$$ssE = sstot - ssA - ssB - ssC - ssD$$

$$- ss(AB) - ss(AC) - \cdots - ss(BCD) - ss(ABCD) ,$$

$$df = (n - 1) - (a - 1) - (b - 1) - \cdots - (a - 1)(b - 1)(c - 1)(d - 1) .$$

7. The mean square for an effect is the corresponding sum of squares divided by the degrees of freedom.

$$\text{Example:} \quad ms(ABD) = ss(ABD)/((a - 1)(b - 1)(d - 1)) ,$$

$$msE = ssE/df$$

8. The decision rule for testing the null hypothesis that an effect is zero against the alternative hypothesis that the effect is nonzero is

$$\text{reject } H_0 \text{ if } \frac{ss/\nu}{ssE/df} = \frac{ms}{msE} > F_{\nu, df, \alpha} \,,$$

where ss is the sum of squares calculated in rule 4, ν is the degrees of freedom in rule 2, and $ms = ss/\nu$.

$$\text{Example}: \quad \text{To test } H_0^{ABD} : \{\text{the interaction } ABD \text{ is negligible}\}$$
$$\text{against } H_A^{ABD} : \{\text{the interaction } ABD \text{ is not negligible}\} \,,$$

the decision rule is

$$\text{reject } H_0^{ABD} \text{ if } \frac{ms(ABD)}{msE} > F_{(a-1)(b-1)(d-1), df, \alpha}.$$

9. An estimable contrast for an interaction or main effect is a linear combination of the corresponding parameters (averaged over all higher-order interactions in the model), where the coefficients add to zero over each of the subscripts in turn.

Example: All estimable contrasts for the ABD interaction are of the form

$$\sum_i \sum_j \sum_l h_{ijl} (\alpha\beta\delta)^*_{ijl}$$

where

$$\sum_i h_{ijl} = 0 \ \text{ for all } j, l; \quad \sum_j h_{ijl} = 0 \text{ for all } i, l; \quad \sum_l h_{ijl} = 0 \text{ for all } i, j \,,$$

and where $(\alpha\beta\delta)^*$ is the parameter representing the ABD interaction averaged over all the higher-order interactions in the model.

10. If the sample sizes are equal, the least squares estimate of an estimable contrast in rule 9 is obtained by replacing each parameter with $\overline{y}$ having the same subscripts and averaging over all subscripts not present.

Example: The least squares estimate of the ABD contrast in rule 9 is

$$\sum_i \sum_j \sum_l h_{ijl} \widehat{(\alpha\beta\delta)}^*_{ijl} = \sum_i \sum_j \sum_l h_{ijl} \, \overline{y}_{ij.l.} \,.$$

11. The variance of an estimable contrast for a factorial effect is obtained by adding the squared contrast coefficients, dividing by the product of r and the numbers of levels of all factors not present in the effect, and multiplying by σ^2.

Example: $\text{Var}\left(\sum_i \sum_j \sum_l h_{ijl}\widehat{(\alpha\beta\delta)}^*_{ijl}\right) = \left(\sum_i \sum_j \sum_l h^2_{ijl}/(cr)\right)\sigma^2$.

12. The "sum of squares" for testing the null hypothesis H^c_0 that a contrast is zero is the square of the normalized contrast estimate.

Example: The sum of squares for testing the null hypothesis that the contrast $\sum\sum\sum h_{ijl}(\alpha\beta\delta)^*_{ijl}$ is zero against the alternative hypothesis that the contrast is nonzero is the square of the least squares estimate of the normalized contrast $\sum\sum\sum h_{ijl}(\alpha\beta\delta)^*_{ijl}/\sqrt{\sum\sum\sum h^2_{ijl}/(cr)}$; that is,

$$ssc = \frac{\left(\sum_i \sum_j \sum_l h_{ijl}\overline{y}_{ij.l.}\right)^2}{\sum_i \sum_j \sum_l h^2_{ijl}/(cr)} \ .$$

13. The decision rule for testing the null hypothesis H^c_0 that an estimable contrast is zero, against the alternative hypothesis that the contrast is nonzero, is

$$\text{reject } H^c_0 \text{ if } \frac{ssc}{msE} > F_{1,df,\alpha/m} \ ,$$

where ssc is the square of the normalized contrast estimate, as in rule 12; msE is the error mean square; df is the number of error degrees of freedom; α is the overall Type I error probability; and m is the number of preplanned hypotheses being tested.

14. Simultaneous confidence intervals for contrasts in the treatment combinations can be obtained from the general formula (4.4.20), p. 83, with the appropriate critical coefficients for the Bonferroni, Scheffé, Tukey, and Dunnett methods.

Example: For the four-way complete model, the general formula for simultaneous $100(1-\alpha)\%$ confidence intervals for a set of contrasts of the form $\Sigma\Sigma\Sigma\Sigma c_{ijkl}\tau_{ijkl}$ is

$$\Sigma\Sigma\Sigma\Sigma c_{ijkl}\tau_{ijkl} \in \left(\Sigma\Sigma\Sigma\Sigma c_{ijkl}\overline{y}_{ijkl.} \pm w\ \sqrt{msE\ (\Sigma\Sigma\Sigma\Sigma c^2_{ijkl}/r)}\right),$$

where the critical coefficient, w, is

$$w_B = t_{df,\alpha/2m} \ ; \ \ w_s = \sqrt{(v-1)F_{v-1,df,\alpha}} \ ;$$

$$w_T = q_{v,df,\alpha}/\sqrt{2} \ ; \ \ w_{D2} = |t|^{(0.5)}_{v-1,df,\alpha} \ ;$$

for the four methods, respectively, and v is the number of treatment combinations, and df is the number of error degrees of freedom.

15. Simultaneous confidence intervals for the true mean effects of the treatment combinations in the complete model can be obtained from the general formula (4.3.12), p. 76, with the appropriate critical coefficients for the Bonferroni or Scheffé methods.

    Example: For the four-way complete model, the general formula for simultaneous $100(1 - \alpha)\%$ confidence intervals for true mean effects of the treatment combinations $\mu + \tau_{ijkl}$ is

    $$\mu + \tau_{ijkl} \in \left(\bar{y}_{ijkl.} \pm w \sqrt{msE/r} \right),$$

    where the critical coefficient, w, is

    $$w_B = t_{df,\alpha/(2v)} \quad \text{or} \quad w_s = \sqrt{v \, F_{v,df,\alpha}}$$

    for the Bonferroni and Scheffé methods, respectively.

16. Simultaneous $100(1 - \alpha)\%$ confidence intervals for contrasts in the levels of a single factor can be obtained by modifying the formulae in rule 14. Replace v by the number of levels of the factor of interest, and r by the number of observations on each level of the factor of interest.

    Example: For the four-way complete model, the general formula for simultaneous confidence intervals for contrasts $\sum_i c_i \bar{\tau}_{i...} = \sum_i \alpha_i^*$ in A is

    $$\sum_i c_i \alpha_i^* \in \left(\sum_i c_i \bar{y}_{i....} \pm w \sqrt{msE \left(\sum_i c_i^2/(bcdr) \right)} \right), \qquad (7.3.3)$$

    where the critical coefficients for the five methods are, respectively,

    $$w_B = t_{df,\alpha/(2m)} \; ; \; w_s = \sqrt{(a-1)F_{a-1,df,\alpha}} \; ; \; w_T = q_{a,df,\alpha}/\sqrt{2} \; ;$$

    $$w_{D1} = w_H = t_{a-1,df,\alpha}^{(0.5)} \; ; \; w_{D2} = |t|_{a-1,df,\alpha}^{(0.5)} \; ;$$

    where df is the number of error degrees of freedom.

7.4 A Real Experiment—Popcorn–Microwave Experiment

The experiment described in this section was run by Jianjian Gong, Chongqing Yan, and Lihua Yang in 1992 to compare brands of microwave popcorn. The details in the following checklist have been extracted from the experimenters' report.

The Design Checklist

(a) **Define the objectives of the experiment**.

The objective of the experiment was to find out which brand gives rise to the best popcorn in terms of the proportion of popped kernels. The experiment was restricted to popcorn produced in a microwave oven.

(b) **Identify all sources of variation**.

  (i) Treatment factors and their levels.

   The first treatment factor was "brand." Three levels were selected, including two national brands (levels 1 and 2) and one local brand (level 3). These brands were the brands most commonly used by the experimenters and their colleagues. All three brands are packaged for household consumers in boxes of 3.5 ounce packages, and a random selection of packages was used in this experiment.

   Power of the microwave oven was identified as a possible major source of variation and was included as a second treatment factor. Three available microwave ovens had power ratings of 500, 600, and 625 W. The experimenters used only one oven for each power level. This means that their conclusions could be drawn only about the three ovens in the study and not about power levels in general.

   Popping time was taken as a third treatment factor. The usual instructions provided with microwave popcorn are to microwave it until rapid popping slows to 2 to 3 seconds between pops. Five preliminary trials using brand 3, a 600 W microwave oven, and times equally spaced from 3 to 5 min suggested that the best time was between 4 and 5 min. Hence, time levels of 4, 4.5, and 5 min were selected for the experiment and coded 1–3, respectively.

 (ii) Experimental units

   The experiment was to be run sequentially over time. The treatment combinations were to be examined in a completely random order. Consequently, the experimental units were the time slots that were to be assigned at random to the treatment combinations.

(iii) Blocking factors, noise factors, and covariates.

   Instead of randomly ordering the observations on all of the treatment combinations, it might have been more convenient to have taken the observations oven by oven. In this case, the experiment would have been a "split-plot design" (see Sect. 2.4.4) with ovens representing the blocks. In this experiment, no blocking factors or covariates were identified by the experimenters. The effects of noise factors, such as room temperature, were thought to be negligible and were ignored.

(c) **Choose a rule by which to assign the experimental units to the treatments**.

A completely randomized design was indicated. The time-slots were randomly assigned to the brand–power–time combinations. Popcorn packages were selected at random from a large batch purchased by the experimenters to represent each brand. Changes in quality, if any, of the packaged popcorn over time could not be detected by this experiment.

(d) **Specify measurements to be made, the experimental procedure, and the anticipated difficulties**.

A main difficulty for the experimenters was to choose the response variable. They considered weight, volume, number, and percentage of successfully popped kernels as possible response variables. In each case, they anticipated difficulty in consistently being able to classify kernels as popped or not. To help control such variation or inconsistency in the measurement process, a

single experimenter made all measurements. For measuring weight, the experimenters needed a more accurate scale than was available, since popcorn is very light. They decided against measuring volume, since brands with smaller kernels would appear to give less volume, as the popcorn would pack more easily into a measuring cylinder. The percentage, rather than number, of successfully popped kernels for each package was selected as the response variable.

(e) **Run a pilot experiment**.

The experimenters ran a very small pilot experiment to check their procedure and to obtain a rough estimate of the error variance. they collected observations on only 9 treatment combinations. Using a three-way main-effects model, they found that the overall average popping rate was about 70% and the error standard deviation was a little less than 10.7%. The highest popping rate occurred when the popping time was at its middle level (4.5 min), suggesting that the range of popping times under consideration was reasonable. Results for 600 and 625 W microwave ovens were similar, with lower response rates for the 500 W microwave oven. However, since all possible interactions had been ignored for this preliminary analysis, the experimenters were cautious about drawing any conclusions from the pilot experiment.

(f) **Specify the model**.

For their main experiment, the experimenters selected the three-way complete model, which includes all main effects and interactions between the three treatment factors. They assumed that the packages selected to represent each brand would be very similar, and package variability for each brand could be ignored.

(a) **— revisited. Define the objectives of the experiment**.

Having identified the treatment factors, response variables, etc., the experimenters were able to go back to step (a) and reformalize the objectives of the experiment. They decided that the three questions of most interest were:

- Which combination of brand, power, and time will produce the highest popping rate? (Thus, pairwise comparisons of all treatment combinations were required.)
- Which brand of popcorn performs best overall? (Pairwise comparison of the levels of brand, averaging over the levels of power and time, was required.)
- How do time and power affect response? (Pairwise comparison of time–power combinations, averaging over brands, was required. Also, main-effect comparisons of power and time were required.)

(g) **Outline the analysis**.

Tukey's method of simultaneous confidence intervals for pairwise comparisons was to be used separately at level 99% for each of the above five sets of contrasts, giving an experimentwise confidence level of at least 95%.

(h) **Calculate the number of observations that need to be taken**.

The data from the pilot study suggested that 10.7% would be a reasonable guess for the error standard deviation. This was calculated using a main-effects model rather than the three-way complete model, but we would expect a model with more terms to reduce the estimated error

variance, not to enlarge it. Consequently, the value $msE = 10.7^2$ was used in the sample-size calculations. The experimenters decided that their confidence intervals for any pairwise main-effect comparison should be no wider than 15% (that is, the half-width, or minimum significant difference, should be less than 7.5%). Using rule 16, p. 212, for Tukey's pairwise comparisons, a set of 99% simultaneous confidence intervals for the pairwise differences between the brands (factor A) is

$$\bar{y}_{i...} - \bar{y}_{s...} \pm w_T \sqrt{msE\,(2/(bcr))}\,,$$

where the critical coefficient is $w_T = q_{3,27r-27,0.01}/\sqrt{2}$. The error degrees of freedom are calculated for a complete model as $df = n - v = 27r - 27$. Consequently, using 10.7^2 as the rough estimate of msE, we need to solve

$$msd = \left(q_{3,27(r-1),.01}/\sqrt{2}\right)\sqrt{(10.7)^2(2/(9r))} \le 7.5\,.$$

Trial and error shows that $r = 4$ is adequate. Thus a total of $n = rv = 108$ observations would be needed.

(i) **Review the above decisions. Revise, if necessary**.

The experimenters realized that it would not be possible to collect 108 observations in the time they had available. Since the effects of power levels of 600 and 625 W were comparable in the pilot study, they decided to drop consideration of the 600 W microwave and to include only power levels of 500 W (level 1) and 625 W (level 2) in the main experiment. Also, they decided to take only $r = 2$ observations (instead of the calculated $r = 4$) on each of the $v = 18$ remaining treatment combinations. The effect of this change is to widen the proposed confidence intervals. A set of 99% simultaneous confidence intervals for pairwise comparisons in the brands using Tukey's method and $msE=10.7^2$ would have half-width

$$msd = (q_{3,18,.01}/\sqrt{2})\sqrt{(10.7)^2(2/(6 \times 2))} = 14.5\,,$$

about twice as wide as in the original plan. It was important, therefore, to take extra care in running the experiment to try to reduce the error variability.

The experiment was run, and the resulting data are shown in Table 7.2. Unfortunately, the error variance does not seem to be much smaller than in the pilot experiment, since the mean squared error was reduced only to $(9.36)^2$. A plot of the standardized residuals against fitted values did not show any pattern of unequal variances or outliers. Likewise, a plot of the standardized residuals against the normal scores was nearly linear, giving no reason to question the model assumption of normality. Unfortunately, the experimenters did not keep information concerning the order of observations, so the independence assumption cannot be checked.

Data Analysis

Table 7.3 contains the analysis of variance for investigating the three-way complete model. If an overall significance level of $\alpha \le 0.07$ is selected, allowing each hypothesis to be tested at level $\alpha^* = 0.01$, the only null hypothesis that would be rejected would be H_0^T :{popping time has no effect on the proportion of popped kernels}. However, at a slightly higher significance level, the brand–time interaction also appears to have an effect on the proportion of popped kernels.

If the equivalent cell-means model is used, the null hypothesis of no difference between the treatment combinations would be rejected at significance level $\alpha = 0.07$. This is shown in the row of Table 7.3 labeled "Treatments." Since the design is equireplicate, the main effects and interactions are estimated independently, and their sums of squares add to the treatment sum of squares. The corresponding numbers of degrees of freedom likewise add up.

Figure 7.6 shows an interaction plot for the factors "Brand" and "Time." The plot suggests that use of time level 2, namely 4.5 min, generally gives a higher popping rate for all three brands. Using level 2 of time, brands 1 and 2 appear to be better than brand 3. The two national brands thus appear to be better than the local brand. Brand 1 appears to be less sensitive than brand 2 to the popping time. (We say that brand 1 appears to be *more robust* to changes in popping time.) Unless this perceived difference is due to error variability, which does not show on the plot, brand 1 is the brand to be recommended.

Having examined the analysis of variance table and Fig. 7.6, the most interesting issue seems to be that the differences in the brands might not be the same at the different popping times. This is not one of the comparisons that had been preplanned at step (g) of the checklist. It is usually advisable to include in the plan of the analysis the use of Scheffé's multiple comparisons for all contrasts that look interesting after examining the data. If we had done this at overall 99% confidence level, then the experimentwise error rate would have been at least 94%. Interaction contrasts and their least squares estimates are defined in rules 9 and 10, p. 211. The interaction contrast of most interest is, perhaps,

$$\overline{\tau}_{1.2} - \overline{\tau}_{1.3} - \overline{\tau}_{2.2} + \overline{\tau}_{2.3} \,,$$

Table 7.2 Percentage y_{ijkl} of kernels popped—popcorn–microwave experiment

Brand (i)	Power (j)	Time (k)		
		1	2	3
1	1	73.8, 65.5	70.3, 91.0	72.7, 81.9
1	2	70.8, 75.3	78.7, 88.7	74.1, 72.1
2	1	73.7, 65.8	93.4, 76.3	45.3, 47.6
2	2	79.3, 86.5	92.2, 84.7	66.3, 45.7
3	1	62.5, 65.0	50.1, 81.5	51.4, 67.7
3	2	82.1, 74.5	71.5, 80.0	64.0, 77.0
		$\overline{y}_{..1.} = 72.9000$	$\overline{y}_{..2.} = 79.8667$	$\overline{y}_{..3.} = 63.8167$

Table 7.3 Three-way ANOVA for the popcorn–microwave experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
B	2	331.1006	165.5503	1.89	0.1801
P	1	455.1111	455.1111	5.19	0.0351
T	2	1554.5756	777.2878	8.87	0.0021
B*P	2	196.0406	98.0203	1.12	0.3485
B*T	4	1433.8578	358.4644	4.09	0.0157
P*T	2	47.7089	23.8544	0.27	0.7648
B*P*T	4	47.3344	11.8336	0.13	0.9673
Treatments	17	4065.7289	239.1605	2.73	0.0206
Error	18	1577.8700	87.6594		
Total	35	5643.5989			

Fig. 7.6 Interaction plot for factors "brand" and "time" averaged over "power" for the popcorn–microwave experiment
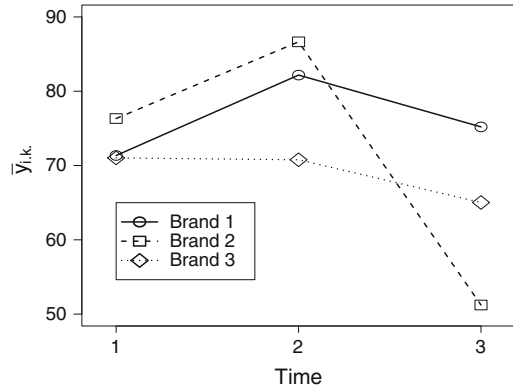
which compares the differences in brands 1 and 2 at popping times 2 and 3. This has least squares estimate

$$\overline{y}_{1.2.} - \overline{y}_{1.3.} - \overline{y}_{2.2.} + \overline{y}_{2.3.} = 82.175 - 75.20 - 86.65 + 51.225 = -28.45\,.$$

The importance of preplanning will now become apparent. Using Scheffé's method (rule 14) at overall level 99%, a confidence interval for this contrast is given by

$$\sum_i \sum_k c_{ik} \overline{y}_{i.k.} \pm w_S \sqrt{msE\left(\sum_i \sum_k c_{ik}^2/(br)\right)}$$
$$= -28.45 \pm \sqrt{17 F_{17,18,.01}} \sqrt{87.6594\,(4/4)}$$
$$= -28.45 \pm 69.69 \quad = \ (-96.41,\ 39.52)\,.$$

Our popping rates are percentages, so our minimum significant difference is 69%. This is far too large to give any useful information. The resulting interval gives the value of the interaction contrast as being between -96.4% and 39.5%! Had this contrast been preplanned for at individual confidence level 99%, we would have used the critical value $w_B = t_{18,0.005} = 2.878$ instead of $w_S = 7.444$, and we would have obtained a minimum significant difference of about 30%, leading to the interval $(-55.40, -1.50)$. Although still wide, this interval would have given more information, and in particular, it would have indicated that the interaction contrast was significantly different from zero.

   The other important effect that showed up in the analysis of variance table was the effect of the different popping times (4, 4.5, or 5 min). Comparisons of popping times did feature as one of the preplanned sets of multiple comparisons, and consequently, we use Tukey's method (rule 16) for pairwise differences $\gamma_k - \gamma_u$ at overall level 99%. The minimum significant difference is

$$msd = w_T \sqrt{msE\ \Sigma c_k^2/(abr)} = (q_{3,18,.01}/\sqrt{2})\ \sqrt{(87.6594)(2/12)} = 12.703\,.$$

The average percentages of popped kernels for the three popping times are shown in Table 7.2 as

$$\overline{y}_{..1.} = 72.9000\,,\quad \overline{y}_{..2.} = 79.8667\,,\quad \overline{y}_{..3.} = 63.8167\,,$$

so the three confidence intervals are

$$\gamma_1 - \gamma_2 \in (-6.9667 \pm 12.7030) = (-5.736, 19.670),$$
$$\gamma_1 - \gamma_3 \in \quad (9.0833 \pm 12.7030) = (-3.620, 21.786),$$
$$\gamma_2 - \gamma_3 \in \ (16.0500 \pm 12.7030) = \quad (3.347, 28.753).$$

We see that at an experimentwise confidence level of at least 94%, use of popping time 2 (4.5 min) produces on average between 3.35% and 28.75% more popcorn than use of popping time 3 (5 min).

The other questions asked by the experimenters appear to be of less interest, and we will omit these. The experimentwise confidence level is still at least 94%, even though we have chosen not to calculate all of the preplanned intervals.

7.5   One Observation per Cell

If the complete model is used for a factorial experiment with one observation per cell, then there are no degrees of freedom available to estimate the error variance. This problem was discussed in Sect. 6.7, where one possible method of analysis was described. The method relies on being able to identify a number of negligible contrasts, which are then excluded from the model. The corresponding sums of squares and degrees of freedom are used to estimate the error variance. With this approach, confidence intervals can be constructed and hypothesis tests conducted. An example with four treatment factors that are believed not to interact with each other is presented in the next section.

Two alternative approaches for the identification of nonnegligible contrasts are provided in the subsequent sections. In Sect. 7.5.2 we show an approach based on the evaluation of a half-normal probability plot of a set of contrast estimates, and in Sect. 7.5.3 we discuss a more formalized approach. These two approaches work well under effect sparsity, that is, when most of the treatment contrasts under examination are negligible.

7.5.1   Analysis Assuming that Certain Interaction Effects are Negligible

For a single replicate factorial experiment, if the experimenter knows ahead of time that certain interactions are negligible, then by excluding those interactions from the model, the corresponding degrees of freedom can be used to estimate the error variance. It must be recognized, however, that if interactions are incorrectly assumed to be negligible, then *msE* will be inflated, in which case the results of the experiment may be misleading.

Table 7.4   Data for the drill advance experiment

$ABCD$	Advance	$y = \log(\text{advance})$	$ABCD$	Advance	$y = \log(\text{advance})$
1111	1.68	.2253	2111	1.98	.2967
1112	2.07	.3160	2112	2.44	.3874
1121	4.98	.6972	2121	5.70	.7559
1122	7.77	.8904	2122	9.43	.9745
1211	3.28	.5159	2211	3.44	.5366
1212	4.09	.6117	2212	4.53	.6561
1221	9.97	.9987	2221	9.07	.9576
1222	11.75	1.0700	2222	16.30	1.2122

Source Applications of Statistics to Industrial Experimentation, by C. Daniel, Copyright © 1976, John Wiley & Sons, New York. Reprinted by permission of John Wiley & Sons, Inc

Table 7.5 Analysis of variance for the drill advance experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
A	1	0.01275	0.01275	7.02	0.0226
B	1	0.25387	0.25387	139.74	0.0001
C	1	1.00550	1.00550	553.46	0.0001
D	1	0.08045	0.08045	44.28	0.0001
Error	11	0.01998	0.00182		
Total	15	1.37254			

Example 7.5.1 `Drill advance experiment`

Daniel (1976) described a single replicate $2 \times 2 \times 2 \times 2$ experiment to study the effects of four treatment factors on the rate of advance of a small stone drill. The treatment factors were "load on the drill" (A), "flow rate through the drill" (B), "speed of rotation" (C), and "type of mud used in drilling" (D). Each factor was observed at two levels, coded 1 and 2. The author examined several different transformations of the response and concluded that the log transform was one of the more satisfactory ones. In the rest of our discussion, y_{ijkl} represents the log (to the base 10) of the units of drill advance, as was illustrated in the original paper. The data are shown in Table 7.4.

In many experiments with a number of treatment factors, experimenters are willing to believe that some or all of the interactions are very small. Had that been the case here, the experimenter would have used the four-way main-effects model. (Analysis of this experiment without assuming negligible interactions is discussed in Example 7.5.2, p. 222.)

Degrees of freedom and sums of squares are given by rules 2 and 4 in Sect. 7.3. For example, the main effect of B has $b-1$ degrees of freedom and

$$ ssB \; = \; acd \sum_i \left(\overline{y}_{.j..} - \overline{y}.... \right)^2 \; = \; acd \sum_i \overline{y}_{.j..}^2 - acbd \overline{y}....^2 \, . $$

The sums of squares for the other effects are calculated similarly and are listed in the analysis of variance table, Table 7.5. The error sum of squares shown in Table 7.5 is the total of all the eleven (negligible) interaction sums of squares and can be obtained by subtraction, as in rule 6, p. 210:

$$ ssE = sstot - ssA - ssB - ssC - ssD = 0.01998 \, . $$

Similarly, the number of error degrees of freedom is the total of the $15 - 4 = 11$ interaction degrees of freedom. An estimate of σ^2 is therefore $msE = ssE/11 = 0.0018$. Since $F_{1,11,.01} = 9.65$, the null hypotheses of no main effects of B, C, and D would all have been rejected at overall significance level $\alpha \leq 0.04$. Alternatively, from a computer analysis we would see that the p-values for B, C, and D are each less than or equal to an individual significance level of $\alpha^* = 0.01$.

Confidence intervals for the $m = 4$ main-effect contrasts using Bonferroni's method at an overall level of at least 95% can be calculated from rule 16. From rules 10 and 11 on p. 211, the least squares estimate for the contrast that compares the effects of the high and low levels of B is $\overline{y}_{.2..} - \overline{y}_{.1..}$, with variance $\sigma^2(2/8)$, giving the confidence interval

$$ \left(\overline{y}_{.2..} - \overline{y}_{.1..} \pm w_B \, \sqrt{msE \, (2/8)} \right) , $$

where the critical coefficient is

$$w_B = t_{11,.025/4} = t_{11,.00625} \approx z + (z^3 + z)/((4)(11)) \approx 2.911 \,,$$

from (4.4.22), p. 83, and

$$msd = w_B \sqrt{msE\,(2/8)} = 2.911 \sqrt{(0.00182)/4} = 0.062 \,.$$

Now, $\overline{y}_{.2..} = 0.820$, $\overline{y}_{.1..} = 0.568$, so the confidence interval for the B contrast is

$$(0.252 \pm 0.062) \approx (0.190, 0.314) \,,$$

where the units are units of log drill advance. Confidence intervals for the other three main effects comparing high with low levels can be calculated similarly as

$$
\begin{aligned}
A &: 0.056 \pm 0.062 = (-0.006, 0.118) \,, \\
C &: 0.501 \pm 0.062 = \ \ (0.439, 0.563) \,, \\
D &: 0.142 \pm 0.062 = \ \ (0.080, 0.204) \,.
\end{aligned}
$$

We see that the high levels of B, C, D give a somewhat higher response in terms of log drill advance (with overall confidence level at least 95%), whereas the interval for A includes zero. $\square$

7.5.2   Analysis Using Half-Normal Probability Plot of Effect Estimates

For a single replicate factorial experiment, with v treatment combinations, one can find a set of $v - 1$ orthogonal contrasts. When these are normalized, the contrast estimators all have variance σ^2. If the assumptions of normality, equal variances, and independence of the response variables are approximately satisfied, the estimates of negligible contrasts are like independent observations from a normal distribution with mean zero and variance σ^2. If we were to plot the normalized contrast estimates against their normal scores (in the same way that we checked for normality of the error variables in Chap. 5), the estimates of negligible effects would tend to fall nearly on a straight line. Any contrast for which the corresponding estimate would appear to be far from the line would be considered to be nonnegligible. The sign of such a nonnegligible contrast could be positive or negative and this depends upon which level of the factor is labeled as the high level and which is labeled as the low level. For many factors (especially qualitative factors), these designations are arbitrary, and so it is common to use a *half-normal probability plot* for detecting nonnegligible contrasts. This is obtained by plotting the *absolute values* of the normalized contrast estimates against their *half-normal scores*.

Half-normal scores are percentiles of the half-normal distribution with $\sigma = 1$, corresponding to the distribution of the absolute value of a standard normal random variable. In particular, the qth half-normal score for $m = v - 1$ contrast absolute estimates is the value ξ_q for which

$$P(Z \le \xi_q) = 0.5 * [1 + q/(m + 1)],$$

where Z is a standard normal random variable. Hence, the qth half-normal score is

$$\xi_q = \Phi^{-1}[0.5 * (1 + q/(m + 1))] \,, \tag{7.5.4}$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution.

If the model assumptions are approximately satisfied, and if the estimated effects are all negligible, then the half-normal plot should show points roughly on a straight line through the origin with slope equal to σ. However, if any of the effects are large, then their estimates should stand out as relatively large, creating a nonlinear pattern. Provided that there is effect sparsity—namely, all but a few contrast estimates are expected to be negligible—it is not difficult to pick out the nonnegligible contrasts. We note that it is possible to use a half-normal probability plot for non-normalized contrasts provided that they all have the *same variance*, so that the line of negligible contrasts still has slope equal to the common contrast standard deviation.

Example 7.5.2 `Drill advance experiment, continued`

The data for the drill advance experiment were given in Table 7.4 in Example 7.5.1. The experiment involved treatment factors "load" (A), "flow" (B), "speed" (C), and "mud" (D) and response "log(drill advance)" (Y). If we have no information about which factors are likely to be negligible, we would use the four-way complete model or the equivalent cell-means model:

$$Y_{ijkl} = \mu + \tau_{ijkl} + \epsilon_{ijkl} \,,$$
$$\epsilon_{ijkl} \sim N(0, \sigma^2) \,,$$
$$\epsilon'_{ijkl}\text{s mutually independent} \,,$$
$$i = 1, 2; \quad j = 1, 2; \quad k = 1, 2; \quad l = 1, 2 \,.$$

The contrast coefficients for the four main effects and some of the interactions in such a cell-means model were listed in Table 7.1, p. 208. The contrast coefficients for the interactions can be obtained by multiplying together the corresponding main-effect coefficients. Each contrast is normalized by dividing the coefficients by $\sqrt{\Sigma c_{ijkl}^2 / r} = \sqrt{16}$ rather than by the divisors of Table 7.1. For example, the normalized BCD interaction contrast has coefficient list

$$\frac{1}{4}[-1, \; 1, \; 1, -1, \; 1, -1, -1, \; 1, -1, \; 1, \; 1, -1, \; 1, -1, -1, \; 1 \,] \,.$$

The least squares estimate of the normalized BCD interaction contrast is then

$$\frac{1}{4}[-(0.2253) + (0.3160) + \cdots - (0.9576) + (1.2122)] \quad = \quad -0.0300 \,.$$

The 15 normalized factorial contrast estimates are given in Table 7.6, and the half-normal probability plot of these estimates is shown in Fig. 7.7, with the main effect estimates labeled. Observe that all the estimates fall roughly on a straight line, except for the estimates for the main-effects of factors D, B, and C. Hence, these three main effects appear to be nonnegligible.

In the construction of the half-normal probability plot, the contrasts must be scaled to have the same variance, and normalization is one way to achieve this. When *all factors* have two levels, and when the contrasts are written in terms of the treatment combination parameters as in Table 7.1, their least squares estimators will all have the same variance, as long as the *same divisor* is used for every contrast. A popular selection for divisor is $v/2$, which is the natural divisor for main-effect contrasts comparing the average treatment combination effects at the two levels of a factor. Thus, rather than using divisor $\sqrt{16}$ in Example 7.5.2, we could have used divisor $v/2 = 8$. If the divisor $v/2$ is used, the estimators all have variance $4\sigma^2/v$. If no divisor is used, the estimators all have variance $v\sigma^2$. As long as the variances are all equal, the half-normal probability plot can be used to identify the important

Table 7.6 Normalized contrast estimates for the drill advance experiment

Effect:	A	B	C	D		
Estimate:	0.1129	0.5039	1.0027	0.2836		
Effect:	AB	AC	AD	BC	BD	CD
Estimate:	−0.0298	0.0090	0.0581	−0.0436	−0.0130	0.0852
Effect:	ABC	ABD	ACD	BCD	ABCD	
Estimate:	0.0090	0.0454	0.0462	−0.0300	0.0335	

Fig. 7.7 Half-normal probability plot of normalized contrast absolute estimates for the drill advance experiment
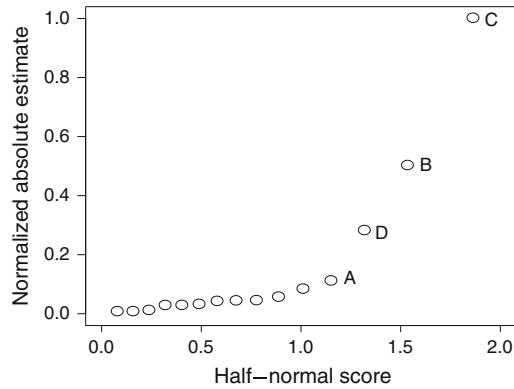
contrasts. In *all other* sizes of experiment, the contrast coefficients are not all ±1, and we recommend that all contrasts be normalized so that their estimators all have variance σ^2.

7.5.3   Analysis Using Confidence Intervals

In this section, an alternative to the half-normal probability plot is presented for the analysis of a single replicate factorial experiment. As with the half-normal probability plot, we require a set of m orthogonal contrasts and effect sparsity, and we make no assumptions as to *which* effects are negligible. The procedure provides confidence intervals for the m contrasts with a simultaneous confidence level of at least $100(1 - \alpha)\%$. For the moment, we recode the treatment combinations as $1, 2, \ldots, v$, their effects as $\tau_1, \tau_2, \ldots \tau_v$, and we generically denote each of the m contrasts by $\sum c_i \tau_i$.

First, let d equal the integer part of $(m + 1)/2$, which is $m/2$ if m is even and is $(m + 1)/2$ if m is odd. The method requires that there be at least d negligible effects (effect sparsity). In general, this will be true if at least one of the factors has no effect on the response (and so does not interact with any of the other factors) or if most of the higher-order interactions are negligible.

We take each of the m contrasts in turn. For the kth contrast $\sum c_i \tau_i$, we calculate its least squares estimate $\sum c_i y_i$ and its sum of squares ssc_k, using rules 10 and 12, p. 211. We then calculate the *quasi mean squared error* msQ_k for the kth contrast by taking the average of the d smallest of $ssc_1, \ldots, ssc_{k-1}, ssc_{k+1}, \ldots, ssc_m$ (that is, the smallest d contrast sums of squares ignoring the kth).

The *Voss−Wang method* gives simultaneous $100(1 - \alpha)\%$ confidence intervals for the m contrasts, the confidence interval for the kth contrast being

$$\sum c_i \tau_i \quad \in \quad \left(\sum c_i y_i \pm w_V \sqrt{msQ_k \sum_i c_i^2} \right). \tag{7.5.5}$$

The critical coefficients $w_V = v_{m,d,\alpha}$ are provided in Appendix A.11. The critical values $v_{m,d,\alpha}$ were obtained by Voss and Wang (1999) as the square root of the percentile corresponding to α in the right-hand tail of the distribution of

$$V^2 = \max \left\{ SSC_k / MSQ_k \right\},$$

where the maximum is over $k = 1, 2, \ldots, m$.

Example 7.5.3 `Drill advance experiment, continued`

Consider again the single replicate drill advance experiment of Examples 7.5.1 and 7.5.2 with four factors having two levels each. We can find $m = 15$ orthogonal factorial contrasts, nine of which are shown in Table 7.1, p. 208. The Voss–Wang method of simultaneous confidence intervals, described above, is reasonably effective as long as there are at least $d = 8$ negligible contrasts in this set.

For an overall 95% confidence level, the critical coefficient is obtained from Appendix A.11 as $w_V = v_{15,8,0.05} = 9.04$. Selecting divisors $v/2 = 8$ for each contrast, we obtain the least squares estimates in Table 7.7.

The sums of squares for the 15 contrasts are also listed in Table 7.7 in descending order. For the contrasts corresponding to each of the seven largest sums of squares, the quasi mean squared error is composed of the eight smallest contrast sums of squares; that is,

$$msQ_k = (0.0000808 + \cdots + 0.0020571)/8 = 0.0009004,$$

and the minimum significant difference for each of these seven contrasts is

Table 7.7 Confidence interval information for the drill advance experiment

Effect	ssc_k	msQ_k	Estimate	msd_k
C	1.0054957	0.0009004	0.5014	0.1356
B	0.2538674	0.0009004	0.2519	0.1356
D	0.0804469	0.0009004	0.1418	0.1356
A	0.0127483	0.0009004	0.0565	0.1356
CD	0.0072666	0.0009004	0.0426	0.1356
AD	0.0033767	0.0009004	0.0291	0.1356
ACD	0.0021374	0.0009004	0.0231	0.1356
ABD	0.0020571	0.0009105	0.0227	0.1364
BC	0.0019016	0.0009299	−0.0218	0.1378
ABCD	0.0011250	0.0010270	0.0168	0.1449
BCD	0.0008986	0.0010553	−0.0150	0.1468
AB	0.0008909	0.0010563	−0.0149	0.1469
BD	0.0001684	0.0011466	−0.0065	0.1531
ABC	0.0000812	0.0011575	0.0045	0.1538
AC	0.0000808	0.0011575	0.0045	0.1537

$$msd_k = v_{15,8,0.05}\sqrt{msQ_k\,(16/(8\times 8))} = (9.04)\sqrt{0.0009004\times 0.25} \approx 0.1356\,.$$

The quasi mean squared errors for the contrasts corresponding to the eight smallest sums of squares are modestly larger, leading to slightly larger minimum significant differences and correspondingly wider intervals. All contrast estimates and minimum significant differences are summarized in Table 7.7.

The four largest contrast estimates in absolute value are 0.5014 for C, 0.2519 for B, 0.1418 for D, and 0.0565 for A, giving the intervals

$$
\begin{aligned}
\text{For } C: \quad & 0.5014 \pm 0.1356 & = & \ \ (0.3658, 0.6370)\,, \\
\text{For } B: \quad & 0.2519 \pm 0.1356 & = & \ \ (0.1163, 0.3875)\,, \\
\text{For } D: \quad & 0.1418 \pm 0.1356 & = & \ \ (0.0062, 0.2774)\,, \\
\text{For } A: \quad & 0.0565 \pm 0.1356 & = & \ (-0.0791, 0.1921)\,.
\end{aligned}
$$

Thus, in the 95% simultaneous set, the intervals for the main-effect contrasts of C, B, and D exclude zero and are declared to be the important effects. The intervals for A and for all of the interaction contrasts include zero, so we conclude that these contrasts are not significantly different from zero. Notice that our conclusion agrees with that drawn from the half-normal probability plot. The benefit of the Voss–Wang method is that we no longer need to guess which contrast estimates lie on the straight line, and also that we have explicit confidence intervals for the magnitudes of the nonnegligible contrasts.   □

7.6   Using SAS Software

The analysis of experiments with three or more factors and at least one observation per cell uses the same types of SAS commands as illustrated for two factors in Sect. 6.8. In Sect. 7.6.1, we illustrate the additional commands needed to obtain a half-normal probability plot of the contrast estimates in the drill advance experiment, and in Sect. 7.6.2, we illustrate computations for the Voss-Wang confidence intervals. In Sect. 7.6.3, we show the complications that can arise when one or more cells are empty.

7.6.1   Half-Normal Probability Plots of Contrast Estimates

In Table 7.8, we show a SAS program for producing a half-normal probability plot similar to that of Fig. 7.7, p. 223, but for the unnormalized contrast estimates of the drill advance experiment. The levels of A, B, C, and D together with the responses ADVANCE are entered via the INPUT statement in the first DATA statement. A log transformation is then taken so that the response Y used in the analysis is the log of the units of drill advance. Note that the function LOG10() calculates log to the base 10, whereas LOG() would calculate log to the base e, which is the more usual transformation. The coded factor levels 1 and 2 are converted to contrast coefficients -0.5 and $+0.5$, respectively (e.g., A = A - 1.5). These coefficients could have been entered directly via the INPUT statement, as shown in Table 6.14, p. 184. The interaction coefficients are obtained by multiplication to also have values ± 0.5 (e.g., AB = 2*A*B). The contrast coefficients are printed as columns similar to those in Table 7.1, p. 208, but with values ± 0.5.

The regression procedure, PROC REG, is used to compute regression coefficient estimates, which are the desired contrast estimates. Though unnormalized, these estimates can be shown to have common variance $\sigma^2/4$. The option OUTEST outputs these contrast estimates to a new data set DRILL2, with one row and many variables. Since the OUTEST option saves more variables than needed, a copy of

Table 7.8 SAS program for a half-normal probability plot for the drill advance 2^4 experiment

```
DATA DRILL;
  INPUT A B C D ADVANCE;
  Y = LOG10(ADVANCE); * log to base 10;
  * Compute contrast coefficients +/-0.5;
  A = A - 1.5; B = B - 1.5; C = C - 1.5; D = D - 1.5;
  AB = 2*A*B; AC = 2*A*C; AD = 2*A*D; BC = 2*B*C; BD = 2*B*D;
  CD = 2*C*D;
  ABC = 4*A*B*C; ABD = 4*A*B*D; ACD = 4*A*C*D; BCD = 4*B*C*D;
  ABCD = 8*A*B*C*D;
  LINES;
  1 1 1 1    1.68
  : : : :      :
  2 2 2 2   16.30
PROC PRINT;

* Compute coefficient estimates corresponding to each contrast,
* and output the estimates to a new data set named "drill2";
PROC REG OUTEST = DRILL2 NOPRINT;
  MODEL Y = A B C D AB AC AD BC BD CD ABC ABD ACD BCD ABCD;

* Keep only the variables containing effect estimates;
DATA DRIILL2; SET DRILL2; KEEP A--ABCD;

* Transpose the data set "drill2" to get the effect estimates in
* a new data set "drill3" under the variable name "est1";
PROC TRANSPOSE PREFIX = EST OUT = DRILL3;

* Compute absolute estimates and corresponding half-normal scores;
DATA DRILL3; SET DRILL3;
  ABS_EST = ABS(EST1); * ABS_EST = 4*ABS(EST1) would normalize;
PROC SORT; BY ABS_EST;
DATA DRILL3; SET DRILL3; P = _N_/16; HP = 0.5*(1+P);
  HNSCORE = PROBIT(HP);

* Generate high-resolution half-normal probability plot;
PROC SGPLOT;
  SCATTER Y = ABS_EST X = HNSCORE;
  TITLE "Half-Normal Probability Plot of Contrast Absolute Estimates";
  YAXIS LABEL = "Absolute Estimate"; XAXIS LABEL = "Half-Normal Score";
```

DRILL2 is made which only keeps the variables with the contrast estimates. The NOPRINT option suppresses printing of the procedure's output. (See Chap. 8 for more information about regression.)

In order to be able to plot the estimates, we need them as the different values of a single variable. This is achieved by PROC TRANSPOSE, which turns the single row of the data set DRILL2 into a column in the new data set DRILL3. The resulting least squares estimates are listed as values of the variable EST1. The absolute estimates are then computed and sorted. The absolute estimates could have been normalized by multiplying by two, though we have not done so here.

In the final DATA step, the half-normal scores corresponding to the values of EST1 are calculated as in (7.5.4) the PROBIT function being the inverse cumulative distribution function (cdf) of the standard normal distribution. This final data set is then printed.

Finally, the last procedure in Table 7.8, the statistical graphics plotting procedure PROC SGPLOT, draws a high resolution plot of the absolute contrast estimates versus the half-normal scores. Assuming

Table 7.9 SAS program for the Voss–Wang method for the drill advance experiment

```
* Use the data set DRILL from the prior program;
DATA DRILL; SET DRILL;

* Fit complete model to obtain the m=15 effect sums of squares;
PROC GLM;
  MODEL Y = A B C D AB AC AD BC BD CD ABC ABD ACD BCD ABCD / SS1;
  ODS select ModelANOVA ParameterEstimates;

* The models fit below depend on the results of the above GLM procedure;
* CIs for C, B, D, A, CD, AD, and ACD: compute estimates and standard
* errors by omitting the (other) d=8 effects with the 8 smallest SSs;
PROC GLM;
  MODEL Y = C B D A CD AD ACD / SS1;
  ODS select OverallANOVA ParameterEstimates;

* CI for ABD: compute est and stde by omitting d=8 other effects;
PROC GLM;
  MODEL Y = C B D A CD AD ABD / SS1;
  ODS select OverallANOVA ParameterEstimates;

* CI for BC: compute est and stde by omitting d=8 other effects;
PROC GLM;
  MODEL Y = C B D A CD AD BC / SS1;
  ODS select OverallANOVA ParameterEstimates;

* Continue as above for the six remaining effects;
```

effect sparsity, the nonnegligible contrasts are those whose estimates do not lie along a straight line through the origin.

7.6.2   Voss–Wang Confidence Interval Method

The SAS program in Table 7.9 does most of the computations needed to obtain the Voss-Wang simultaneous confidence intervals (7.5.5) used for analysis of a single-replicate experiment (Sect. 7.5.3). Using the data of the drill advance experiment, the program starts with the DRILL data set created in the program in Table 7.8.

The first call of PROC GLM fits the complete model, providing the estimates and sum of squares for each of the 15 effects, matching the values given in Table 7.7 (p. 224). While this is sufficient information to facilitate applying the Voss-Wang method by hand, the rest of the program provides additional useful computations based on the results of the first procedure call.

First though, a few comments regarding the calls of PROC GLM. With no CLASS statement, PROC GLM fits a regression model, as did PROC REG in Sect. 7.6.1. Analogously, by again using contrast coefficients ± 0.5, the resulting regression coefficient estimates are again the desired contrast estimates displayed in Table 7.7. The option SS1, while unnecessary, requests output of the Type I sums of squares, thereby suppressing output of the matching Type III sums of squares. The output delivery system statement ODS, also unnecessary, selectively limits output by ODS table name.

Now, given the results of the first call of PROC GLM, consider using SAS software for additional computations for formula (7.5.5). For the kth effect, the corresponding quasi mean squared error msQ_k

is most easily calculated as the error mean square obtained from the submodel that omits the terms corresponding to the d smallest contrast sums of squares besides ssc_k. The second and subsequent calls of PROC GLM illustrate these computations for some of the effects.

The second call of PROC GLM fits the model excluding the $d = 8$ effects with the smallest sums of squares from the first call. This yields the value $mse = msQ_k = 0.00090004$ given in Table 7.7, needed to compute the confidence intervals (7.5.5) for the other seven effects C, B, D, A, CD, AD and ACD. Moreover, for each of these seven effects, the SAS software provides a standard error value 0.01500370. While this is not a standard error per se, since msQ_k is not an estimate of σ^2, it is the value of $\sqrt{msQ_k \sum_i c_i^2}$ in formula (7.5.5). The critical values $v_{m,d,\alpha}$ for the Voss–Wang method are not directly available through SAS, so the intervals must be completed by hand.

The third call of GLM provides the information to compute the confidence interval (7.5.5) for the effect ABD. Since ABD had one of the eight smallest sums of squares, ABD is included in the model in place of the term ACD that had the next smallest sum of squares. This call yields $mse = msQ_k = 0.00091049$, matching the value given in Table 7.7, and standard error value 0.01508716 for ABD.

Similarly, the fourth call of GLM provides the corresponding information to compute the confidence interval (7.5.5) for the effect BC, and additional calls could be made for the six remaining effects.

7.6.3 Experiments with Empty Cells

We now illustrate the use of SAS software for the analysis of an experiment with empty cells. No new SAS procedures or commands are introduced, but the empty cells can cause complications. For illustration, we use the following experiment.

Example 7.6.1 Rail weld experiment

S. M. Wu (1964) illustrated the usefulness of two-level factorial designs using the data listed in the SAS program of Table 7.10. Under investigation were the effects of three factors—ambient temperature (T), wind velocity (V), and rail steel bar size (S)—on the ultimate tensile strength of welds. The factor levels were 0° and 70 °F for temperature, 0 and 20 miles per hour for wind velocity, and 4/11 and 11/11 in. for bar size, each coded as levels 1 and 2, respectively. Only six of the possible eight treatment combinations were observed, but $r = 2$ observations were taken on each of these six.

Some SAS commands for analyzing the rail weld experiment are presented in Table 7.10. Notice that rather than listing the two observations for each treatment combination on separate lines, we have listed them as Y1 and Y2 on the same line. We have then combined the observations into the response variable Y. The new variable REP, which will be ignored in the model, is merely a device to keep the observations distinct. This method of input is often useful if the data have been stored in a table, with the observations for the same treatment combinations listed side by side, as in Table 7.2, p. 216.

The three-way complete model is requested in the first call of PROC GLM in Table 7.10. The output is shown in Fig. 7.8. With two cells empty, there are data on only six treatment combinations, so there are only five degrees of freedom available for comparing treatments. This is not enough to measure the three main effects, the three two-factor interactions, and the three-factor interaction. This is indicated in the output, since two effects have zero degrees of freedom. The ESTIMATE statement for the contrast under the first call of PROC GLM generates no output. Instead, it generates a note in the SAS log indicating that the contrast is not estimable.

The only model that can be used is one that uses at most five degrees of freedom. Of course, this should be anticipated ahead of time during step (g) of the checklist (Chap. 2). Figure 7.9 illustrates with a solid ball at the corresponding corners of the cube the treatment combinations for which data are

collected. One might guess that the TV interaction effect is not estimable, since data are only collected at three of the four combinations of levels of these two factors.

One possibility is to exclude from the complete model those interactions for which the Type I degrees of freedom are zero, namely the TV and TVS interaction effects. The contrast coefficient lists for the seven factorial effects are shown in Table 7.11. It is clear that the T and V contrasts are not orthogonal to the TV interaction contrast, and that the S, TS, and VS contrasts are not orthogonal to the TVS interaction contrast. Consequently, an incorrect omission of TV and TVS from the model will bias the estimates of all the other contrasts. If we do decide to exclude both the TV and TVS interaction effects, then the model is of the form

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{ji} + \epsilon_{ijkl} \, .$$

We illustrate analysis of this model using the second call of PROC GLM in Table 7.10. Some of the output is shown in Fig. 7.10. The contrasts for T and V are not orthogonal to each other, but they can be estimated (although with a small positive correlation). Similar comments apply to the S, TS, and VS contrasts. None of the factorial effects appears particularly strong in Fig. 7.10.

The ESTIMATE statements under the second call of PROC GLM generate the information shown in Fig. 7.10 for testing or constructing confidence intervals for the usual main effects and two-factor interaction effects under the given model.   □

Table 7.10 SAS program for the rail weld experiment with two empty cells

```
DATA;
  INPUT T V S Y1 Y2;
  REP = 1; Y = Y1; OUTPUT; * create SAS observation for y = y1;
  REP = 2; Y = Y2; OUTPUT; * create SAS observation for y = y2;
  LINES;
  1 1 1 84.0 91.0
  1 1 2 77.7 80.5
  2 1 1 95.5 84.0
  2 1 2 99.7 95.4
  2 2 1 76.0 98.0
  2 2 2 93.7 81.7
PROC PRINT;
  VAR T V S REP Y;
* try to fit a 3-way complete model;
PROC GLM;
  CLASS T V S;
  MODEL Y = T | V | S;
  ESTIMATE 'TEMPERATURE' T  -1 1;
* fit a sub-model using 5 degrees of freedom;
PROC GLM;
  CLASS T V S;
  MODEL Y = T V S T*S V*S;
  ESTIMATE 'TEMPERATURE' T -1 1;
  ESTIMATE 'VELOCITY'     V -1 1;
  ESTIMATE 'SIZE'         S -1 1;
  ESTIMATE 'TEMPERATURE*SIZE' T*S 1 -1 -1 1 / DIVISOR = 2;
  ESTIMATE 'VELOCITY*SIZE'    V*S 1 -1 -1 1 / DIVISOR = 2;
```

Source Data are from Wu (1964). Copyright © 1964 American Welding Society. Reprinted with permission. (Reprinted University of Wisconsin Engineering Experiment Station, Reprint 684)

Fig. 7.8 Output from the first call of PROC GLM for the rail weld experiment

Fig. 7.9 Treatment combinations included in the design of the rail weld experiment
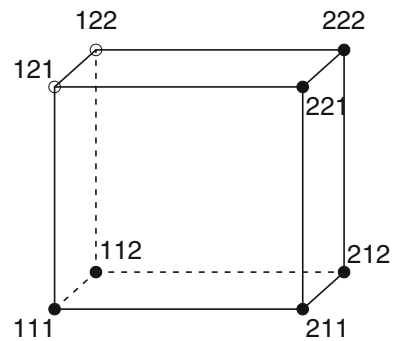
Table 7.11 Contrast coefficients for the observed treatment combinations (T.C.) in the rail weld experiment

TC	T	V	TV	S	TS	VS	TVS
111	−1	−1	1	−1	1	1	−1
112	−1	−1	1	1	−1	−1	1
211	1	−1	−1	−1	−1	1	1
212	1	−1	−1	1	1	−1	−1
221	1	1	1	−1	−1	−1	−1
222	1	1	1	1	1	1	1

Fig. 7.10 Output from the
second call of `PROC GLM`

7.7   Using R Software

The analysis of experiments with three or more factors and at least one observation per cell uses the same types of R commands as illustrated for two factors in Sect. 6.9. In Sect. 7.7.1, we illustrate the additional commands needed to obtain a half-normal probability plot of the normalized contrast estimates in the drill advance experiment, and in Sect. 7.7.2, we illustrate computations for the Voss-Wang confidence intervals. In Sect. 7.7.3, we show the complications that can arise when one or more cells are empty.

7.7.1   Half-Normal Probability Plots of Contrast Estimates

In Table 7.12, we show an R program for producing a half-normal probability plot similar to that of Fig. 7.7, p. 223, for the normalized contrast estimates of the drill advance experiment. The levels of A, B, C, and D together with the responses `Advance` are read from file into the data set `drill.data`. A log transformation is then taken so that the response `y = log10(Advance)` used in the analysis is the log of the units of drill advance. Note that the function `log10()` calculates log to the base 10, whereas `log()` would calculate log to the base e, which is the more usual transformation. Also in the second block of code, the coded factor levels 1 and 2 are converted to contrast coefficients -1 and $+1$, respectively, (e.g. for each factor, `2*(1) - 3 -> -1` and `2*(2) - 3 -> +1`). These coefficients could have been entered directly into and read directly from the data file.

In the second block of code, the linear model function `lm` fits the linear regression model `y ~ A*B*C*D`, saving the fitted model coefficient estimates and other information as `model1`. The syntax `A*B*C*D` causes all main effect and interaction coefficients for `A, B, C,` and `D` to be included in the model as regressors or predictors of `y`. The interaction coefficients are obtained by multiplication to also have values ±1 (e.g., `AB = A*B`). The statement `model1$coefficients`

Table 7.12 R program for a half-normal probability plot for the drill advance 2^4 experiment

```
# Input data for A, B, C, D and Advance
drill.data = read.table("data/drill.advance.txt", header=T)
# Compute log advance, and convert levels 1 and 2 to coeff's -1 and 1, resp.
drill.data = within(drill.data,
  { y = log10(Advance); A = 2*A-3; B = 2*B-3; C = 2*C-3; D = 2*D-3 })
head(drill.data, 3)

# Fit regression model with interactions to obtain estimates
model1 = lm(y ~ A*B*C*D, data=drill.data)
model1$coefficients # Show estimates

# Generate half-normal plot of effect estimates
# install.packages("gplots")
library(gplots)
qqnorm.aov(model1, xlab="Half-Normal Scores",
           ylab="Normalized Absolute Estimates")
```

displays the regression coefficient estimates which are half the value of the contrast estimates, including the intercept estimate. (See Chap. 8 for more information about regression.)

The last block of code calls the `qqnorm.aov` function of the `gplots` package. This function takes the estimates of the coefficients excluding the intercept, normalizes them, then generates the desired half-normal plot, plotting the normalized absolute contrast estimates versus the half-normal scores. Assuming effect sparsity, the nonnegligible contrasts are those whose estimates do not lie along a straight line through the origin.

7.7.2   Voss–Wang Confidence Interval Method

The R program in Table 7.13 illustrates computation of the Voss-Wang simultaneous confidence intervals (7.5.5) for analysis of a single-replicate experiment (Sect. 7.5.3), using the data of the drill advance experiment. The levels of *A*, *B*, *C*, and *D* together with the responses `Advance` are read from file into the data set `drill.data`, and the response `y = log10(Advance)`—the log base 10 of `Advance`—is added to the data set. The levels of the factors *A*–*D* are then converted from 1 and 2 to -1 and $+1$, respectively.

In the second block of code, the linear model function `lm` is used to fit the linear regression model `y ~ A*B*C*D`, saving the results as `model1`. The model is a regression model because the factors are not factor variables (See Chap. 8). The syntax `A*B*C*D` causes all main effects and interactions involving the variables *A*, *B*, *C* and *D* to be included in the model. The information saved includes the regression coefficient estimate and corresponding Type I sum of square for each of the 15 effects. The regression coefficients represent half the corresponding treatment effects of interest, because of the use of contrast coefficients $\pm$ which are 2 units apart. The statement

```
estimate = 2*(model1$coefficients[2:16])
```

doubles the coefficient estimates to obtain the usual treatment contrast (effect) estimates (i.e. the difference of two averages), discards the first coefficient estimate corresponding to the intercept, and saves the 15 treatment contrast estimates as `estimate`. If one would display the analysis of variance

Table 7.13   R program and output for the Voss–Wang method for the drill advance experiment

```
# Input data for A, B, C, D and Advance
drill.data = read.table("data/drill.advance.txt", header=T)
# Compute log advance, and convert levels 1 and 2 to coeff's -1 and 1, resp.
drill.data = within(drill.data,
  { y = log10(Advance); A = 2*A-3; B = 2*B-3; C = 2*C-3; D = 2*D-3 })

# Fit regression model with interactions to get estimates and SS's
model1 = lm(y ~ A*B*C*D, data=drill.data)
# Save estimates, scaled to be difference of 2 averages
estimate = 2*(model1$coefficients[2:16])
# Save sums of squares for effects
SS = anova(model1)[1:15,2]
# Order estimates and SS's in deceasing magnitude
estimate = estimate[rev(order(SS))]
SS = SS[rev(order(SS))]

# For each effect, compute msQ, stde, msd, and CIs
msQ = numeric(15) # A column with 15 cells
sse = sum(SS[8:15]) # sse = SS(8)+...+SS(15), a scalar
for(i in 1:7){msQ[i]=sse/8} # Compute msQ[1]--msQ[7]
for(i in 8:15){msQ[i] = (sse - SS[i] + SS[7])/8} # msQ[8]--msQ[15]
stde = sqrt(msQ/4); msd = 9.04*stde
LCL = estimate - msd; UCL = estimate + msd # CIs
# Display results to 5 decimal places
round( cbind(estimate, SS, msQ, stde, msd, LCL, UCL), digits=5)

        estimate      SS     msQ    stde     msd      LCL     UCL
C        0.50137 1.00550 0.00090 0.01500 0.13563  0.36574 0.63701
B        0.25193 0.25387 0.00090 0.01500 0.13563  0.11629 0.38756
D        0.14182 0.08045 0.00090 0.01500 0.13563  0.00618 0.27745
A        0.05645 0.01275 0.00090 0.01500 0.13563 -0.07918 0.19209
C:D      0.04262 0.00727 0.00090 0.01500 0.13563 -0.09301 0.17826
A:D      0.02905 0.00338 0.00090 0.01500 0.13563 -0.10658 0.16469
A:C:D    0.02312 0.00214 0.00090 0.01500 0.13563 -0.11252 0.15875
A:B:D    0.02268 0.00206 0.00091 0.01509 0.13639 -0.11371 0.15907
B:C     -0.02180 0.00190 0.00093 0.01525 0.13784 -0.15964 0.11603
A:B:C:D  0.01677 0.00113 0.00103 0.01602 0.14485 -0.12808 0.16162
B:C:D   -0.01499 0.00090 0.00106 0.01624 0.14683 -0.16182 0.13184
A:B     -0.01492 0.00089 0.00106 0.01625 0.14690 -0.16182 0.13198
B:D     -0.00649 0.00017 0.00115 0.01693 0.15305 -0.15954 0.14656
A:B:C    0.00451 0.00008 0.00116 0.01701 0.15378 -0.14927 0.15828
A:C      0.00450 0.00008 0.00116 0.01701 0.15378 -0.14929 0.15828
```

table via the statement `anova(model1)`, one would see that the sums of squares for the treatment contrasts are in rows 1–15 of column 2 of the table. The statement

```
SS = anova(model1)[1:15,2]
```

saves these 15 sums of squares as the column `SS`. The last two statements in the second block of code use the reverse order function `rev(order())` to reorder the elements of `estimate` and `SS` to be in decreasing order of magnitude. At this stage, the command `cbind(estimate, SS)` if given would display the `estimate` and `SS` columns shown in the bottom of Table 7.13, where each row corresponds to the contrast or effect identified by the row label.

The third block of code uses this information saved as `estimate` and `SS` to complete the computations. First, `msQ` is defined to be a numeric column with 15 cells—one for each effect. To help

compute the ith value `msQ[i]` for `msQ`, `sse` is assigned the value of the sum of the eight smallest sums of squares. For the ith effect, `msQ[i]` is the average of the eight smallest sums of squares, excluding the value `SS[i]` corresponding to the effect. The first `for` statement assigns the common value `msQ[i]=sse/8` to each of the first seven cells of `msQ`, since the corresponding effects do not have sum of squares `SS[i]` in the smallest eight. The second `for` statement computes `msQ[i]` for each of the last eight effects, i.e. for $i = 8, \ldots, 15$, where `msQ[i]` for the ith effect excludes the corresponding sum of squares `SS[i]` but includes the ninth smallest, `SS[7]`. The estimates were scaled to correspond to estimators with variance $\sigma^2/4$ so, using the quasi mean squared error `msQ` like an estimate of σ^2, the standard errors are estimated as `stde = sqrt(msQ/4)`. While these are not standard errors per se, they are the values of $\sqrt{msQ_k \sum_i c_i^2}$ in formula (7.5.5). So, for each effect, the minimum significant difference is `msd = 9.04*stde`, where the critical value $v_{15,8,0.05} = 9.04$ is obtained from Appendix A.11. The lower and upper confidence limit columns `LCL` and `UCL` are then computed as `estimate±msQ`. Finally, the pertinent information is column-bound and displayed in the bottom of Table 7.13, with values rounded to five decimal places.

7.7.3   Experiments with Empty Cells

We now illustrate the use of R software for the analysis of an experiment with empty cells. No new R procedures or commands are introduced, but the empty cells can cause complications. For illustration, we use the following experiment.

Example 7.7.1 `Rail weld experiment`

Wu (1964) illustrated the usefulness of two-level factorial designs using the data listed in Table 7.11. Under investigation were the effects of three factors—ambient temperature (T), wind velocity (V), and rail steel bar size (S)—on the ultimate tensile strength of welds. The factor levels were 0° and 70 °F for temperature, 0 and 20 miles per hour for wind velocity, and 4/11 and 11/11 in. for bar size, each coded as levels 1 and 2, respectively. Only six of the possible eight treatment combinations were observed, but $r = 2$ observations were taken on each of these six.

Some R commands for analyzing the rail weld experiment are presented in Table 7.14. In the first block of code, the data are read from file into the data set `rail.data`, factor variables are added to the data set, then three lines of data are displayed.

In the second block of code, an attempt is made to fit the three-way complete model, and `lsmeans` is used in an attempt to estimate the main effect of temperature `T`. Partial output is shown in Table 7.15. With two cells empty, there are data on only six treatment combinations, so there are only five degrees of freedom available for comparing treatments. This is not enough to measure the three main effects, the three two-factor interactions, and the three-factor interaction. This is indicated by the analysis of variance table, since it includes only five degrees of freedom for effects, with the effects `fT:fS` and `fT:fV:fS` unlisted. Also, the estimate of the least squares mean for $T = 1$ is listed as not applicable (NA) because it is not estimable, due to the lack of data at two of the four VS combinations at level 1 of T. Consequently, the command

```
summary(contrast(lsmT, list(T=c(-1,1))), infer=c(T,T))
```

is also non-applicable so generates no output.

To estimate contrasts, one must use a model with at most five estimable degrees of freedom. Of course, this should be anticipated ahead of time during step (g) of the checklist (Chap. 2). Figure 7.9 (p. 230) illustrates with a solid ball at the corresponding corners of the cube the treatment combinations

Table 7.14   R program for the rail weld experiment with two empty cells

```
# Input data for T V S y
rail.data = read.table("data/rail.weld.txt", header=T)
# Create factor variables, then display first 3 lines of rail.data
rail.data = within(rail.data,
  { fT = factor(T); fV = factor(V); fS = factor(S) })
head(rail.data, 3)

# Try to fit a 3-way complete model
model1 = aov(y ~ fT*fV*fS, data=rail.data); anova(model1)
# See main effects non-estimable under complete model if empty cells
library(lsmeans)
lsmT = lsmeans(model1, ~ fT)
lsmT; summary(contrast(lsmT, method="pairwise"), infer=c(T,T))

# Fit a model using 5 degrees of freedom
options(contrasts=c("contr.sum","contr.poly"))
model2 = aov(y ~ fT + fV + fS + fT:fS + fV:fS, data=rail.data)
anova(model2) # Type I ANOVA
drop1(model2, ~., test="F") # Type III ANOVA

# Estimate main effects of T, V, and S
lsmT = lsmeans(model2, ~ fT); lsmT
summary(contrast(lsmT, method="pairwise"), infer=c(T,T))
lsmV = lsmeans(model2, ~ fV); lsmV
summary(contrast(lsmV, method="pairwise"), infer=c(T,T))
lsmS = lsmeans(model2, ~ fS); lsmS
summary(contrast(lsmS, method="pairwise"), infer=c(T,T))

# Estimating interaction contrasts
lsmTS = lsmeans(model2, ~ fT:fS); lsmTS
summary(contrast(lsmTS, list(TS=c(1,-1,-1,1)/2)), infer=c(T,T))
lsmVS = lsmeans(model2, ~ fV:fS); lsmVS
summary(contrast(lsmVS, list(VS=c(1,-1,-1,1)/2)), infer=c(T,T))

# Multiple comparisons of treatment combinations
summary(contrast(lsmTS, method="pairwise"), infer=c(T,T))
```

Source Data is from Wu (1964). Copyright © 1964 American Welding Society. Reprinted with permission. (Reprinted University of Wisconsin Engineering Experiment Station, Reprint 684)

for which data are collected. One might guess that the *T V* interaction effect is not estimable, since data are only collected at three of the four combinations of levels of these two factors.

One possibility is to exclude from the complete model those interactions for which the type I degrees of freedom are zero, namely the *TV* and *TVS* interaction effects. The contrast coefficient lists for the seven factorial effects are shown in Table 7.16. It is clear that the *T* and *V* contrasts are not orthogonal to the *TV* interaction contrast, and that the *S*, *TS*, and *VS* contrasts are not orthogonal to the *TVS* interaction contrast. Consequently, the incorrect omission of *TV* and *TVS* from the model will bias the estimates of all the other contrasts. If we do decide to exclude both the *TV* and *TVS* interaction effects, then the model is of the form

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{ji} + \epsilon_{ijkl} \,.$$

We illustrate analysis of this model beginning with the third block of code in Table 7.14. The aov function fits the above model, saving the results as model2. The options statement imposes "sum

Table 7.15   Output from 3-way complete model for the rail weld experiment

```
> # Try to fit a 3-way complete model
> model1 = aov(y ~ fT*fV*fS, data=rail.data); anova(model1)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
fT         1    138   138.2    1.99   0.21
fV         1     79    79.4    1.14   0.33
fS         1      0     0.0    0.00   0.99
fT:fS      1    107   106.7    1.53   0.26
fV:fS      1     25    25.2    0.36   0.57
Residuals  6    418    69.6

> library(lsmeans)
> lsmT = lsmeans(model1, ~ fT)

NOTE: Results may be misleading due to involvement in interactions

> lsmT

 fT lsmean     SE df lower.CL upper.CL
 1     NA     NA NA       NA       NA
 2   90.5 2.9502  6   83.281   97.719

Results are averaged over the levels of: fV, fS
Confidence level used: 0.95
```

Table 7.16   Contrast coefficients for the observed treatment combinations (TC) in the rail weld experiment

TC	*T*	*V*	*TV*	*S*	*TS*	*VS*	*TVS*
111	−1	−1	1	−1	1	1	−1
112	−1	−1	1	1	−1	−1	1
211	1	−1	−1	−1	−1	1	1
212	1	−1	−1	1	1	−1	−1
221	1	1	1	−1	−1	−1	−1
222	1	1	1	1	1	1	1

to zero" constraints on least squares estimates as needed to generate the correct Type III analysis of variance. The `anova` and `drop1` statements generate Type I and Type III analyses, respectively, shown in the top of Table 7.17.

The contrasts for *T* and *V* are not orthogonal to each other, but they can be estimated (although with a small positive correlation). Similar comments apply to the *S*, *TS*, and *VS* contrasts. The `lsmeans`, `summary`, and `contrast` statements of the `lsmeans` package are used to estimate least square means and contrasts, and also for multiple comparisons of the levels of a factor or the combinations of levels of multiple factors. Appropriate syntax is illustrated by the last three blocks of code in Table 7.14. Sample output for factor *T*, including least squares means and the pairwise contrast, is given in the bottom of Table 7.17.

Table 7.17 Partial output from the second call of `aov`

```
> anova(model2) # Type I ANOVA

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
fT         1    138   138.2    1.99   0.21
fV         1     79    79.4    1.14   0.33
fS         1      0     0.0    0.00   0.99
fT:fS      1    107   106.7    1.53   0.26
fV:fS      1     25    25.2    0.36   0.57
Residuals  6    418    69.6

> drop1(model2, ~., test="F") # Type III ANOVA

Single term deletions

Model:
y ~ fT + fV + fS + fT:fS + fV:fS
      Df Sum of Sq RSS  AIC F value Pr(>F)
<none>             418 54.6
fT     1     214.2 632 57.6    3.08   0.13
fV     1      79.4 497 54.7    1.14   0.33
fS     1      29.6 447 53.4    0.43   0.54
fT:fS  1     131.2 549 55.9    1.88   0.22
fV:fS  1      25.2 443 53.3    0.36   0.57

> lsmT = lsmeans(model2, ~ fT); lsmT

 fT lsmean     SE df lower.CL upper.CL
 1   80.15 5.1100  6   67.646   92.654
 2   90.50 2.9502  6   83.281   97.719

Results are averaged over the levels of: fV, fS
Confidence level used: 0.95

> summary(contrast(lsmT, method="pairwise"), infer=c(T,T))

 contrast estimate     SE df lower.CL upper.CL t.ratio p.value
 1 - 2      -10.35 5.9005  6  -24.788    4.088  -1.754  0.1300

Results are averaged over the levels of: fV, fS
Confidence level used: 0.95
```

Consistent with the results of the Type III analysis of variance, none of the contrasts estimates generated by the R program in Table 7.14 would appear particularly strong, (only one is shown in Table 7.14). □

Exercises

1. For the following hypothetical data sets of Sect. 7.2.2 reproduced below, draw interaction plots to evaluate the *BC* and *ABC* interaction effects, with levels of *B* on the horizontal axis and levels of *C*

for labels. In each case, comment on the apparent presence or absence of BC and ABC interaction effects.

(a)

ijk :	111	112	121	122	211	212	221	222	311	312	321	322
$\overline{y}_{ijk.}$:	3.0	4.0	1.5	2.5	2.5	3.5	3.0	4.0	3.0	4.0	1.5	2.5

(b)

ijk :	111	112	121	122	211	212	221	222	311	312	321	322
$\overline{y}_{ijk.}$:	3.0	2.0	1.5	4.0	2.5	3.5	3.0	4.0	3.0	5.0	3.5	6.0

2. In planning a five-factor experiment, it is determined that the factors A, B, and C might interact and the factors D and E might interact but that no other interaction effects should be present. Draw a line graph for this experiment and give an appropriate model.

3. Consider an experiment with four treatment factors, A, B, C, and D, at a, b, c, and d levels, respectively, with r observations per treatment combination. Assume that the four-way complete model is a valid representation of the data. Use the rules of Sect. 7.3 to answer the following.

   (a) Find the number of degrees of freedom associated with the AC interaction effect.
   (b) Obtain an expression for the sum of squares for AC.
   (c) Give a rule for testing the hypothesis that the AC interaction is negligible against the alternative hypothesis that it is not negligible. How should the results of the test be interpreted, given the other terms in the model?
   (d) Write down a contrast for measuring the AC interaction. Give an expression for its least squares estimate and associated variance.
   (e) Give a rule for testing the hypothesis that your contrast in part (d) is negligible.

4. **Popcorn–microwave experiment, continued**

   In the popcorn–microwave experiment of Sect. 7.4 (p. 213), the experimenters studied the effects of popcorn brand, microwave oven power, and cooking time on the percentage of popped kernels in packages of microwave popcorn. Suppose that, rather than using a completely randomized design, the experimenters first collected all the observations for one microwave oven, followed by all observations for the other microwave oven. Would you expect the assumptions on the three-way complete model to be satisfied? Why or why not?

5. **Weathering experiment**

   An experiment is described in the paper "Accelerated weathering of marine fabrics"(Moore, M. A. and Epps, H. H., *Journal of Testing and Evaluation* 20, 1992, 139–143). The purpose of the experiment was to compare the effects of different types of weathering on the breaking strength of marine fabrics used for sails. The factors of interest were

   F: Fabric at 3 levels (1 = polyester, 2 = acrylic, 3 = nylon).
   E: Exposure conditions (1 = continuous light at 62.7 °C, 2 = alternating 30 min light and 15 min condensation).
   A: Exposure levels (1 = 1200 AFU, 2 = 2400 AFU, 3 = 3600 AFU).
   D: Direction of cut of the fabric (1 = warp direction, 2 = filling direction).

   In total there were $v = 3 \times 2 \times 3 \times 2 = 36$ treatment combinations, and $r = 2$ observations were taken on each. The response variable was "percent change in breaking strength of fabric after exposure to weathering conditions." The average response for each of the 36 treatment combinations is shown in Table 7.18. The error mean square was calculated to be 6.598 with 36 degrees of freedom.

Table 7.18 Percent change in breaking strength of fabrics after exposure

| Exposure | AFU | Direction | Fabric (F) | | |
(E)	(A)	(D)	1	2	3
1	1	1	−43.0	−1.7	−74.7
		2	−46.1	+11.7	−86.7
	2	1	−45.3	−4.2	−87.9
		2	−51.3	+10.0	−97.9
	3	1	−53.3	−5.1	−98.2
		2	−54.5	+7.5	−100.0
2	1	1	−48.1	−6.8	−85.0
		2	−43.6	−3.3	−91.7
	2	1	−52.3	−4.2	−100.0
		2	−53.8	−3.3	−100.0
	3	1	−56.5	−5.9	−100.0
		2	−56.4	−6.7	−100.0

Source Moore and Epps (1992). Copyright © ASTM. Reprinted with permission

(a) How would you decide whether or not the error variables have approximately the same variance for each fabric?

(b) Using the cell-means model, test the hypothesis $H_0 : [\tau_1 = \cdots = \tau_{36}]$ against the alternative hypothesis $H_A :$ [at least two τ_i's differ]. What can you conclude?

(c) Write down a contrast in the treatment combinations that compares the polyester fabric with the nylon fabric. Is your contrast estimable?

(d) If your contrast in (c) is estimable, give a formula for the least squares estimator and its variance. Otherwise, go to part (e).

(e) Assuming that you are likely to be interested in a very large number of contrasts and you want your overall confidence level to be 95%, calculate a confidence interval for any pairwise comparison of your choosing. What does the interval tell you?

(f) Calculate a 90% confidence bound for σ^2.

(g) If you were to repeat this experiment and you wanted your confidence interval in (e) to be of length at most 20%, how many observations would you take on each treatment combination?

6. **Weathering experiment, continued**

Suppose you were to analyze the weathering experiment described in Exercise 5 using a four-way complete model.

(a) What conclusions can you draw from the analysis of variance table?

(b) Give an explicit formula for testing that the *FA*-interaction is negligible.

(c) Would confidence intervals for differences in fabrics be of interest? If not, why not? If so, how would they be interpreted? Give a formula for such confidence intervals assuming that these intervals are preplanned and are the only intervals envisaged, and the overall level is to be at least 99%.

(d) In the original paper, the authors write "Fabric direction (*D*) had essentially no effect on percent change in breaking strength for any of the fabrics." Do you agree with this statement? Explain.

Table 7.19 Data for the coating experiment

A	2	2	2	2	2	2	2	2
B	2	2	2	2	1	1	1	1
C	2	2	1	1	2	2	1	1
D	2	1	2	1	2	1	2	1
y_{ijkl}	5.95	4.57	4.03	2.17	3.43	1.02	4.25	2.13
A	1	1	1	1	1	1	1	1
B	2	2	2	2	1	1	1	1
C	2	2	1	1	2	2	1	1
D	2	1	2	1	2	1	2	1
y_{ijkl}	12.28	9.57	6.73	6.07	8.49	4.92	6.95	5.31

Source Data adapted from Saravanan et al. (2001). Published by the Journal of Physics D: Applied Physics

7. **Coating experiment**

   P. Saravanan, V. Selvarajan, S. V. Joshi, and G. Sundararajan (2001, *Journal of Physics D: Applied Physics*) described an experiment to study the effect of different spray parameters on thermal spray coating properties. In the experiment, the authors attempted to produce high-quality alumina (Al_2O_3) coatings by controlling the fuel ratio (factor A at 1:2.8 and 1:2.0), carrier gas flow rate (factor B at 1.33 and 3.21 L s^{-1}), frequency of detonations (factor C at 2 and 4 Hz), and spray distance (factor D at 180 and 220 mm). To quantify the quality of the coating, the researchers measured multiple response variables. In this example we will examine the porosity (vol. %). The data are shown in Table 7.19.

   (a) Assuming that 3- and 4-factor interactions are negligible, outline an analysis that you would wish to perform for such an experiment (step (g) of the checklist; see Chap. 2).
   (b) Check the assumptions on your model.
   (c) Carry out the analysis that you outlined in part (a), including drawing any interaction plots that may be of interest. State your conclusions clearly.

8. **Paper towel strength experiment**

   Burt Beiter, Doug Fairchild, Leo Russo, and Jim Wirtley, in 1990, ran an experiment to compare the relative strengths of two similarly priced brands of paper towel under varying levels of moisture saturation and liquid type. The treatment factors were "amount of liquid" (factor A, with levels 5 and 10 drops coded 1 and 2), "brand of towel" (factor B, with levels coded 1 and 2), and "type of liquid" (factor C, with levels "beer" and "water" coded 1 and 2). A $2 \times 2 \times 2$ factorial experiment with $r = 3$ was run in a completely randomized design. The resulting data, including run order, are given in Table 7.20.

   (a) The experimenters assumed only factors A and B would interact. Specify the corresponding model.
   (b) List all treatment contrasts that are likely to be of primary interest to the experimenters.
   (c) Use residual plots to evaluate the adequacy of the model specified in part (a).
   (d) Provide an analysis of variance table for this experiment, test the various effects, show plots of significant main effects and interactions, and draw conclusions.

Table 7.20 Data for paper towel strength experiment: A = "amount of liquid," B = "brand of towel," and C = "liquid type"

ABC	Strength	(Order)	Strength	(Order)	Strength	(Order)
111	3279.0	(3)	4330.7	(15)	3843.7	(16)
112	3260.8	(11)	3134.2	(20)	3206.7	(22)
121	2889.6	(5)	3019.5	(6)	2451.5	(21)
122	2323.0	(1)	2603.6	(2)	2893.8	(14)
211	2964.5	(4)	4067.8	(10)	3327.0	(18)
212	3114.2	(12)	3009.3	(13)	3242.0	(19)
221	2883.4	(9)	2581.4	(23)	2385.9	(24)
222	2142.3	(7)	2364.9	(8)	2189.9	(17)

Table 7.21 Thrust duration (in seconds) for the rocket experiment

A	B	C_0				C_1			
		D_0	D_1	D_2	D_3	D_0	D_1	D_2	D_3
0	0	21.60	11.54	19.09	13.11	21.60	11.50	21.08	11.72
0	1	21.09	11.14	21.31	11.26	22.17	11.32	20.44	12.82
1	0	21.60	11.75	19.50	13.72	21.86	9.82	21.66	13.03
1	1	19.57	11.69	20.11	12.09	21.86	11.18	20.24	12.29
Total		83.86	46.12	80.01	50.18	87.49	43.82	83.42	49.86

Source Wood and Hartvigsen (1973). Copyright © 1964 American Society for Quality. Reprinted with permission

(e) Construct confidence intervals for each of the treatment contrasts that you listed in part (b), using an appropriate method of multiple comparisons. Discuss the results.

9. **Rocket experiment**

S. R. Wood and D. E. Hartvigsen describe an experiment in the 1964 issue of *Industrial Quality Control* on the testing of an auxiliary rocket engine. According to the authors, the rocket engine must be capable of satisfactory operation after exposure to environmental conditions encountered during storage, transportation, and the in-flight environment. Four environmental factors were deemed important. These were vibration (Factor A; absent, present, coded 0, 1), temperature cycling (Factor B; absent, present, coded 0, 1), altitude cycling (Factor C; absent, present, coded 0, 1) and firing temperature/altitude (Factor D, 4 levels, coded 0, 1, 2, 3). The response variable was "thrust duration," and the observations are shown in Table 7.21, where C_k and D_l denote the kth level of C and the lth level of D, respectively.

The experimenters were willing to assume that the 3-factor and 4-factor interactions were negligible.

(a) State a reasonable model for this experiment, including any assumptions on the error term.
(b) How would you check the assumptions on your model?
(c) Calculate an analysis of variance table and test any relevant hypotheses, stating your choice of the overall level of significance and your conclusions.
(d) Levels 0 and 1 of factor D represent temperatures $-75°F$ and $170°F$, respectively at sea level. Level 2 of D represents $-75°F$ at 35,000 ft. Suppose the experimenters had been interested in

Table 7.22   Manganese data for the spectrometer experiment

			A_1		A_2		A_3	
C	D	E	B_1	B_2	B_1	B_2	B_1	B_2
1	1	1	0.9331	0.9214	0.8664	0.8729	0.8711	0.8627
1	1	2	0.9253	0.9399	0.8508	0.8711	0.8618	0.8785
1	2	1	0.8472	0.8417	0.7948	0.8305	0.7810	0.8009
1	2	2	0.8554	0.8517	0.7810	0.7784	0.7887	0.7853
2	1	1	0.9253	0.9340	0.8879	0.8729	0.8618	0.8692
2	1	2	0.9301	0.9272	0.8545	0.8536	0.8720	0.8674
2	2	1	0.8435	0.8674	0.7879	0.8009	0.7904	0.7793
2	2	2	0.8463	0.8526	0.7784	0.7863	0.7939	0.7844
3	1	1	0.9146	0.9272	0.8769	0.8683	0.8591	0.8683
3	1	2	0.9399	0.9488	0.8739	0.8729	0.8729	0.8481
3	2	1	0.8499	0.8417	0.7893	0.8009	0.7893	0.7904
3	2	2	0.8472	0.8300	0.7913	0.7904	0.7956	0.7827
	Total		10.6578	10.6836	9.9331	9.9991	9.9376	9.9172

Source Inman et al. (1992). Reprinted with Permission from Journal of Quality Technology © 1992 ASQ, www.asq.org

two preplanned contrasts. The first compares the effects of levels 0 and 1 of D, and the second compares the effects the levels 0 and 2 of D. Using an overall level of at least 98%, give a set of simultaneous confidence intervals for these two contrasts.

(e) Test the hypotheses that each contrast identified in part (d) is negligible. Be explicit about which method you are using and your choice of the overall level of significance.

(f) If the contrasts in part (d) had not been preplanned, would your answer to (d) have been different? If so, give the new calculations.

(g) Although it may not be of great interest in this particular experiment, draw an interaction plot for the CD interaction and explain what it shows.

(h) If the experimenters had included the 3-factor and 4-factor interactions in the model, how could they have decided upon the important main effects and interactions?

10. **Spectrometer experiment**

A study to determine the causes of instability of measurements made by a Baird spectrometer during production at North Star Steel Iowa was reported by J. Inman, J. Ledolter, R. V. Lenth, and L. Niemi in the *Journal of Quality Technology* in 1992. A brainstorming session with members of the Quality Assurance and Technology Department of the company produced a list of five factors that could be controlled and could be the cause of the observed measurement variability. The factors and their selected experimental levels were:

A: Temperature of the lab. (67°, 72°, 77°).
B: Cleanliness of entrance window seal (clean, one week's use).
C: Placement of sample (sample edge tangential to edge of disk, sample completely covering disk, sample partially covering disk).
D: Wear of boron nitride disk (new, one month old).
E: Sharpness of counter electrode tip (newly sharpened, one week's wear).

Spectrometer measurements were made on several different elements. The manganese measurements are shown in Table 7.22, where A_i and B_j denote the ith level of A and the jth level of B,

respectively. The experimenters were willing to assume that the 4-factor and 5-factor interactions were negligible.

(a) Test any relevant hypotheses, at a 0.05 overall level of significance, and state your conclusions.
(b) Draw an interaction plot for the AE interaction. Does the plot show what you expected it to show? Why or why not? (Mention AE, A, and E.)
(c) The spectrometer manual recommends that the placement of the sample be at level 2. Using level 2 as a control level, give confidence intervals comparing the other placements with the control placement. You may assume that these comparisons were preplanned. State which method you are using and give reasons for your choice. Use an overall confidence level of at least 98% for these two intervals.
(d) Test the hypotheses of no linear and quadratic trends in the manganese measurements due to temperature. Use a significance level of 0.01 for each test.

11. **Antifungal antibiotic experiment**

M. Gupte and P. Kulkarni (2003, *Journal of Chemical Technology and Biotechnology*) described an experiment to maximize the yield of an antifungal antibiotic from the isolate *Thermomonospora sp MTCC 3340*. The researchers examined the effect of the three factors temperature of incubation (factor A at 25, 30, and 37 C), concentration of carbon (factor B at 2, 5, and 7.5%), and concentration of nitrogen (factor C at 0.5, 1, and 3%) on the antifungal yield which was measured in terms of activity against *Candida albicans*, a type of fungus that can be detrimental to humans. The data are shown in Table 7.23.

(a) Construct appropriate plots to assess whether any of the main effects seem to have a significant effect on the response. What do you conclude?
(b) What assumption regarding interactions did you make while drawing your conclusion in part (a)?
(c) Construct an appropriate plot to asses the significance of two-way interactions. Do any two-way interactions seem to have a significant effect on the response? If so, does this affect your conclusions from part (c)?
(d) Suppose there is reason to believe that the three factors do not jointly interact. Fit a model that includes all main effects and two-way interactions. What effects do you find to be significant? How does this compare with your conjectures from part (c)?
(e) Do the assumptions of normality and equal error variances hold for the model considered in part (d)? Are there any outliers?

12. **Antifungal antibiotic experiment, continued**

Consider the data from Table 7.23, but without assuming that the three-factor interaction is negligible. Also, for the purposes of this particular exercise, we change the third levels of factors A and B to be 35 and 8, respectively, so that their levels are equally spaced.

(a) Make a table similar to that of Table 7.1, p. 208, with the first column containing the 27 treatment combinations for the antifungal antibiotic experiment in ascending order. List the contrast coefficients for the main effect trend contrasts: Linear A, Quadratic A, Linear B, and Quadratic B. Also list the contrast coefficients for the interaction trend contrasts Linear $A \times$ Linear B, Linear $A \times$ Quadratic B, Quadratic $A \times$ Linear B, Quadratic $A \times$ Quadratic B.

Table 7.23 Data for the antifungal antibiotic experiment

A	1	1	1	1	1	1	1	1	1
B	1	1	1	2	2	2	3	3	3
C	1	2	3	1	2	3	1	2	3
y_{ijk}	25.84	51.86	32.59	20.48	25.84	12.87	20.48	25.84	10.20
A	2	2	2	2	2	2	2	2	2
B	1	1	1	2	2	2	3	3	3
C	1	2	3	1	2	3	1	2	3
y_{ijk}	51.86	131.33	41.11	41.11	104.11	32.59	65.42	82.53	51.86
A	3	3	3	3	3	3	3	3	3
B	1	1	1	2	2	2	3	3	3
C	1	2	3	1	2	3	1	2	3
y_{ijk}	41.11	104.11	32.59	32.59	82.53	25.84	51.86	65.42	41.11

Source Gupte and Kulkarni (2003). Journal of Chemical Technology and Biotechnology Published by John Wiley and Sons. Reprinted with permission

(b) What divisors are needed to normalize each of the contrasts? Calculate, by hand, the least squares estimates for the normalized contrasts Linear A and Quadratic A.

(c) The levels of C are not equally spaced. Select two orthogonal contrasts that compare the levels of C and add these to your table in part (a).

(d) Use a computer program (similar to that of Table 7.8 or 7.12) to calculate the least squares estimates of a complete set of 26 orthogonal normalized contrasts that measure the main effects of A, B, and C and their interactions. Prepare a half-normal probability plot of the 26 contrast estimates. Explain what you can conclude from the plot.

(e) Use the method of Voss and Wang (Sect. 7.5.3) to examine a complete set of 26 orthogonal normalized contrasts that measure the main effects of A, B, and C and their interactions. Compare your conclusions with those obtained from part 12.

13. **Paper towel experiment, continued**

Consider the paper towel strength experiment of Exercise 8. Suppose that *only the first ten observations* had been collected. These are labeled (1)–(10) in Table 7.20, p. 240.

(a) Is it possible to perform an analysis of variance of these data, using a model that includes main effects and the AB interaction as required by the experimenters? If so, analyze the experiment.

(b) Use a computer program to fit a three-way complete model. Can all of the main effects and interactions be measured? If not, investigate which models could have been used in the analysis of such a design with two empty cells and unequal numbers of observations in the other cells.

14. **Abrasive wear experiment**

To improve the characteristics of some metals, scientists combine them with another metal or element. The resulting mixture is called an alloy. Alloys are widely used in engineering applications in order to reduce cost, improve physical or chemical properties of materials, etc. O. P. Modi, R. P. Yadav, D. P. Mondal, R. Dasgupta, S. Das, and A. H. Yegneswaran (2001, *Journal of Materials Science*) described an experiment to study the effects of three factors on the wear rate (m³/m) of a zinc-aluminum alloy. The three factors were sliding distance (factor A at 25 and 125 m), applied

Table 7.24 Data for the abrasive wear experiment

A	1	1	1	1	2	2	2	2
B	1	1	2	2	1	1	2	2
C	1	2	1	2	1	2	1	2
y_{ijk}	0.049	0.041	0.220	0.358	0.044	0.030	0.133	0.192

Source Data adapted from Modi et al. (2001), Journal of Materials Science. Published by Kluwer Academic Publishers

load (factor B at 1 and 7 N), and abrasive size (factor C at 23 and 275 μm). The data were run in a random order and are listed in Table 7.24.

(a) Calculate the least squares estimates for a set of seven orthogonal contrasts, measuring the main effects and interactions of the three factors.

(b) Draw a half-normal probability plot of the seven contrast estimates. Although $m = 7$ contrasts is too few to be able to draw good conclusions about the main effects and interactions, which contrasts should be investigated in more detail later?

(c) Use the Voss–Wang procedure to examine the seven contrasts used in part (b).

(d) Based on parts (b) and (c), what conclusions can you draw about the effects of the three factors?

15. **Steel bar experiment**

Baten (1956, *Industrial Quality Control*) described an experiment that investigated the cause of variability of the length of steel bars in a manufacturing process. Each bar was processed with one of two different heat treatments (factor A, levels 1, 2) and was cut on one of four different screw machines (factor B, levels 1, 2, 3, 4) at one of three different times of day (factor C, levels 8 am, 11 am, 3 pm, coded 1, 2, 3). There were considerable differences in the lengths of the bars after cutting, and a purpose for this experiment was to try to determine whether there were assignable causes for this variation.

(a) Discuss possible ways to design and analyze this experiment, but assume that it needs to be run in a working factory. In your discussion, consider using

   (i) a completely randomized design,
   (ii) a randomized block design,
   (iii) a design with times of day (factor C) regarded as a block factor,

(b) The randomization employed by the experimenter is not specified in the published article, and we proceed as though it were run as a completely randomized design with the three factors A, B, and C described above. List some of the sources of variation that must have been deemed as minor and ignored.

(c) The data that were collected by the experimenter are shown in Table 7.25. There are $r = 4$ observations on each of the $v = 24$ treatment combinations. The data values are "$y_{ijkt} =$ (length $- 4.38$) $\times$ 1000 in." Check the assumptions on the three-way complete model for these data. (You may wish to remove an outlier). If the assumptions are satisfied, calculate an analysis of variance table. What are your conclusions?

(d) The desired length for each bar was 4.385±0.005 in., which means that the desired value for the response y_{ijkt} is 5 units. Calculate confidence intervals for the true mean lengths of the bars cut on the four machines. Which machines appear to give bars closest to specification?

Table 7.25 Data for the steel bar experiment

ABC	y_{1jk1}	y_{1jk2}	y_{1jk3}	y_{1jk4}	ABC	y_{2jk1}	y_{2jk2}	y_{2jk3}	y_{2jk4}
111	6	9	1	3	211	4	6	0	1
112	6	3	1	−1	212	3	1	1	−2
113	5	4	9	6	213	6	0	3	7
121	7	9	5	5	221	6	5	3	4
122	8	7	4	8	222	6	4	1	3
123	10	11	6	4	223	8	7	10	0
131	1	2	0	4	231	−1	0	0	1
132	3	2	1	0	232	2	0	−1	1
133	−1	2	6	1	233	0	−2	4	−4
141	6	6	7	3	241	4	5	5	4
142	7	9	11	6	242	9	4	6	3
143	10	5	4	8	243	4	3	7	0

Source Baten (1956). Copyright 1956 American Society for Quality. Reprinted with permission

16. **Ice melting experiment**

An experiment to gain a better understanding of the role of various substances in melting ice was run in 2004 by Shuangling He, Mimi Lou, Xiaozhou Xiong, Li Yu, and Yihong Zhao. For each observation, a 10 ml block of ice was placed into water containing a given concentration of sugar, or salt, or a sugar and salt mix. The melting time for the ice block was measured in seconds and then converted to minutes. The three factors of interest were shape of ice block (factor A with levels 1 - lozenge, 2 - cylinder, and 3 - cube), solute (factor B with levels 1 - sugar, 2 - salt, and 3 - equal parts sugar and salt), and concentration (factor C at 5%, 10% 15%, 20%; coded 1, 2, 3, 4).

The experiment was run as a completely randomized design and the resulting data are shown in Table 7.26. There are two observations on each treatment combination.

Table 7.26 Data (in minutes) for the ice melting experiment

Concentration	Lozenge					
	Sugar		Salt		Sugar/Salt	
5%	54.25	53.33	37.17	37.50	42.75	43.83
10%	52.50	52.25	28.33	28.83	39.33	40.00
15%	47.25	48.00	21.17	21.33	35.83	37.00
20%	43.83	44.33	15.50	16.33	25.50	26.17
Concentration	Cylinder					
	Sugar		Salt		Sugar/Salt	
5%	87.00	85.83	64.33	62.83	76.50	77.17
10%	83.50	82.00	51.83	51.50	67.08	67.33
15%	78.50	79.50	36.00	37.33	55.50	55.67
20%	69.00	67.83	25.00	25.50	45.83	46.50
Concentration	Cube					
	Sugar		Salt		Sugar/Salt	
5%	65.83	65.00	54.00	53.17	55.17	54.83
10%	61.83	62.50	42.50	43.83	47.00	48.50
15%	57.50	58.67	33.00	32.50	43.50	44.83
20%	54.50	55.00	27.50	28.33	36.72	35.15

(a) Explain briefly what you would randomize if you were running this experiment and why randomization might be important here.

(b) If the experiment had to be run in two labs with a different technician in each lab, how would you change the design and the randomization?

(c) Use the three-way complete model (7.2.2) and state the error assumptions on the model. Check these assumptions are valid.

(d) Test the single hypothesis that lozenge shaped ice blocks melt faster than the other two shapes on average (averaged over solutes and concentrations). State your null and alternative hypotheses and use significance level 0.01.

(e) Now consider the equivalent cell-means model shown in (7.2.1) with the error assumptions that you listed in part (c) and $i = 1, 2, 3; \; j = 1, 2, 3; \; k = 1, 2, 3, 4; \; t = 1, 2$.

  (i) Give a formula for a set of 95% set of confidence intervals for the true differences in the effects of the treatment combinations. Which method are you using?

  (ii) Using the method in (i), calculate a confidence interval for the difference between the melting times for a cylinder-shaped ice block with 5% concentration of sugar, and a cube-shaped ice block with 20% concentration of salt.

(f) Obtain a 90% upper confidence bound for σ^2.

8.1 Introduction

In each of the previous chapters we were concerned with experiments that were run as completely randomized designs for the purpose of investigating the effects of one or more treatment factors on a response variable. Analysis of variance and methods of multiple comparisons were used to analyze the data. These methods are applicable whether factor levels are qualitative or quantitative.

In this chapter, we consider an alternative approach for quantitative factors, when the set of possible levels of each factor is real-valued rather than discrete. We restrict attention to a single factor and denote its levels by x. The mean response $E[Y_{xt}]$ is modeled as a polynomial function of the level x of the factor, and the points $(x, E[Y_{xt}])$ are called the *response curve*. For example, if $E[Y_{xt}] = \beta_0 + \beta_1 x$ for unknown parameters β_0 and β_1, then the mean response is a linear function of x and the response curve is a line, called the *regression line*. Using data collected at various levels x, we can obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the intercept and slope of the line. Then $\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$ provides an estimate of $E[Y_{xt}]$ as a function of x, and it can be used to estimate the mean response or to predict the values of new observations for any factor level x, including values for which no data have been collected. We call $\hat{y}_x$ the *fitted model* or the *estimated mean response* at the level x.

In Sect. 8.2, we look at polynomial regression and the fit of polynomial response curves to data. Estimation of the parameters in the model, using the method of least squares, is discussed in the optional Sect. 8.3. In Sect. 8.4, we investigate how well a regression model fits a given set of data via a "lack-of-fit" test. In Sect. 8.5, we look at the analysis of a simple linear regression model and test hypotheses about the values of the model parameters. Confidence intervals are also discussed. The general analysis of a higher-order polynomial regression model using a computer package is discussed in Sect. 8.6. Investigation of linear and quadratic trends in the data via orthogonal polynomials is the topic of optional Sect. 8.7. An experiment is examined in detail in Sect. 8.8, and analysis using the SAS and R software packages is done in Sects. 8.9 and 8.10, respectively.

Polynomial regression methods can be extended to experiments involving two or more quantitative factors. The mean response $E[Y_{xt}]$ is then a function of several variables and defines a *response surface* in three or more dimensions. Specialized designs are usually required for fitting response surfaces, and consequently, we postpone their discussion to Chap. 16.

8.2 Models

The standard model for polynomial regression is

$$Y_{xt} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \epsilon_{xt} , \tag{8.2.1}$$

$$\epsilon_{xt} \sim N(0, \sigma^2) ,$$

$$\epsilon_{xt}\text{'s are mutually independent}$$

$$t = 1, \ldots, r_x; \quad x = x_1, \ldots, x_v.$$

The treatment factor is observed at v different levels $x_1, \ldots, x_v$. There are r_x observations taken when the treatment factor is at level x, and Y_{xt} is the response for the tth of these. The responses Y_{xt} are modeled as independent random variables with mean

$$E[Y_{xt}] = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p ,$$

which is a pth-degree polynomial function of the level x of the treatment factor. Since $\epsilon_{xt} \sim N(0, \sigma^2)$, it follows that

$$Y_{xt} \sim N(\beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p, \ \sigma^2) .$$

Typically, in a given experiment, the exact functional form of the true response curve is unknown. In polynomial regression, the true response curve is assumed to be well approximated by a polynomial function. If the true response curve is relatively smooth, then a low-order polynomial function will often provide a good model, at least for a limited range of levels of the treatment factor.

If $p = 1$ in the polynomial regression function, we have the case known as *simple linear regression*, for which the mean response is

$$E[Y_{xt}] = \beta_0 + \beta_1 x ,$$

which is a linear function of x. This model assumes that an increase of one unit in the level of x produces a mean increase of β_1 in the response, and is illustrated in Fig. 8.1. At each value of x, there is a normal distribution of possible values of the response, the mean of which is the corresponding point, $E[Y_{xt}] = \beta_0 + \beta_1 x$, on the regression line and the variance of which is σ^2.

Consider now the data plotted in Fig. 8.2, for which polynomial regression might be appropriate. Envisage a normal distribution of possible values of Y_{xt} for each level x, and a smooth response curve connecting the distribution of their means, $E[Y_{xt}]$. It would appear that a quadratic response curve

Fig. 8.1 Simple linear regression model
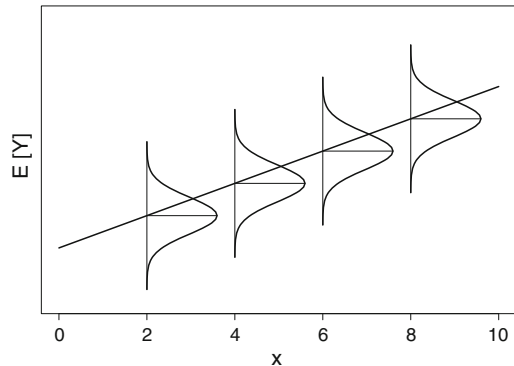
Fig. 8.2   Three
hypothetical observations
y_{xt} at each of five
treatment factor levels
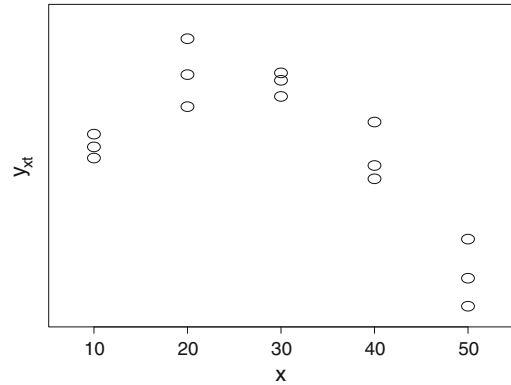
may provide a good fit to these data. This case, for which

$$E[Y_{xt}] = \beta_0 + \beta_1 x + \beta_2 x^2,$$

is called *quadratic regression*. If this model is adequate, the fitted quadratic model can be used to
estimate the value of x for which the mean response is maximized, even though it may not occur at
one of the x values for which data have been collected.

Although regression models can be used to estimate the mean response at values of x that have not
been observed, estimation outside the range of observed x values must be done with caution. There is
no guarantee that the model provides a good fit outside the observed range.

If observations are collected for v distinct levels x of the treatment factor, then any polynomial
regression model of degree $p \leq v - 1$ (that is, with v or fewer parameters) can be fitted to the
data. However, it is generally preferable to use the simplest model that provides an adequate fit. So
for polynomial regression, lower-order models are preferred. Higher-order models are susceptible to
overfit, a circumstance in which the model fits the data too well at the expense of having the fitted
response curve vary or fluctuate excessively between data points. Over-fit is illustrated in Fig. 8.3,
which contains plots for a simple linear regression model and a sixth-degree polynomial regression
model, each fitted to the same set of data. The sixth-degree polynomial model provides the better fit
in the sense of providing a smaller value for the sum of squared errors. However, since we may be
looking at natural fluctuation of data around a true linear model, it is arguable that the simple linear
regression model is actually a better model—better for predicting responses at new values of x, for
example. Information concerning the nature of the treatment factor and the response variable may shed
light on which model is more likely to be appropriate.

Least Squares Estimates

Once data are available, we can use the method of least squares to find estimates $\hat{\beta}_j$ of the parameters
β_j of the chosen regression model. The fitted model is then

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \cdots + \hat{\beta}_p x^p,$$

and the error sum of squares is

$$ssE = \sum_x \sum_t (y_{xt} - \hat{y}_x)^2.$$

Fig. 8.3 Data and fitted
linear and sixth-degree
polynomial regression
models
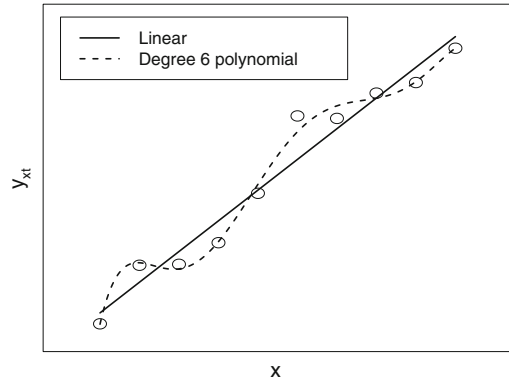
The number of error degrees of freedom is the number of observations minus the number of parameters
in the model; that is, $n - (p + 1)$. The mean squared error,

$$msE = \sum_x \sum_t (y_{xt} - \hat{y}_x)^2 / (n - p - 1) \,,$$

provides an unbiased estimate of σ^2.

In the following optional section, we obtain the least squares estimates of the parameters β_0 and β_1
in a simple linear regression model. However, in general we leave the determination of least squares
estimates to a computer, since the formulae are not easily expressed without the use of matrices, and
the hand computations are generally tedious. An exception to this occurs with the use of orthogonal
polynomial models, discussed in Sect. 8.7.

Checking Model Assumptions

Having made an initial selection for the degree of polynomial model required in a given scenario, the
model assumptions should be checked. The first assumption to check is that the proposed polynomial
model for $E[Y_{xt}]$ is indeed adequate. This can done either by examination of a plot of the residuals
versus x or by formally testing for model lack of fit. The standard test for lack of fit is discussed in
Sect. 8.4.

If no pattern is apparent in a plot of the residuals versus x, this indicates that the model is adequate.
Lack of fit is indicated if there is a clear function-like pattern. For example, suppose a quadratic model
is fitted but a cubic model is needed. Any linear or quadratic pattern in the data would then be explained
by the model and would not be evident in the residual plot, but the residual plot would show the pattern
of a cubic polynomial function unexplained by the fitted model (see Fig. 8.4).

Residual plots can also be used to assess the assumptions on the random error terms in the model in
the same way as discussed in Chap. 5. The residuals are plotted versus run order to evaluate indepen-
dence of the error variables, plotted versus fitted values $\hat{y}_x$ to check the constant variance assumption
and to check for outliers, and plotted versus the normal scores to check the normality assumption.

If the error assumptions are not valid, the fitted line still provides a model for mean response.
However, the results of confidence intervals and hypothesis tests can be misleading. Departures from
normality are generally serious problems only when the true error distribution has long tails or when
prediction of a single observation is required. Nonconstant variance can sometimes be corrected via
transformations, as in Chap. 5, but this may also change the order of the model that needs to be fitted.

If no model assumptions are invalidated, then analysis of variance can be used to determine whether
or not a simpler model would suffice than the one postulated by the experimenter (see Sect. 8.6).
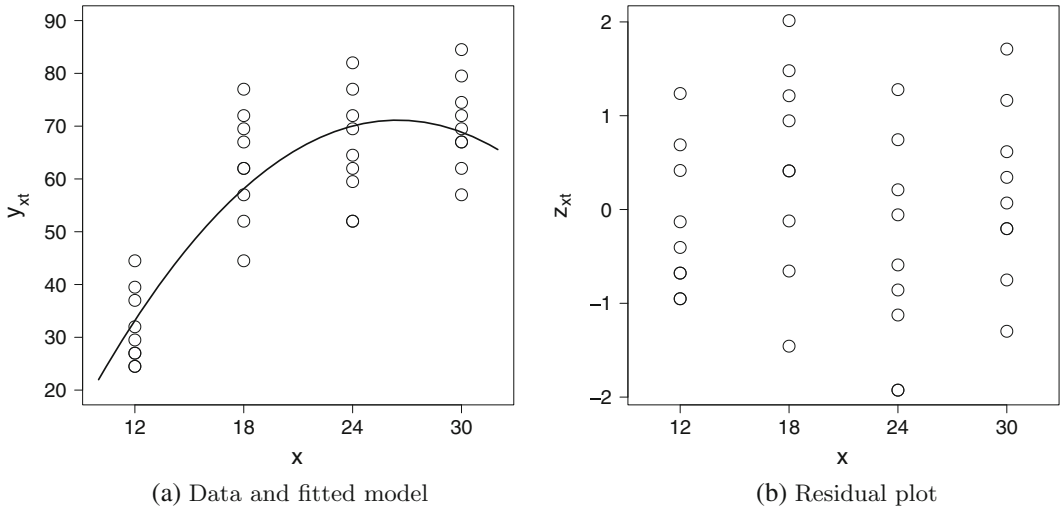
Fig. 8.4 Plots for a quadratic polynomial regression model fitted to data from a cubic model

8.3 Least Squares Estimation (Optional)

In this section, we derive the normal equations for a general polynomial regression model. These equations can be solved to obtain the set of least squares estimates $\hat{\beta}_j$ of the parameters β_j. We illustrate this for the case of simple linear regression.

8.3.1 Normal Equations

For the pth-order polynomial regression model (8.2.1), the normal equations are obtained by differentiating the sum of squared errors

$$\sum_x \sum_t e_{xt}^2 = \sum_x \sum_t (y_{xt} - \beta_0 - \beta_1 x - \cdots - \beta_p x^p)^2$$

with respect to each parameter and setting each derivative equal to zero. For example, if we differentiate with respect to β_j, set the derivative equal to zero, and replace each β_i with $\hat{\beta}_i$, we obtain the jth normal equation as

$$\sum_x \sum_t x^j y_{xt} = \sum_x \sum_t x^j \left(\hat{\beta}_0 + x\hat{\beta}_1 + \cdots + x^p \hat{\beta}_p \right) . \tag{8.3.2}$$

We have one normal equation of this form for each value of j, $j = 0, 1, \ldots, p$. Thus, in total, we have $p + 1$ equations in $p + 1$ unknowns $\hat{\beta}_j$. Provided that the number of levels of the treatment factor exceeds the number of parameters in the model (that is, $v \geq p + 1$), there is a unique solution to the normal equations giving a unique set of least squares estimates, with the result that all parameters are estimable.

8.3.2   Least Squares Estimates for Simple Linear Regression

For the simple linear regression model, we have $p = 1$, and there are two normal equations obtained from (8.3.2) with $j = 0, 1$. These are

$$\sum_x \sum_t y_{xt} = n\hat{\beta}_0 + \sum_x \sum_t x\hat{\beta}_1 \, ,$$

$$\sum_x \sum_t xy_{xt} = \sum_x \sum_t x\hat{\beta}_0 + \sum_x \sum_t x^2\hat{\beta}_1 \, ,$$

where $n = \Sigma_x r_x$ denotes the total number of observations in the experiment. Dividing the first equation by n, we obtain

$$\hat{\beta}_0 = \overline{y}_{..} - \hat{\beta}_1\overline{x}_{..} \, , \tag{8.3.3}$$

where $\overline{x}_{..} = \sum_x r_x x/n$. Substituting this into the second equation gives

$$\hat{\beta}_1 = \frac{\sum_x \sum_t xy_{xt} - n\overline{x}_{..}\overline{y}_{..}}{ss_{xx}} \, , \tag{8.3.4}$$

where $ss_{xx} = \sum_x r_x(x - \overline{x}_{..})^2$.

8.4   Test for Lack of Fit

We illustrate the lack-of-fit test via the quadratic regression model

$$E[Y_{xt}] = \beta_0 + \beta_1 x + \beta_2 x^2 \, .$$

If data have been collected for only three levels $x = x_1, x_2, x_3$ of the treatment factor, then the fitted model $\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ will pass through the sample means $\overline{y}_{x.}$ computed at each value of x. This means that the predicted response $\hat{y}_x$ at the observed values of x is $\hat{y}_x = \overline{y}_{x.}$ (for $x = x_1, x_2, x_3$). This is the same fit as would be obtained using the one-way analysis of variance model, so we know that it is the best possible fit of a model to the data in the sense that no other model can give a smaller sum of squares for error, ssE.

   If observations have been collected at more than three values of x, however, then the model is unlikely to fit the data perfectly, and in general, $\hat{y}_x \neq \overline{y}_{x.}$. If the values $\hat{y}_x$ and $\overline{y}_{x.}$ are too far apart relative to the amount of variability inherent in the data, then the model does not fit the data well, and there is said to be model *lack of fit*. In other words, in our example, the quadratic function is not sufficient to model the mean response $E[Y_{xt}]$.

   If there is replication at one or more of the x-values, and if data are collected at more than three x-values, then it is possible to conduct a test for lack-of-fit of the quadratic model. The null hypothesis is that the quadratic model is adequate for modeling mean response; that is,

$$H_0^Q : E[Y_{xt}] = \beta_0 + \beta_1 x + \beta_2 x^2 \, .$$

The alternative hypothesis is that a more general model (the one-way analysis of variance model) is needed; that is,

$$H_A^Q : E[Y_{xt}] = \mu + \tau_x \, ,$$

where τ_x is the effect on the response of the treatment factor at level x. We fit the quadratic regression model and obtain ssE and $msE = ssE/(n-3)$. Now, MSE is an unbiased estimator of the error variance if the quadratic model is correct, but otherwise it has expected value larger than σ^2.

At each level x where more than one observation has been taken, we can calculate the sample variance s_x^2 of the responses. Each sample variance s_x^2 is an unbiased estimator of the error variance, σ^2, and these can be pooled to obtain the *pooled sample variance*,

$$s_p^2 = \left[\sum_x (r_x - 1)s_x^2 \right] /(n - v) . \tag{8.4.5}$$

Provided that the assumption of equal error variances is valid, the pooled sample variance is an unbiased estimator of σ^2 even if the model does not fit the data well. This pooled sample variance is called the *mean square for pure error* and denoted by $msPE$. An alternative way to compute $msPE$ is as the mean square for error obtained by fitting the one-way analysis of variance model.

The test of lack of fit, which is the test of H_0^Q versus H_A^Q, is based on a comparison of the two fitted models (the quadratic model and the one-way analysis of variance model), using the difference in the corresponding error sums of squares. We write ssE for the error sum of squares obtained from the quadratic regression model and $ssPE$ for the error sum of squares from the one-way analysis of variance model. Then the *sum of squares for lack of fit* is

$$ssLOF = ssE - ssPE .$$

The sum of squares for pure error has $n - v$ degrees of freedom associated with it, whereas the sum of squares for error has $n - (p + 1) = n - 3$ (since there are $p + 1 = 3$ parameters in the quadratic regression model). The number of degrees of freedom for lack of fit is therefore $(n-3)-(n-v) = v-3$. The corresponding *mean square for lack of fit*,

$$msLOF = ssLOF/(v - 3),$$

measures model lack of fit because it is an unbiased estimator of σ^2 if the null hypothesis is true but has expected value larger than σ^2 otherwise.

Under the polynomial regression model (8.2.1) for $p = 2$, the decision rule for testing H_0^Q versus H_A^Q at significance level α is

$$\text{reject } H_0^Q \text{ if } msLOF/msPE > F_{v-3,n-v,\alpha} .$$

In general, a polynomial regression model of degree p can be tested for lack of fit as long as $v > p + 1$ and there is replication for at least one of the x-levels. A test for lack of fit of the pth-degree polynomial regression model is a test of the null hypothesis

$$H_0^p : \{ E[Y_{xt}] = \beta_0 + \beta_1 x + \cdots + \beta_p x^p; \ x = x_1, \ldots, x_v \}$$

versus the alternative hypothesis

$$H_A^p : \{ E[Y_{xt}] = \mu + \tau_x; \ x = x_1, \ldots, x_v \} .$$

The decision rule at significance level α is

Table 8.1 Hypothetical data for one continuous treatment factor

x		y_{xt}		$\bar{y}_{x.}$	s_x^2
10	69.42	66.07	71.70	69.0633	8.0196
20	79.91	81.45	85.52	82.2933	8.4014
30	88.33	82.01	84.43	84.9233	10.1681
40	62.59	70.98	64.12	65.8967	19.9654
50	25.86	32.73	24.39	27.6600	19.8189

Table 8.2 Test for lack of fit of quadratic regression model for hypothetical data

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Lack of fit	2	30.0542	15.0271	1.13	0.3604
Pure error	10	132.7471	13.2747		
Error	12	162.8013			

$$\text{reject } H_0^p \text{ if } msLOF/msPE > F_{v-p-1,n-v,\alpha},$$

where

$$msLOF = ssLOF/(v - p - 1) \quad \text{and} \quad ssLOF = ssE - ssPE.$$

Here, ssE is the error sum of squares obtained by fitting the polynomial regression model of degree p, and $ssPE$ is the error sum of squares obtained by fitting the one-way analysis of variance model.

Example 8.4.1 `Lack-of-fit test for quadratic regression`

In this example we conduct a test for lack of fit of a quadratic polynomial regression model, using the hypothetical data that were plotted in Fig. 8.2 (p. 251). Table 8.1 lists the $r = 3$ observations for each of $v = 5$ levels x of the treatment factor, together with the sample mean and sample variance. The pooled sample variance (8.4.5) is

$$s_p^2 = msPE = \sum_x 2s_x^2/(15 - 5) = 13.2747,$$

and the sum of squares for pure error is therefore

$$ssPE = (15 - 5)msPE = 132.7471.$$

Alternatively, this can be obtained as the sum of squares for error from fitting the one-way analysis of variance model.

The error sum of squares ssE is obtained by fitting the quadratic polynomial regression model using a computer program (see Sects. 8.9 and 8.10 for achieving this via SAS and R software, respectively). We obtain $ssE = 162.8013$. Thus

$$ssLOF = ssE - ssPE = 162.8013 - 132.7471 = 30.0542$$

with

$$v - p - 1 = 5 - 2 - 1 = 2$$

degrees of freedom. The test for lack of fit is summarized in Table 8.2. Since the p-value is large, there is no significant lack of fit. The quadratic model seems to be adequate for these data. □

8.5   Analysis of the Simple Linear Regression Model

Suppose a linear regression model has been postulated for a given scenario, and a check of the model assumptions finds no significant violations including lack of fit. Then it is appropriate to proceed with analysis of the data.

It was shown in the optional Sect. 8.3 that the least squares estimates of the intercept and slope parameters in the simple linear regression model are

$$\hat{\beta}_0 = \overline{y}_{..} - \hat{\beta}_1 \overline{x}_{..} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_x \sum_t x y_{xt} - n \overline{x}_{..} \overline{y}_{..}}{ss_{xx}} , \tag{8.5.6}$$

where $\overline{x}_{..} = \sum_x r_x x / n$ and $ss_{xx} = \sum_x r_x (x - \overline{x}_{..})^2$. The corresponding estimators (random variables), which we also denote by $\hat{\beta}_0$ and $\hat{\beta}_1$, are normally distributed, since they are linear combinations of the normally distributed random variables Y_{xt}. In Exercise 1, the reader is asked to show that the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ are equal to

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}_{..}^2}{ss_{xx}} \right) \quad \text{and Var}(\hat{\beta}_1) = \sigma^2 \left(\frac{1}{ss_{xx}} \right) . \tag{8.5.7}$$

If we estimate σ^2 by

$$msE = \frac{\sum_x \sum_t (y_{xt} - (\hat{\beta}_0 + \hat{\beta}_1 x))^2}{n - 2} , \tag{8.5.8}$$

it follows that

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{msE \left(\frac{1}{n} + \frac{\overline{x}_{..}^2}{ss_{xx}} \right)}} \sim t_{n-2} \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{msE \left(\frac{1}{ss_{xx}} \right)}} \sim t_{n-2} .$$

Thus, the decision rule at significance level α for testing whether or not the intercept is equal to a specific value a ($H_0^{\text{int}} : \{\beta_0 = a\}$ versus $H_A^{\text{int}} : \{\beta_0 \neq a\}$) is

$$\text{reject } H_0^{\text{int}} \quad \text{if} \quad \frac{\hat{\beta}_0 - a}{\sqrt{msE \left(\frac{1}{n} + \frac{\overline{x}_{..}^2}{ss_{xx}} \right)}} > t_{n-2, \alpha/2} \text{ or } < t_{n-2, 1-\alpha/2} . \tag{8.5.9}$$

The decision rule at significance level α for testing whether or not the slope of the regression model is equal to a specific value b ($H_0^{\text{slp}} : \{\beta_1 = b\}$ versus $H_A^{\text{slp}} : \{\beta_1 \neq b\}$) is

$$\text{reject } H_0^{\text{slp}} \quad \text{if} \quad \frac{\hat{\beta}_1 - b}{\sqrt{msE \left(\frac{1}{ss_{xx}} \right)}} > t_{n-2, \alpha/2} \text{ or } < t_{n-2, 1-\alpha/2} . \tag{8.5.10}$$

Corresponding one-tailed tests can be constructed by choosing the appropriate tail of the t distribution and replacing $\alpha/2$ by α.

Confidence intervals at individual confidence levels of $100(1-\alpha)\%$ for β_0 and β_1 are, respectively,

$$\hat{\beta}_0 \;\pm\; t_{n-2,\alpha/2}\sqrt{msE\left(\frac{1}{n}+\frac{\overline{x}_{..}^2}{ss_{xx}}\right)} \tag{8.5.11}$$

and

$$\hat{\beta}_1 \;\pm\; t_{n-2,\alpha/2}\sqrt{msE\left(\frac{1}{ss_{xx}}\right)}. \tag{8.5.12}$$

We can use the regression line to estimate the expected mean response $E[Y_{xt}]$ at any particular value of x, say x_a; that is,

$$\widehat{E}[Y_{x_at}] = \hat{y}_{x_a} = \hat{\beta}_0 + \hat{\beta}_1 x_a \,.$$

The variance associated with this estimator is

$$\mathrm{Var}(\hat{Y}_{x_a}) = \sigma^2\left(\frac{1}{n}+\frac{(x_a-\overline{x}_{..})^2}{ss_{xx}}\right).$$

Since $\hat{Y}_{x_a}$ is a linear combination of the normally distributed random variables $\hat{\beta}_0$ and $\hat{\beta}_1$, it, too, has a normal distribution. Thus, if we estimate σ^2 by msE given in (8.5.8), we obtain a $100(1-\alpha)\%$ confidence interval for the expected mean response at x_a as

$$\hat{\beta}_0 + \hat{\beta}_1 x_a \;\pm\; t_{n-2,\alpha/2}\sqrt{msE\left(\frac{1}{n}+\frac{(x_a-\overline{x}_{..})^2}{ss_{xx}}\right)}. \tag{8.5.13}$$

A confidence "band" for the entire regression line can be obtained by calculating confidence intervals for the mean response at all values of x. Since this is an extremely large number of intervals, we need to use Scheffé's method of multiple comparisons. So, a $100(1-\alpha)\%$ confidence band for the regression line is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x_a \;\pm\; \sqrt{2F_{2,n-2,\alpha}}\sqrt{msE\left(\frac{1}{n}+\frac{(x_a-\overline{x}_{..})^2}{ss_{xx}}\right)}. \tag{8.5.14}$$

The critical coefficient here is $w = \sqrt{2\,F_{2,n-2,\alpha}}$ rather than the value $w = \sqrt{(v-1)\,F_{v-1,n-v,\alpha}}$ that we had in the one-way analysis of variance model, since there are only two parameters of interest in our model (instead of linear combinations of $v-1$ pairwise comparisons) and the number of error degrees of freedom is $n-2$ rather than $n-v$.

Finally, we note that it is also possible to use the regression line to predict a future observation at a particular value x_a of x. The predicted value $\hat{y}_{x_a}$ is the same as the estimated mean response at x_a obtained from the regression line; that is,

$$\hat{y}_{x_a} = \hat{\beta}_0 + \hat{\beta}_1 x_a \,.$$

The variance associated with this prediction is larger by an amount σ^2 than that associated with the estimated mean response, since the model acknowledges that the data values are distributed around their mean according to a normal distribution with variance σ^2. Consequently, we may adapt (8.5.13) to obtain a $100(1-\alpha)\%$ *prediction interval* for a future observation at x_a, as follows:

Table 8.3 Fluid flow in liters/minute for the heart–lung pump experiment

rpm	Liters per minute				
50	1.158	1.128	1.140	1.122	1.128
75	1.740	1.686	1.740		
100	2.340	2.328	2.328	2.340	2.298
125	2.868	2.982			
150	3.540	3.480	3.510	3.504	3.612

$$\hat{\beta}_0 + \hat{\beta}_1 x_a \;\pm\; t_{n-2,1-\alpha/2} \sqrt{msE \left(1 + \frac{1}{n} + \frac{(x_a - \bar{x}_{..})^2}{ss_{xx}} \right)}. \tag{8.5.15}$$

Alternatively, the prediction interval follows, because

$$\frac{\hat{Y}_{x_a} - Y_{x_a}}{\sqrt{msE \left(1 + \frac{1}{n} + \frac{(x_a - \bar{x}_{..})^2}{ss_{xx}} \right)}} \sim t(n-2)$$

under our model.

Example 8.5.1 Heart-lung pump experiment, continued

In Example 4.2.3, p. 73, a strong linear trend was discovered in the fluid flow rate as the number of revolutions per minute increases in a rotary pump head of an Olson heart–lung pump. Consequently, a simple linear regression model may provide a good model for the data. The data are reproduced in Table 8.3. It can be verified that

$$\bar{x}_{..} \;=\; \sum_x r_x x/n \;=\; [5(50) + 3(75) + 5(100) + 2(125) + 5(150)]/20 \;=\; 98.75\,,$$

and

$$\bar{y}_{..} = 2.2986 \text{ and } \sum_x \sum_t x y_{xt} = 5212.8\,.$$

So,

$$\begin{aligned} ss_{xx} \;&=\; [5(-48.75)^2 + 3(-23.75)^2 + 5(1.25)^2 + 2(26.25)^2 + 5(51.25)^2] \\ &=\; 28,093.75\,, \end{aligned}$$

giving

$$\begin{aligned} \hat{\beta}_1 \;&=\; [5212.8 \;-\; 20(98.75)(2.2986)]/[28,093.75] \\ &=\; 673.065/28,093.75 \;=\; 0.02396\,. \end{aligned}$$

The mean square for error (8.5.8) for the regression model is best calculated by a computer package. It is equal to $msE = 0.001177$, so the estimated variance of $\hat{\beta}_1$ is

$$\text{Var}(\hat{\beta}_1) \;=\; msE/ss_{xx} \;=\; (0.001177)/28,093.75 \;=\; 0.000000042\,.$$

A 95% confidence interval for β_1 is then given by (8.5.12), as

$$
\begin{aligned}
0.02396 \;&\pm\; t_{18,.025}\sqrt{0.000000042}\,, \\
0.02396 \;&\pm\; (2.101)(0.00020466)\,, \\
&(\,0.02353,\; 0.02439\,)\,.
\end{aligned}
$$

To test the null hypothesis $H_0^{\mathrm{slp}} : \{\beta_1 = 0\}$, against the one-sided alternative hypothesis $H_A^{\mathrm{slp}} : \{\beta_1 > 0\}$ that the slope is greater than zero at significance level $\alpha = 0.01$, we use a one-sided version of the decision rule (8.5.10) and calculate

$$
\frac{\hat{\beta}_1 - 0}{\sqrt{msE\left(\frac{1}{ss_{xx}}\right)}} \;=\; \frac{0.02396}{0.00020466} \;=\; 117.07\,,
$$

and since this is considerably greater than $t_{18,0.01} = 2.552$, we reject H_0^{slp}. We therefore conclude that the slope of the regression line is greater than zero, and the fluid flow increases as the revolutions per minute increase.   □

8.6 Analysis of Polynomial Regression Models

8.6.1 Analysis of Variance

Suppose a polynomial regression model has been postulated for a given experiment, and the model assumptions appear to be satisfied, including no significant lack of fit. Then it is appropriate to proceed with analysis of the data. A common objective of the analysis of variance is to determine whether or not a lower-order model might suffice. One reasonable approach to the analysis, which we demonstrate for the quadratic model ($p = 2$), is as follows.

First, test the null hypothesis $H_0^L : \beta_2 = 0$ that the highest-order term $\beta_2 x^2$ is not needed in the model so that the simple linear regression model is adequate. If this hypothesis is rejected, then the full quadratic model is needed. Otherwise, testing continues and attempts to assess whether an even simpler model is suitable. Thus, the next step is to test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$. If this is rejected, the simple linear regression model is needed and adequate. If it is not rejected, then x is apparently not useful in modeling the mean response.

Each test is constructed in the usual way, by comparing the error sum of squares of the full (quadratic) model with the error sum of squares of the reduced model corresponding to the null hypothesis being true. For example, to test the null hypothesis $H_0^L : \beta_2 = 0$ that the simple linear regression model is adequate versus the alternative hypothesis H_A^L that the linear model is not adequate, the decision rule at significance level α is

$$\text{reject } H_0^L \quad \text{if} \quad ms(\beta_2)/msE > F_{1,n-v,\alpha}\,,$$

where the mean square $ms(\beta_2) = ss(\beta_2)/1$ is based on one degree of freedom, and

$$ss(\beta_2) = ssE_1 - ssE_2\,,$$

where ssE_1 and ssE_2 are the error sums of squares obtained by fitting the models of degree one and two, respectively.

Table 8.4 Analysis of variance table for polynomial regression model of degree p. Here ssE_b denotes the error sum of squares obtained by fitting the polynomial regression model of degree b

Source of variation	Degrees of freedom	Sum of square	Mean squares	Ratio
β_p	1	$ssE_{p-1} - ssE$	$ms(\beta_p)$	$ms(\beta_p)/msE$
β_{p-1}, β_p	2	$ssE_{p-2} - ssE$	$ms(\beta_{p-1}, \beta_p)$	$ms(\beta_{p-1}, \beta_p)/msE$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\beta_2, \ldots, \beta_p$	$p - 1$	$ssE_1 - ssE$	$ms(\beta_2, \ldots, \beta_p)$	$ms(\beta_2, \ldots, \beta_p)/msE$
Model	p	$ssE_0 - ssE$	$ms(\beta_1, \ldots, \beta_p)$ $ms(\beta_1, \ldots, \beta_p)/msE$	
Error	$n - p - 1$	ssE	msE	
Total	$n - 1$	$sstot$		

Similarly, the decision rule at significance level α for testing $H_0 : \beta_1 = \beta_2 = 0$ versus the alternative hypothesis that H_0 is false is

$$\text{reject } H_0 \quad \text{if} \quad ms(\beta_1, \beta_2)/msE > F_{2,n-v,\alpha} \,,$$

where the mean square $ms(\beta_1, \beta_2) = ss(\beta_1, \beta_2)/2$ is based on 2 degrees of freedom, and

$$ss(\beta_1, \beta_2) = (ssE_0 - ssE_2)/2 \,,$$

and ssE_0 and ssE_2 are the error sums of squares obtained by fitting the models of degree zero and two, respectively.

The tests are generally summarized in an analysis of variance table, as indicated in Table 8.4 for the polynomial regression model of degree p. In the table, under sources of variability, "Model" is listed rather than "$\beta_1, \ldots, \beta_p$" for the test of $H_0 : \beta_1 = \cdots = \beta_p = 0$, since this is generally included as standard output in a computer package. Also, to save space, we have written the error sum of squares as ssE for the full model, rather than indicating the order of the model with a subscript p. Analysis of variance for quadratic regression ($p = 2$) is illustrated in the following example.

Example 8.6.1 `Analysis of variance for quadratic regression`

Consider the hypothetical data in Table 8.1, p. 256, with three observations for each of the levels $x = 10, 20, 30, 40, 50$. For five levels, the quartic model is the highest-order polynomial model that can be fitted to the data. However, a quadratic model was postulated for these data, and a test for lack of fit of the quadratic model, conducted in Example 8.4.1, suggested that this model is adequate.

The analysis of variance for the quadratic model is given in Table 8.5. The null hypothesis $H_0^L :$ $\{\beta_2 = 0\}$ is rejected, since the p-value is less than 0.0001. So, the linear model is not adequate, and the quadratic model is needed. This is no surprise, based on the plot of the data shown in Fig. 8.5. Now, suppose the objective of the experiment was to determine how to maximize mean response. From the data plot, it appears that the maximum response occurs within the range of the levels x that were observed. The fitted quadratic regression model can be obtained from a computer program, as illustrated in Sects. 8.9 and 8.10 for the SAS and R programs, respectively. The fitted model is

Table 8.5 Analysis of variance for the quadratic model

Source of variation	Degrees of freedom	Sum of square	Mean squares	Ratio	p-value
β_2	1	3326.2860	3326.2860	245.18	0.0001
Model	2	6278.6764	3139.3382	231.40	0.0001
Error	12	162.8013	13.5668		
Total	14	6441.4777			

Fig. 8.5 Quadratic polynomial regression model fitted to hypothetical data
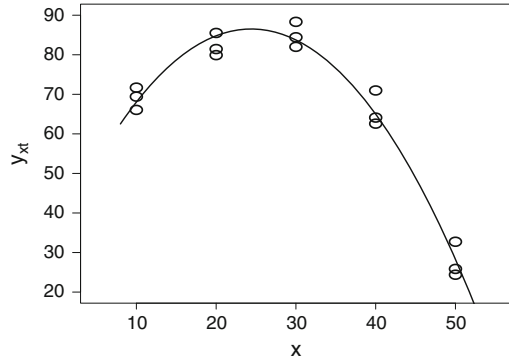
$$\hat{y}_x = 33.43333 + 4.34754x - 0.08899x^2 \,,$$

and is plotted in Fig. 8.5 along with the raw data. The fitted curve achieves its maximum value when x is around 24.4, which should provide a good estimate of the level x that maximizes mean response. Further experimentation involving levels around this value could now be done. □

The adequacy of a regression model is sometimes assessed in terms of the proportion of variability in the response variable that is explained by the model. This proportion, which is the ratio of the model sum of squares to the sum of squares total, is called the *coefficient of multiple determination*, or the R^2-*value*. In the notation of Table 8.4,

$$R^2 = (ssE_0 - ssE)/sstot = ss(\beta_1, \ldots, \beta_p)/sstot. \tag{8.6.1}$$

For simple linear regression,

$$R^2 = ss(\beta_1)/sstot$$

is called the *coefficient of determination*, and in this case $R^2 = r^2$, where

$$r = ss_{xy}/\sqrt{ss_{xx}ss_{yy}}$$

is the *sample correlation coefficient*, or *Pearson product-moment correlation coefficient*.

8.6.2   Confidence Intervals

When the model is fitted via a computer program, the least squares estimates of $\hat{\beta}_j$ and their corresponding standard errors (estimated standard deviations) usually form part of the standard computer

output. If the model assumptions are satisfied, then

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{\mathrm{Var}}\,(\hat{\beta}_j)}} \sim t_{n-p-1}\,.$$

Individual confidence intervals can be obtained for the model parameters, as we illustrated in Sect. 8.5 for the simple linear regression model. The general form is

$$\hat{\beta}_j \ \pm \ t_{n-2,\alpha/2}\,\sqrt{\widehat{\mathrm{Var}}\,(\hat{\beta}_j)}\,.$$

Most programs will also allow calculation of the estimated mean response at any value of $x = x_a$ together with its standard error, and also calculation of the predicted response at $x = x_a$ plus its standard error. Confidence and prediction intervals for these can again be calculated using the t_{n-p-1} distribution. The confidence interval formula for mean response at $x = x_a$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_a + \cdots + \hat{\beta}_p x_a^p \ \pm \ t_{n-p-1,\alpha/2}\,\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}_{x_a})}$$

and the prediction interval formula for a new observation at $x = x_a$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_a + \cdots + \hat{\beta}_p x_a^p \ \pm \ t_{n-p-1,\alpha/2}\,\sqrt{\hat{\sigma}^2 + \widehat{\mathrm{Var}}\,(\hat{Y}_{x_a})}\,.$$

The overall confidence level for all the intervals combined should be computed via the Bonferroni method as usual. A confidence band for the regression line is obtained by calculating confidence intervals for the estimated mean response at all values of x, using the critical coefficient for Scheffé's method; that is,

$$\hat{\beta}_0 + \hat{\beta}_1 x + \cdots + \hat{\beta}_p x^p \ \pm \ \sqrt{(p+1)\,F_{p+1,n-p-1,\alpha}}\,\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}_x)}\,.$$

8.7   Orthogonal Polynomials and Trend Contrasts (Optional)

The normal equations for polynomial regression were presented in Eq. (8.3.2). It was noted that solving the equations can be tedious. However, the factor levels can be transformed in such a way that the least squares estimates have a simple algebraic form and are easily computed. Furthermore, the parameter estimators become uncorrelated and are multiples of the corresponding trend contrast estimators. This transformation is illustrated in this section for simple linear regression and for quadratic regression, when the factor levels x are equally spaced with equal numbers r of observations per level.

8.7.1   Simple Linear Regression

Consider the simple linear regression model, for which

$$Y_{xt} = \beta_0 + \beta_1 x + \epsilon_{xt}\,; \quad x = x_1, \ldots, x_v; \quad t = 1, \ldots, r\,. \tag{8.7.17}$$

When there are r observations on each of the v quantitative levels x of the treatment factor, the average value of x is $\overline{x}_{..} = r \sum_x x/n = \sum_x x/v$. The transformation $z_x = x - \overline{x}_{..}$ centers the levels x at zero, so that $\sum_x z_x = 0$. This makes the estimates of the slope and intercept parameters uncorrelated (or orthogonal). We can replace x in model (8.7.17) by z_x, so that the "centered" form of the model is

$$Y_{xt} = \beta_0^* + \beta_1^* z_x + \epsilon_{xt} ; \quad x = x_1, \ldots, x_v; \quad t = 1, \ldots, r . \tag{8.7.18}$$

A transformation of the independent variable changes the interpretation of some of the parameters. For example, in the simple linear regression model (8.7.17), β_0 denotes mean response when $x = 0$, whereas in the transformed model (8.7.18), β_0^* denotes mean response when $z_x = 0$, which occurs when $x = \overline{x}_{..}$.

The normal equations corresponding to $j = 0$ and $j = 1$ for the centered model are obtained from (8.3.2) with z_x in place of x. Thus, we have

$$\sum_x \sum_t y_{xt} = \sum_x \sum_t \left(\hat{\beta}_0^* + z_x \hat{\beta}_1^* \right) , \; = vr\hat{\beta}_0^*$$
$$\sum_x \sum_t z_x y_{xt} = \sum_x \sum_t z_x \left(\hat{\beta}_0^* + z_x \hat{\beta}_1^* \right) = \sum_x rz_x^2 \hat{\beta}_1^* .$$

Solving these equations gives the least squares estimates as

$$\hat{\beta}_0^* = \overline{y}_{..} \quad \text{and} \quad \hat{\beta}_1^* = \frac{1}{r \sum_x z_x^2} \sum_x \sum_t z_x y_{xt} .$$

Now,

$$\text{Cov}\left(Y_{..}, \sum_x \sum_t z_x Y_{xt} \right) = \sum_x \sum_t z_x \text{Cov}(Y_{xt}, Y_{xt}) = r\sigma^2 \sum_x z_x = 0 ,$$

so the estimators $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ are uncorrelated.

We now consider a special case to illustrate the relationship of the slope estimator with the linear trend contrast that we used in Sect. 4.2.4. Suppose equal numbers of observations are collected at the three equally spaced levels

$$x_1 = 5, \quad x_2 = 7, \quad \text{and} \quad x_3 = 9 .$$

Then $\overline{x}_{..} = 7$, so

$$z_5 = -2, \quad z_7 = 0, \quad \text{and} \quad z_9 = 2 .$$

These values are twice the corresponding linear trend contrast coefficients $(-1, 0, 1)$ listed in Appendix A.2. Now, $r = 2$, so $r \sum_x z_x^2 = 8r$, and

$$\hat{\beta}_1^* = \frac{1}{r \sum_x z_x^2} \sum_x \sum_t z_x y_{z_x t} = \frac{1}{8r} (2y_{9.} - 2y_{5.})$$
$$= \frac{1}{4} (\overline{y}_{9.} - \overline{y}_{5.}) ,$$

which is a quarter of the value of the linear trend contrast estimate. It follows that β_1^* and the linear trend contrast have the same normalized estimate and hence also the same sum of squares. Thus, testing $H_0 : \beta_1^* = 0$ under model (8.7.18) is analogous to testing the hypothesis $H_0 : \tau_3 - \tau_1 = 0$ of no linear

trend effect under the one-way analysis of variance model

$$Y_{it} = \mu + \tau_i + \epsilon_{it} \, ; \quad i = 1, 2, 3 \, ; \quad t = 1, 2 \, ,$$

where τ_i is the effect on the response of the ith coded level of the treatment factor. The one difference is that in the first case, the model is the linear regression model ($p = 1$), while in the second case, the model is the one-way analysis of variance model, which is equivalent to a model of order $p = v - 1 = 2$. Thus the two models will not yield the same mean squared error, so the F-statistics will not be identical.

8.7.2   Quadratic Regression

Consider the quadratic regression model, for which

$$Y_{xt} = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon_{xt} \, . \tag{8.7.19}$$

Assume that the treatment levels $x = x_1, \ldots, x_v$ are equally spaced, with r observations per level. To achieve orthogonality of estimates, it is necessary to transform both the linear and the quadratic independent variables.

Let $z_x = x - \bar{x}_{..}$ as in the case of simple linear regression, so that again $\sum_x z_x = 0$. Similarly, define

$$z_x^{(2)} = z_x^2 - \sum_x z_x^2 / v \, .$$

Then $\sum_x z_x^{(2)} = 0$. Also, writing z_i for the ith value of z_x in rank order, we note that since the levels x are equally spaced,

$$z_i = -z_{v+1-i} \quad \text{and} \quad z_i^{(2)} = z_{v+1-i}^{(2)} \, ,$$

so $\sum_x z_x z_x^{(2)} = 0$. These conditions give uncorrelated parameter estimators. To see this, consider the transformed model

$$Y_{xt} = \beta_0^* + \beta_1^* z_x + \beta_2^* z_x^{(2)} + \epsilon_{xt} \, . \tag{8.7.20}$$

The normal equations (8.3.2) become

$$
\begin{aligned}
\sum_x \sum_t y_{xt} &= \sum_x \sum_t \left(\hat{\beta}_0^* + z_x \hat{\beta}_1^* + z_x^{(2)} \hat{\beta}_2^* \right) &&= vr\hat{\beta}_0^* \, , \\
\sum_x \sum_t z_x y_{xt} &= \sum_x \sum_t z_x \left(\hat{\beta}_0^* + z_x \hat{\beta}_1^* + z_x^{(2)} \hat{\beta}_2^* \right) &&= r \sum_x z_x^2 \hat{\beta}_1^* \, , \\
\sum_x \sum_t z_x^{(2)} y_{xt} &= \sum_x \sum_t z_x^{(2)} \left(\hat{\beta}_0^* + z_x \hat{\beta}_1^* + z_x^{(2)} \hat{\beta}_2^* \right) &&= r \sum_x (z_x^{(2)})^2 \hat{\beta}_2^* \, .
\end{aligned}
$$

The least squares estimates, obtained by solving the normal equations, are

$$\hat{\beta}_0^* = \bar{y}_{..} \, , \quad \hat{\beta}_1^* = \frac{\sum_x \sum_t z_x y_{xt}}{r \sum_x z_x^2} \, , \quad \hat{\beta}_2^* = \frac{\sum_x \sum_t z_x^{(2)} y_{xt}}{r \sum_x (z_x^{(2)})^2} \, .$$

The estimators $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ are unchanged from the simple linear regression model (8.7.18), so they remain uncorrelated. Similarly, $\hat{\beta}_0^*$ and $\hat{\beta}_2^*$ are uncorrelated, because

$$\text{Cov}\left(Y_{..},\ \sum_x \sum_t z_x^{(2)} Y_{xt}\right) = r\sigma^2 \sum_x z_x^{(2)} = 0.$$

Observe that $\text{Cov}(\hat{\beta}_1^*, \hat{\beta}_2^*)$ is also zero, since it is proportional to

$$\text{Cov}\left(\sum_x \sum_t z_x Y_{xt},\ \sum_x \sum_t z_x^{(2)} Y_{xt}\right) = r\sigma^2 \sum_x z_x z_x^{(2)} = 0.$$

The transformed variables z_x and $z_x^{(2)}$ are called *orthogonal polynomials*, because they are polynomial functions of the levels x and give rise to uncorrelated parameter estimators $\hat{\beta}_0^*$, $\hat{\beta}_1^*$, and $\hat{\beta}_2^*$. It was illustrated in the previous subsection on simple linear regression that the values z_x are multiples of the coefficients of the linear trend contrast. Likewise, the values $z_x^{(2)}$ are multiples of the coefficients of the quadratic trend contrast. For example, suppose we have $r = 17$ observations on the equally spaced levels

$$x_1 = 12, \quad x_2 = 18, \quad x_3 = 24, \quad x_4 = 30.$$

Then $z_x = x - \overline{x}_{..}$, so

$$z_{12} = -9, \quad z_{18} = -3, \quad z_{24} = 3, \quad z_{30} = 9.$$

These are 3 times the linear trend contrast coefficients listed in Appendix A.2. Also, $\sum_x z_x^2/v = 45$, so

$$z_{12}^{(2)} = 36, \quad z_{18}^{(2)} = -36, \quad z_{24}^{(2)} = -36, \quad z_{30}^{(2)} = 36,$$

which are 36 times the quadratic trend contrasts.

As in the simple linear regression case, one can likewise show that the least squares estimates $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ are constant multiples of the corresponding linear and quadratic trend contrast estimates $\hat{\tau}_3 - \hat{\tau}_1$ and $\hat{\tau}_1 - 2\hat{\tau}_2 + \hat{\tau}_3$ that would be used in the one-way analysis of variance model. Consequently, the sums of squares for testing no quadratic trend and no linear trend are the same, although again, the error mean square will differ.

8.7.3 Comments

We have illustrated via two examples the equivalence between the orthogonal trend contrasts in analysis of variance and orthogonal polynomials in regression analysis for the case of equispaced, equireplicated treatment levels. While both are convenient tools for data analysis, identification of orthogonal trend contrasts and orthogonal polynomials can be rather complicated for higher-order trends, unequally spaced levels, or unequal numbers of observations per level. Fortunately, analogous testing information can also be generated by fitting appropriate full and reduced models, as was discussed in Sect. 8.6.1. This is easily accomplished using computer regression software. Use of SAS and R software for such tests will be illustrated in Sects. 8.9 and 8.10.

8.8 A Real Experiment—Bean-Soaking Experiment

The bean-soaking experiment was run by Gordon Keeler in 1984 to study how long mung bean seeds ought to be soaked prior to planting in order to promote early growth of the bean sprouts. The experiment was run using a completely randomized design, and the experimenter used a one-way

analysis of variance model and methods of multiple comparisons to analyze the data. In Sect. 8.8.2, we present the one-way analysis of variance, and then in Sect. 8.8.3, we reanalyze the data using polynomial regression methods.

8.8.1 Checklist

The following checklist has been drawn from the experimenter's report.

(a) **Define the objectives of the experiment.**

The objective of the experiment is to determine whether the length of the soaking period affects the rate of growth of mung bean seed sprouts. The directions for planting merely advise soaking overnight, and no further details are given.

As indicated in Fig. 8.6, I expect to see no sprouting whatsoever for short soaking times, as the water does not have sufficient time to penetrate the bean coat and initiate sprouting. Then, as the soaking time is increased, I would expect to see a transition period of sprouting with higher rates of growth as water begins to penetrate the bean coat. Eventually, the maximum growth rate would be reached due to complete saturation of the bean. A possible decrease in growth rates could ensue from even longer soaking times due to bacterial infection and "drowning" the bean.

(b) **Identify all sources of variation**.

(i) Treatment factors and their levels.

There is just one treatment factor in this experiment, namely soaking time. A pilot experiment was run to obtain an indication of suitable times to be examined in the main experiment. The pilot experiment examined soaking times from 0.5 to 16 h. Many beans that had been soaked for less than 6 h failed to germinate, and at 16 h the saturation point had not yet been reached. Consequently, the five equally spaced soaking times of 6, 12, 18, 24 and 30 h will be selected as treatment factor levels for the experiment.

(ii) Experimental units.

The experimental units are the mung bean seeds selected at random from a large sack of approximately 10,000 beans.

(iii) Blocking factors, noise factors, and covariates.

Sources of variation that could affect growth rates include: individual bean differences; protozoan, bacterial, fungal, and viral parasitism; light; temperature; humidity; water quality.

Differences between beans will hopefully balance out in the random assignment to soaking times. Light, temperature, humidity, and water quality will be kept constant for all beans in the experiment. Thus, no blocking factors or covariates will be needed in the model.

Bacterial infection could differ from one treatment factor level to another due to soaking the beans in different baths. However, if the beans assigned to different treatment factor levels are soaked in the same bath, this introduces the possibility of a chemical signal from beans ready to germinate to the still dormant beans that sprouting conditions are prime. Consequently, separate baths will be used.

(c) **Choose a rule by which to assign experimental units to treatments**.

A completely randomized design will be used with an equal number of beans assigned to each soaking time.

Fig. 8.6 Anticipated
results from the
bean-soaking experiment
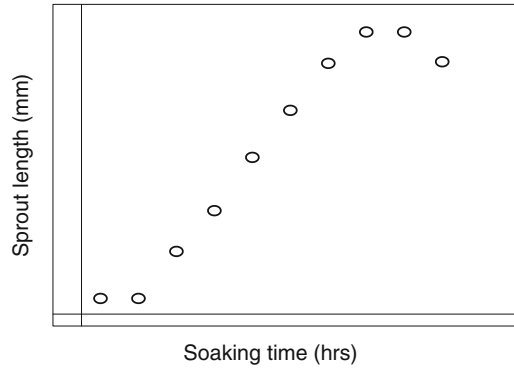
(d) **Specify the measurements to be made, the experimental procedure, and the anticipated difficulties**.

The soaking periods will be started at 6-h intervals, so that the beans are removed from the water at the same time. They will then be allowed to grow in the same environmental conditions for 48 h, when the lengths of the bean sprouts will be measured (in millimeters).

The main difficulty in running the experiment is in controlling all the factors that affect growth. The beans themselves will be randomly selected and randomly assigned to soaking times. Different soaking dishes for the different soaking times will be filled at the same time from the same source. On removal from the soaking dishes, the beans will be put in a growth chamber with no light but high humidity. During the pilot experiment, the beans were rinsed after 24 h to keep them from dehydrating. However, the procedure cannot be well controlled from treatment to treatment, and will not be done in the main experiment.

A further difficulty is that of accurately measuring the shoot length.

(e) **Run a pilot experiment**.

A pilot study was run and the rest of the checklist was completed. As indicated in step (b), the results were used to determine the soaking times to be included in the experiment.

(f) **Specify the model**.

The one-way analysis of variance model (3.3.1) will be used, and the assumptions will be checked after the data are collected.

(g) **Outline the analysis**.

Confidence intervals for the pairwise differences in the effects of soaking time on the 48-h shoot lengths will be calculated. Also, in view of the expected results, linear, quadratic and cubic trends in the shoot length will be examined. Tukey's method will be used for the pairwise comparisons with $\alpha_1 = 0.01$, and Bonferroni's method will be used for the three trend contrasts with overall level $\alpha_2 \leq 0.01$. The experimentwise error rate will then be at most 0.02.

Table 8.6 Length of shoots of beans after 48 h for the bean-soaking experiment

Soaking time (h)	r			Length (mm)				Average length	Sample variance
12	17	5	11	8	11	4	4	5.9412	7.0588
		8	3	6	4	7	3		
		5	4	6	9	3			
18	17	11	16	18	24	18	18	18.4118	12.6324
		21	14	21	19	17	24		
		14	20	16	20	22			
24	17	17	16	26	18	14	24	19.5294	15.6397
		18	14	24	26	21	21		
		22	19	14	19	19			
30	17	20	18	22	20	21	17	21.2941	8.5956
		16	23	25	19	21	20		
		27	25	22	23	23			

Fig. 8.7 Plot of sprout length y_{xt} against soaking time x for the bean-soaking experiment
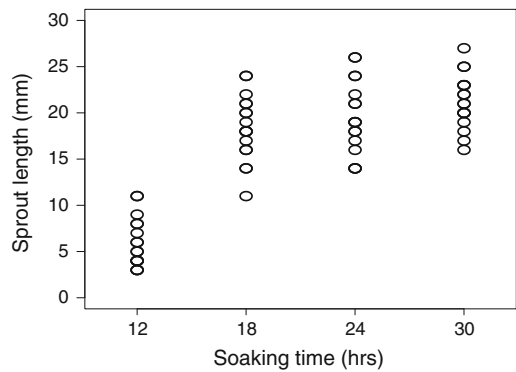
(h) **Calculate the number of observations that need to be taken**.

Using the results of the pilot experiment, a calculation showed that 17 observations should be taken on each treatment (see Example 4.5.1, p. 93).

(i) **Review the above decisions. Revise, if necessary**.

Since 17 observations could easily be taken for the soaking time, there was no need to revise the previous steps of the checklist.

The experiment was run, and the resulting data are shown in Table 8.6. The data for soaking time 6 h have been omitted from the table, since none of these beans germinated.

The data are plotted in Fig. 8.7 and show that the trend expected by the experimenter is approximately correct. For the soaking times included in the study, sprout length appears to increase with soaking time, with soaking times of 18, 24, and 30 h yielding similar results, but a soaking of time of only 12 h yielding consistently shorter sprouts.

8.8.2   One-Way Analysis of Variance and Multiple Comparisons

The experimenter used Tukey's method with a 99% simultaneous confidence level to compare the effects of soaking the beans for 12, 18, 24, or 30 h. The formula for Tukey's method for the one-way analysis of variance model was given in (4.4.28) as

$$\tau_i - \tau_s \in \left(\bar{y}_{i.} - \bar{y}_{s.} \pm w_T \sqrt{\left(\frac{2}{r}\right) msE} \right),$$

where $w_T = q_{v,n-v,\alpha}/\sqrt{2}$.

The treatment sample means are shown in Table 8.6. There are $r = 17$ observations on each of the $v = 4$ levels of the treatment factor. The formula for the sum of squares for error in the one-way analysis of variance model was given in (3.4.5), p. 39. Using the data in Table 8.6 we have

$$msE = ssE/(n - v) = 10.9816 \,.$$

From Table A.8, $q_{4,64,0.01} = 4.60$. Thus, in terms of the coded factor levels, the 99% simultaneous confidence intervals for pairwise comparisons are

$$
\begin{aligned}
\tau_4 - \tau_3 &\in (-1.93, \ \ 5.46)\,, &\tau_3 - \tau_2 &\in (-2.58, \ \ 4.81)\,, \\
\tau_4 - \tau_2 &\in (-0.81, \ \ 6.58)\,, &\tau_3 - \tau_1 &\in \ \ (9.89, 17.29)\,, \\
\tau_4 - \tau_1 &\in \ \ (11.66, 19.05)\,, &\tau_2 - \tau_1 &\in \ \ (8.77, 16.17)\,.
\end{aligned}
$$

From these, we can deduce that soaking times of 18, 24, and 30 h yield significantly longer sprouts on average after 48 h than does a soaking time of only 12 h. The three highest soaking times are not significantly different in their effects on the sprout lengths, although the plot (Fig. 8.7) suggests that the optimum soaking time might approach or even exceed 30 h.

The one-way analysis of variance for the data is given in Table 8.7 and includes the information for testing for linear, quadratic, and cubic trends. The coefficients for the trend contrasts, when there are $v = 4$ equally spaced levels and equal sample sizes, are listed in Table A.2. The linear contrast is $[-3, -1, \ 1, \ 3]$, and the hypothesis of no linear trend is $H_0^L : \{-3\tau_1 - \tau_2 + \tau_3 + 3\tau_4 = 0\}$. Obtaining the treatment sample means from Table 8.6, the estimate of the linear trend is

$$\sum_i c_i \bar{y}_i = -3\bar{y}_{1.} - \bar{y}_{2.} + \bar{y}_{3.} + 3\bar{y}_{4.} = 47.1765 \,,$$

with associated variance

$$\Sigma_i (c_i^2/r)\sigma^2 = (1/17)(9 + 1 + 1 + 9)\sigma^2 = (20/17)\sigma^2 \,.$$

The sum of squares is calculated from (4.3.14), p. 77; that is,

$$ssc = \left(\sum_i c_i \bar{y}_{i.} \right)^2 \Big/ \left(\sum_i c_i^2/17 \right).$$

So, the sum of squares for the linear trend is

Table 8.7 One-way ANOVA for the bean-soaking experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Soaking time	3	2501.29	833.76	75.92	0.0001
Linear trend	1	1891.78	1891.78	172.27	0.0001
Quadratic trend	1	487.12	487.12	44.36	0.0001
Cubic trend	1	122.40	122.40	11.15	0.0014
Error	64	702.82	10.98		
Total	67	3204.12			

$$ssc = (47.1765)^2/(20/17) = 1891.78.$$

The quadratic and cubic trends correspond to the contrasts $[\ 1,\ -1,\ -1,\ \ 1\ ]$ and $[-1,\ \ 3,\ -3,\ \ 1\ ]$, respectively, and their corresponding sums of squares are calculated in a similar way and are listed in Table 8.7. If we test the hypotheses that each of these three trends is zero with an overall significance level of $\alpha = 0.01$ using the Bonferroni method, then, using (4.4.24) on p. 84 for each trend, the null hypothesis that the trend is zero is rejected if $ssc/msE > F_{1,64,0.01/3}$. This critical value is not tabulated, but since $F_{1,64,0.0033} = t^2_{1,64,0.00166}$, it can be approximated using (4.4.22) as follows:

$$t_{1,64,0.00166} \approx 2.935 + (2.935^3 + 2.935)/(4 \times 64) = 3.0454,$$

so the critical value is $F_{1,64,0.0033} \approx 9.2747$. (Alternatively, the critical value could be obtained from a computer package using the "inverse cumulative distribution function" of the F-distribution.)

To test the null hypothesis H_0^L that the linear trend is zero against the alternative hypothesis $H_A^L : -3\tau_1 - \tau_2 + \tau_3 + 3\tau_4 \neq 0$ that the linear trend is nonzero, the decision rule is to

$$\text{reject } H_0 \text{ if } ssc/msE = 172.27 > F_{1,64,.0033} \approx 9.2747\,.$$

Thus, using a simultaneous significance level $\alpha = 0.01$ for the three trends, the linear trend is determined to be nonzero.

The corresponding test ratios for the quadratic and cubic trends are given in Table 8.7. There is sufficient evidence to conclude that the linear, quadratic, and cubic trends are all significantly different from zero. The probability that one or more of these hypotheses would be incorrectly rejected by this procedure is at most $\alpha = 0.01$.

8.8.3   Regression Analysis

In the previous subsection, the bean-soaking experiment was analyzed using the one-way analysis of variance and multiple comparison methods. In this subsection, we reanalyze the experiment using regression analysis. Since there are four levels of the treatment factor "soaking time," the highest-order polynomial regression model that can be (uniquely) fitted to the data is the cubic regression model, namely,

Table 8.8 Cubic regression ANOVA for the bean-soaking experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
β_3	1	122.40	122.40	11.15	0.0014
β_2, β_3	2	609.52	304.76	27.76	0.0001
Model	3	2501.29	833.76	75.92	0.0001
Error	64	702.82	10.98		
Total	67	3204.12			

$$Y_{xt} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon_{xt} \,,$$

$$\epsilon_{xt} \sim N(0, \sigma^2) \,,$$

$$\epsilon_{xt}\text{'s are mutually independent} \,,$$

$$x = 12, 18, 24, 30; \quad t = 1, \dots, 17.$$

Using the data given in Table 8.6, the fitted model can be obtained from a computer program (see Sects. 8.9 and 8.10) as

$$\hat{y}_x = -101.058824 + 15.475490x - 0.657680x^2 + 0.009259x^3 \,.$$

Table 8.8 contains the analysis of variance for the bean experiment data based on the cubic regression model. The cubic model provides the same fit as does the one-way analysis of variance model, since $p+1 = v = 4$. Thus, $\hat{y}_x = \overline{y}_{x.}$ for $x = 12, 18, 24, 30$, and the number of degrees of freedom, the sum of squares, and the mean square for the major sources of variation—the treatment factor ("Model"), error, and total—are the same in the regression analysis of variance as in the one-way analysis of variance. It is not possible to test for model lack of fit, since the postulated model is of order $p = 3 = v - 1$. We can, however, test to see whether a lower-order model would suffice.

We first test the null hypothesis $H_0^Q : \beta_3 = 0$, or equivalently, that the quadratic regression model $E[Y_{xt}] = \beta_0 + \beta_1 x + \beta_2 x^2$ would provide an adequate fit to the data. The result of the test is summarized in Table 8.8. The test ratio is 11.15 with a p-value of 0.0014. So, we reject H_0^Q and conclude that the cubic model is needed. Since the cubic regression model provides the same fit as the analysis of variance model, this test is identical to the test that the cubic trend contrast is zero in the one-way analysis of variance, shown in Table 8.7.

If $H_0^Q : \beta_3 = 0$ had not been rejected, then the next step would have been to have tested the null hypothesis $H_0^L : \beta_2 = \beta_3 = 0$, or equivalently, that the simple linear regression model is adequate. If neither $H_0^Q : \beta_3 = 0$ nor $H_0^L : \beta_2 = \beta_3 = 0$ had been rejected, the next step would have been to have tested $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

Based on the previous analysis, the cubic model is needed to provide an adequate fit to the data. Figure 8.8 illustrates the cubic model fitted to the data. We may now see the dangers of using a model to predict the value of the response beyond the range of observed x values. The cubic model predicts that mean sprout length will increase rapidly as soaking time is increased beyond 30 h! Clearly, this model is extremely unlikely to be reliable for extrapolation beyond 30 h.

Recall that Tukey's method of multiple comparisons did not yield any significant differences in mean response between the soaking times of 18, 24, and 30 h. Yet the plot of the data in Fig. 8.8 suggests that a trend over these levels might well exist. There is a lot of variability inherent in the data that prevents significant differences between the soaking times from being detected. Nevertheless, a

Fig. 8.8 Plot of data and fitted cubic polynomial regression model for the bean-soaking experiment
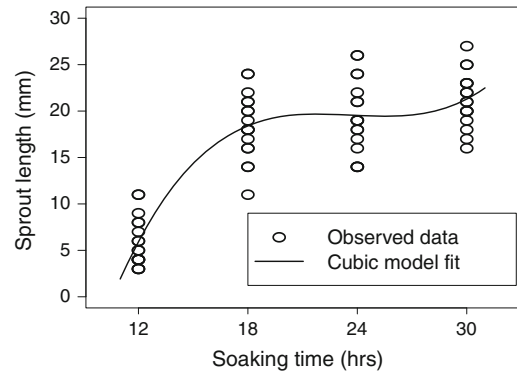
followup experiment examining soaking times from 18 to, say, 48 h might provide the information needed to determine the best range of soaking times.

8.9   Using SAS Software

Polynomial regression models can be fitted using the SAS regression procedure `PROC REG`. The procedure provides least squares estimates of the regression parameters. Predicted (fitted) values and residuals can be saved to an output data set, as can 95% confidence limits for mean response, 95% prediction limits for new observations for given treatment levels x, and corresponding standard errors.

A sample SAS program to analyze the data from the bean-soaking experiment of Sect. 8.8 is shown in Table 8.9 . In the first `DATA` statement, the variables x^2 and x^3 are created for the cubic regression model. `PROC REG` is used to fit the cubic regression model, and the output is shown in Fig. 8.9.

An analysis of variance table is automatically generated and includes information needed for testing the hypothesis that the treatment factor "soaking time" has no predictive value for mean growth length, namely, $H_0 : \{\beta_1 = \beta_2 = \beta_3 = 0\}$. The information for this test is listed with source of variation "Model". We see that the p-value is less than 0.0001, so H_0 would be rejected.

Below the analysis of variance table, parameter estimates for the fitted model are given. Using these, we have the fitted cubic regression model

$$\hat{y}_x = -101.05882 + 15.47549x - 0.65768x^2 + 0.00926x^3 .$$

The standard error of each estimate is also provided, together with the information for conducting a t-test of each individual hypothesis $H_0 : \{\beta_i = 0\}$, $i = 1, 2, 3$.

Inclusion of the option `SS1` in the `MODEL` statement of `PROC REG` causes printing of the Type I (sequential) sums of squares in the output. Each Type I sum of squares is the variation explained by entering the corresponding variable into the model, given that the previously listed variables are already in the model. For example, the Type I sum of squares for X is $ssE_0 - ssE_1$, where ssE_0 is the error sum of squares for the model with $E[Y_{xt}] = \beta_0$, and ssE_1 is the error sum of squares for the simple linear regression model $E[Y_{xt}] = \beta_0 + \beta_1 x$; that is,

$$ss(\beta_1|\beta_0) = ssE_0 - ssE_1 = 1891.77647 .$$

Likewise, the Type I sum of squares for X2 is the difference in error sums of squares for the linear and quadratic regression models; that is,

Fig. 8.9 Output generated by PROC REG
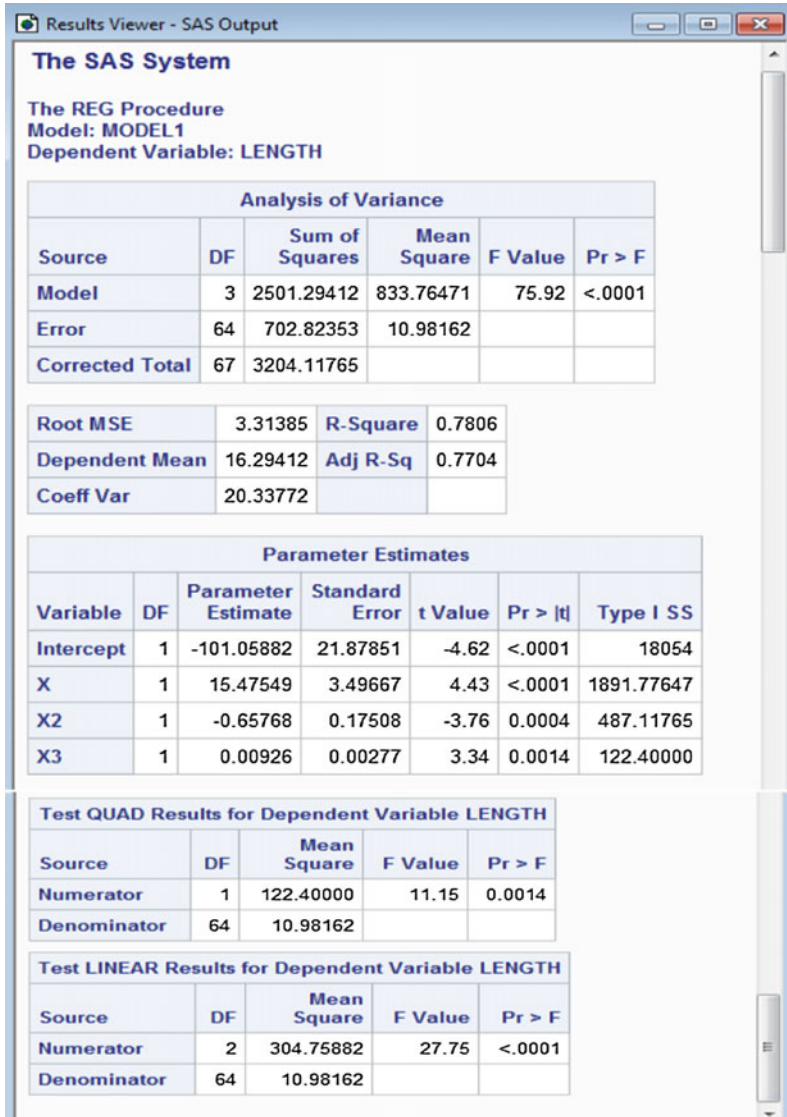
$$ss(\beta_2|\beta_0, \beta_1) = ssE_1 - ssE_2 = 487.11765 \,,$$

and for X3, the Type I sum of squares is the difference in error sums of squares for the quadratic and cubic regression models; that is,

$$ss(\beta_3|\beta_0, \beta_1, \beta_2) = ssE_2 - ssE = 122.40000 \,,$$

where we have written ssE for the error sum of squares for the full cubic model (rather than ssE_3). Thus, the ratio used to test the null hypothesis $H_0^Q : \{\beta_3 = 0\}$ versus $H_A^Q : \{\beta_3 \neq 0\}$ is

$$ss(\beta_3)/msE = ss(\beta_3|\beta_0, \beta_1, \beta_2)/msE = 122.4/10.98162 = 11.1459 \,.$$

Table 8.9 SAS program for analysis of the bean-soaking experiment

```
DATA BEAN;
  INPUT X LENGTH;
  X2 = X**2; X3 = X**3;
  LINES;
  12    5
  12   11
  12    8
   :    :
  30   23
  30   23
;
* create extra x-values for plotting the fitted curve;
DATA TOPLOT;
  DO X = 8 TO 34;  X2 = X**2; X3 = X**3;
    LENGTH = .; * "." denotes a missing value;
    OUTPUT;
  END; * X loop;
* concatenate data sets BEAN and TOPLOT;
DATA; SET BEAN TOPLOT;
* do the analysis;
PROC REG;  MODEL LENGTH = X X2 X3 / SS1;
  QUAD:   TEST X3 = 0; * test adequacy of quadratic model;
  LINEAR: TEST X2 = 0, X3 = 0; * test adequacy of linear model;
  OUTPUT PREDICTED = LHAT RESIDUAL = E
     L95M = L95M U95M = U95M STDP = STDM
     L95 = L95I U95 = U95I STDI = STDI;
* plot the data and fitted model, overlayed on one plot;
PROC SGPLOT;
  SCATTER Y = LENGTH X = X / LEGENDLABEL = 'Observed data'
   MARKEREATTRS = (SIZE = 0.25cm COLOR = BLACK);
  SCATTER Y = LHAT X = X / LEGENDLABEL = 'Cubic model fit'
   MARKEREATTRS = (SYMBOL = SQUARE SIZE = 0.25cm COLOR = BLACK);
  YAXIS LABEL = "Sprout length (mm)" VALUES = (-20 TO 30 by 5);
  XAXIS LABEL = "Soaking time (hrs)" VALUES = (8 TO 36 by 4);
* 95% confidence intervals and standard errors for mean response;
PROC PRINT; VAR X L95M LHAT U95M STDM;
* 95% prediction intervals and standard errors for new observations;
PROC PRINT; VAR X L95I LHAT U95I STDI;
* generate residual plots;
PROC RANK NORMAL = BLOM; VAR E; RANKS NSCORE;
PROC SGPLOT; SCATTER Y = E X = X;
PROC SGPLOT; SCATTER Y = E X = LHAT;
PROC SGPLOT; SCATTER Y = E X = NSCORE;
```

The output of the TEST statement labeled QUAD provides the same information, as well as the p-value 0.0014. The null hypothesis H_0^Q is thus rejected, so the quadratic model is not adequate—the cubic model is needed. Hence, there is no reason to test further reduced models, but the information for such tests will be discussed for illustrative purposes.

To test $H_0^L : \beta_2 = \beta_3 = 0$, the full model is the cubic model and the reduced model is the linear model, so the numerator sum of squares of the test statistic is

$$ss(\beta_2, \beta_3) = ssE_1 - ssE \; = \; ss(\beta_2|\beta_0, \beta_1) + ss(\beta_3|\beta_0, \beta_1, \beta_2)$$
$$= 487.117647 + 122.400000 \; = \; 609.517647 \,,$$

and the decision rule for testing H_0^L against the alternative hypothesis H_A^L that the cubic model is needed is

$$\text{reject } H_0^L \text{ if } ms(\beta_2, \beta_3)/msE > F_{2,64,\alpha} \,,$$

where

$$ms(\beta_2, \beta_3) = ss(\beta_2, \beta_3)/2 \,.$$

The information for this test of adequacy of the linear model is also generated by the TEST statement labeled LINEAR.

The OUTPUT statement in PROC REG saves into an output data set the upper and lower 95% confidence limits for mean response and the corresponding standard error under the variable names L95M, U95M and STDM. This is done for each x-value in the input data set for which all regressors are available. Similarly, the upper and lower 95% prediction limits for a new individual observation and the corresponding standard error are saved under the variable names L95I, U95I and STDI. These could be printed or plotted, though we do not do so here.

The plot produced by PROC SGPLOT is not shown but is similar to the plot in Fig. 8.8. Overlaid on the same axes are plots of the raw data and the fitted cubic polynomial regression curve. A trick was used to generate data to plot the fitted curve. Actual x values range from 12 to 30. In the DATA TOPLOT step in Table 8.9, additional observations were created in the data set corresponding to the integer x values ranging from 8 to 34 but with missing values for the dependent variable length. While observations with missing length values cannot be used to fit the model, the regression procedure does compute the corresponding predicted values LHAT. The OUTPUT statement includes these fitted values in the newly created output data set, so they can be plotted to show the fitted model.

In this example, it is not possible to test for lack of fit of the cubic model, since data were collected at only four x-levels. If we had been fitting a quadratic model, then a lack-of-fit test would have been possible. An easy way to generate the relevant output using the SAS software is as follows. In line 4 of the program, add a classification variable A, using the statement "A = X;". Then insert a PROC GLM procedure before PROC REG as follows.

```
PROC GLM;
  CLASS A;
  MODEL LENGTH = X X2 A;
```

Then the Type I sum of squares for A is the appropriate numerator *ssLOF* for the test ratio.

Statements for generation of residual plots for checking the error assumptions are included in the sample SAS program in Table 8.9, but the output is not shown here.

8.10 Using R Software

Polynomial regression models can be fitted using the R function lm that fits linear models. The function provides least squares estimates of the regression parameters. Predicted (fitted) values and residuals are available, as are 95% confidence limits for mean response, 95% prediction limits for new observations for given treatment levels x, and corresponding standard errors.

A sample R program to analyze the data from the bean-soaking experiment of Sect. 8.8 is shown in Table 8.10. In the first block of code, the data are read from file into the data set bean.data.

Subsequently, though the results are not shown here, the `head` command would display the first six rows of data, showing for example that the data set contains the two variables `x` and `Length`, then the dimension command would reveal that the data set contains 68 observations, and a scatterplot of the data would be generated.

In the second block of code, the linear model function `lm` is used to fit the cubic, quadratic, and linear regression models, saving the respective results as `model3`, `model2` and `model1`, and related commands are used to generate the output shown in Tables 8.11 and 8.12. Since the data set contains the soaking time, x, the syntax `I(x^2)` allows inclusion of the quadratic term x^2 as a predictor variable in a model without creating a corresponding variable in the data set, and likewise `I(x^3)` for the cubic term x^3. The command `summary(model3)` displays the parameter least squares estimates shown in the middle of Table 8.11. From these, we have the fitted cubic regression model

$$\hat{y}_x = -101.05882 + 15.47549x - 0.65768x^2 + 0.00926x^3 \,.$$

The standard error of each estimate is also provided, together with the information for conducting a t-test of each individual hypothesis $H_0 : \{\beta_i = 0\}$, $i = 1, 2, 3$.

The `summary` command also generates the analysis of variance F-test of the hypothesis that the treatment factor "soaking time" has no predictive value for mean growth length, namely, $H_0 : \{\beta_1 = \beta_2 = \beta_3 = 0\}$. The information for this test is listed after "`F-statistic`". We see that the p-value is very small, only 2×10^{-16}, so H_0 would be rejected.

Having saved the results of the cubic fit as `model3`, the statement `anova(model3)` causes display of the Type I (sequential) sums of squares, provided in an analysis of variance table in the bottom of Table 8.11. Each Type I sum of squares is the variation explained by entering the corresponding variable into the model, given that the previously listed variables are already in the model. For example, the Type I sum of squares for `x` is $ssE_0 - ssE_1$, where ssE_0 is the error sum of squares for the model with $E[Y_{xt}] = \beta_0$, and ssE_1 is the error sum of squares for the simple linear regression model $E[Y_{xt}] = \beta_0 + \beta_1 x$; that is,

$$ss(\beta_1|\beta_0) = ssE_0 - ssE_1 \approx 1892 \,.$$

Likewise, the Type I sum of squares for `x^2` is the difference in error sums of squares for the linear and quadratic regression models; that is,

$$ss(\beta_2|\beta_0, \beta_1) = ssE_1 - ssE_2 \approx 487 \,,$$

and for `x^3`, the Type I sum of squares is the difference in error sums of squares for the quadratic and cubic regression models; that is,

$$ss(\beta_3|\beta_0, \beta_1, \beta_2) = ssE_2 - ssE \approx 122 \,,$$

where we have written ssE for the error sum of squares for the full cubic model (rather than ssE_3). Thus, the ratio used to test the null hypothesis $H_0^Q : \{\beta_3 = 0\}$ versus $H_A^Q : \{\beta_3 \neq 0\}$ is

$$ss(\beta_3)/msE = ss(\beta_3|\beta_0, \beta_1, \beta_2)/msE \approx 122/11 \approx 11.2 \,,$$

with corresponding p-value 0.0014. The null hypothesis H_0^Q is thus rejected, so the quadratic model is not adequate—the cubic model is needed. The same information is generated by the statement `anova(model2, model3)`, which compares the reduced quadratic and full cubic models, with

Table 8.10 R program for analysis of the bean-soaking experiment

```
bean.data = read.table("data/bean.txt", header=T)
head(bean.data); dim(bean.data); plot(Length ~ x, data=bean.data)

# Fit regression models and generate ANOVA info
model3 = lm(Length ~ x + I(x^2) + I(x^3), data=bean.data) # Fit cubic model
summary(model3) # Display least squares estimates, overall F test
anova(model3) # Display type 1 SS
# Would a lower order model suffice?
model2 = lm(Length ~ x + I(x^2), data=bean.data) # Fit quadratic model
model1 = lm(Length ~ x, data=bean.data) # Fit simple linear reg model
anova(model2, model3) # Can cubic term be dropped?
anova(model1, model3) # Can both cubic and quadratic terms be dropped?

# Compute predicted values, CIs, PIs, and std errors for x=8, 8.01, ..., 34
# Set up a grid of x's for prediction: x=8, 8.01, 8.02, ..., 34
xPred = data.frame(x=seq(8, 34, 0.01))
# Calculate fitted values, 95% CIs for mean response, se.fit
preds = predict(model3, xPred, se.fit=T, interval=c("confidence"))
# Calculate 95% PIs for new observations
preds2 = predict(model3, xPred, interval = c("prediction"))
# preds; preds2 # (Reader: display preds and preds2 to see contents)
se.fit = preds$se.fit # to remove "preds$" from column header name
# Compute standard error for prediction
rmse = preds$residual.scale # used to compute se.pred
se.pred = sqrt(preds$se.fit^2 + rmse^2)
# Consolidate results for display
stats = cbind(xPred, preds$fit, se.fit, preds2[,2:3], se.pred)
head(stats) # display first six rows of results

# Plot data (length vs x), plus fitted model for x=8:34
plot(Length ~ x, xlim = c(8, 34), ylim = c(-10, 30), data=bean.data)
lines(xPred$x, preds$fit[, 1])

# Some plots to check model assumptions
bean.data$e = residuals(model3); # Obtain residuals
bean.data$pred = fitted(model3) # Obtain predicted values
# Plot residuals vs x
plot(e ~ x, ylab = "Residuals", las=1, xaxt="n", data=bean.data)
axis(1, at = c(12,18,24,30)); abline(h=0)
# Plot residuals vs predicted values
plot(e ~ pred, xlim=c(5,25), las=1, xaxt="n", data=bean.data,
     xlab="Predicted Values", ylab = "Residuals")
axis(1, at=seq(5,25,5)); abline(h=0)
# Normal probability plot of residuals
qqnorm(model3$res, ylim=c(-10,10), xlim=c(-4,4)); abline(h=0, v=0)
```

output shown in the top of Table 8.12. Since the cubic model is needed, there is no reason to test further reduced models, but the information for such tests will be discussed for illustrative purposes.

To test $H_0^L : \beta_2 = \beta_3 = 0$, the full model is the cubic model and the reduced model is the simple linear regression model, so the numerator sum of squares of the test statistic computed from the Type I sums of squares is

Table 8.11 Output generated for the cubic model

```
> # Fit regression models and generate ANOVA info
> model3 = lm(Length ~ x + I(x^2) + I(x^3), data=bean.data) # Fit cubic model
> summary(model3) # Display least squares estimates, overall F test

Call:
lm(formula = Length ~ x + I(x^2) + I(x^3), data = bean.data)

Residuals:
   Min    1Q Median    3Q    Max
-7.412 -2.029 -0.412  2.059  6.471

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -101.05882   21.87851   -4.62 0.000019
x             15.47549    3.49667    4.43 0.000038
I(x^2)        -0.65768    0.17508   -3.76  0.00037
I(x^3)         0.00926    0.00277    3.34  0.00141

Residual standard error: 3.31 on 64 degrees of freedom
Multiple R-squared:  0.781, Adjusted R-squared:  0.77
F-statistic: 75.9 on 3 and 64 DF,  p-value: <2e-16

> anova(model3) # Display type 1 SS
Analysis of Variance Table

Response: Length
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1   1892    1892   172.3 < 2e-16
I(x^2)     1    487     487    44.4 7.3e-09
I(x^3)     1    122     122    11.2  0.0014
Residuals 64    703      11
```

$$ss(\beta_2, \beta_3) = ssE_1 - ssE = ss(\beta_2|\beta_0, \beta_1) + ss(\beta_3|\beta_0, \beta_1, \beta_2)$$
$$\approx 487 + 122 = 609,$$

and the decision rule for testing H_0^L against the alternative hypothesis H_A^L that the cubic model is needed is

$$\text{reject } H_0^L \text{ if } ms(\beta_2, \beta_3)/msE > F_{2,64,\alpha},$$

where

$$ms(\beta_2, \beta_3) = ss(\beta_2, \beta_3)/2.$$

The information for this test of adequacy of the simple linear regression model is also generated by the statement anova(model1, model3), with results shown in the bottom of Table 8.12. Here, the numerator sum of squares is rounded to 610, yielding $F = 27.8$ and $p = 2.1 \times 10^{-09}$.

The third block of code in Table 8.10 saves a grid of x values from 8 to 34 in step of 0.01 in a data set xPred, in order to compute confidence and prediction intervals at these x values. The predict function is called twice, each time using the results of the cubic fit saved as model3. For each x value in the grid, the first call of predict computes the predicted length, the standard error for predicting mean length, and the 95% confidence interval for mean response, saving these results asvadjust preds. By

Table 8.12 Output generated for the cubic model (continued)

```
> # Would a lower order model suffice?
> model2 = lm(Length ~ x + I(x^2), data=bean.data) # Fit quadratic model
> model1 = lm(Length ~ x, data=bean.data) # Fit simple linear reg model
> anova(model2, model3) # Can cubic term be dropped?
Analysis of Variance Table

Model 1: Length ~ x + I(x^2)
Model 2: Length ~ x + I(x^2) + I(x^3)
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1     65 825
2     64 703  1       122 11.2 0.0014
> anova(model1, model3) # Can both cubic and quadratic terms be dropped?
Analysis of Variance Table

Model 1: Length ~ x
Model 2: Length ~ x + I(x^2) + I(x^3)
  Res.Df  RSS Df Sum of Sq    F  Pr(>F)
1     66 1312
2     64  703  2       610 27.8 2.1e-09
```

displaying `preds`, one would see that the predicted values and confidence limits are saved as the three columns of the object `preds$fit`, the standard error is saved as the lone column of `preds$se.fit`, and the root mean squared error is saved as a scalar as `preds$residual.scale`. Due to the `"prediction"` option, the second call of `predict` computes the predicted length and the 95% prediction interval for a new observation for each $x = 8, \ldots, 34$, saving these results as the three columns of `preds2`. The standard error for prediction is not provided directly, but is subsequently computed for each x from the standard error for estimation of mean length for the given x value and from the common root mean squared error value, both available from `preds`. The column bind command `cbind` is used to combine the desired information into the columns of the object `stats`. These could be printed or plotted, though we do not do so here.

The output of the `plot` function in the fourth block of code is not shown here, but is similar to the plot show in Fig. 8.8. The `plot` command causes the raw data to be plotted. Then the `lines` subcommand augments the plot with the line corresponding to the predicted values at the grid points $x = 8, 8.01, 8.02, \ldots, 34$, giving a sense of the fitted cubic polynomial regression curve.

In this example, it is not possible to test for lack of fit of the cubic model, since data were collected at only four x-levels. If we had been fitting a quadratic model, then a lack-of-fit test would have been possible. An easy way to generate the relevant output using the R software is as follows. Anyplace after saving the fitted quadratic model as `model2` in the second block of code, add the following code.

```
bean.data$fA = factor(bean.data$x)
modelA = lm(Length ~ fA, data=bean.data)
anova(model2, modelA))
```

The first command adds a factor variable `fA` to the data set, the second fits the one-way model with a different mean for each x value (i.e. for each level of `fA`), and the third generates the F-test for lack of fit by comparing the reduced quadratic model to the full one-way model, which is the fullest model one can fit here.

Statements for generation of residual plots for checking the error assumptions are included in the sample R program in Table 8.10, but the output is not shown here.

Table 8.13 Data for the bicycle experiment

Treatment (mph)		Crank rates y_{xt}	
x		y_{xt}	
5	15	19	22
10	32	34	27
15	44	47	44
20	59	61	61
25	75	73	75

Exercises

1. For the simple linear regression model

$$E[Y_{xt}] = \beta_0 + \beta_1 x \,,$$

the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the parameters β_0 and β_1 are given in (8.5.6), p. 257. Show that their variances are

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}_{..}^2}{ss_{xx}} \right) \quad \text{and} \quad \text{Var}(\hat{\beta}_1) = \sigma^2 \left(\frac{1}{ss_{xx}} \right) \,,$$

where $ss_{xx} = \sum_x r_x (x - \overline{x}_{..})^2$, as given in (8.5.7).

2. **Bicycle experiment, continued**

   The bicycle experiment was run to compare the crank rates required to keep a bicycle at certain speeds, when the bicycle (a Cannondale SR400) was in twelfth gear on flat ground. The speeds chosen were $x = 5$, 10, 15, 20, and 25 mph. The data are given in Table 8.13. (See also Exercise 6 of Chap. 5.)

   (a) Fit the simple linear regression model to the data, and use residual plots to check the assumptions of the simple linear regression model.
   (b) If a transformation of the data is needed, choose a transformation, refit the simple linear regression model, and check for lack of fit.
   (c) Using your results from parts (a) and (b), select a model for the data. Use this model to obtain an estimate for the mean crank rate needed to maintain a speed of 18 mph in twelfth gear on level ground.
   (d) Calculate a 95% confidence interval for the mean crank rate needed to maintain a speed of 18 mph in twelfth gear on level ground.
   (e) Find the 95% confidence band for the regression line. Draw a scatter plot of the data and superimpose the regression line and the confidence band on the plot.
   (f) Would you be happy to use your model to estimate the mean crank rate needed to maintain a speed of 35 mph in twelfth gear on level ground. Why or why not?

Table 8.14 Systolic blood pressure measurements—(order of collection in parentheses)

Jogging time in seconds (order of collection)					
10	20	25	30	40	50
120 (1)	125 (2)	127 (10)	128 (3)	137 (5)	143 (6)
118 (9)	126 (4)		131 (7)		
	123 (8)				

Table 8.15 Data for the trout experiment

x	Hemoglobin (grams per 100 ml)									
00	6.7	7.8	5.5	8.4	7.0	7.8	8.6	7.4	5.8	7.0
05	9.9	8.4	10.4	9.3	10.7	11.9	7.1	6.4	8.6	10.6
10	10.4	8.1	10.6	8.7	10.7	9.1	8.8	8.1	7.8	8.0
15	9.3	9.3	7.2	7.8	9.3	10.2	8.7	8.6	9.3	7.2

Source Gutsell, J.S. (1951). Copyright© 1951 International Biometric Society. Reprinted with permission

3. **Systolic blood pressure experiment**

A pilot experiment was run by John Spitak in 1987 to investigate the effect of jogging on systolic blood pressure. Only one subject was used in the pilot experiment, and a main experiment involving a random sample of subjects from a population of interest would need to be run in order to draw more general conclusions. The subject jogged in place for a specified number of seconds and then his systolic blood pressure was measured. The subject rested for at least 5 min, and then the next observation was taken.

The data and their order of observation are given in Table 8.14.

(a) Fit a simple linear regression model to the data and test for model lack of fit.
(b) Use residual plots to check the assumptions of the simple linear regression model.
(c) Give a 95% confidence interval for the slope of the regression line.
(d) Using the confidence interval in part (c), test at significance level $\alpha = 0.05$ whether the linear term is needed in the model.
(e) Repeat the test in part (d) but using the formula for the orthogonal polynomial linear trend coefficients for unequally spaced levels and unequal sample sizes given in Sect. 4.2.4. Do these two tests give the same information?
(f) Estimate the mean systolic blood pressure of the subject after jogging in place for 35 sec and calculate a 99% confidence interval.
(g) The current experiment was only a pilot experiment. Write out a checklist for the main experiment.

4. **Trout experiment, continued**

The data in Table 8.15 show the measurements of hemoglobin (grams per 100 ml) in the blood of brown trout. (The same data were used in Exercise 15 of Chap. 3.) The trout were placed at random in four different troughs. The fish food added to the troughs contained, respectively, $x = 0, 5, 10,$ and 15 grams of sulfamerazine per 100 pounds of fish. The measurements were made on ten randomly selected fish from each trough after 35 days.

(a) Fit a quadratic regression model to the data.
(b) Test the quadratic model for lack of fit.

(c) Use residual plots to check the assumptions of the quadratic model.

(d) Test whether the quadratic term is needed in the model.

(e) Use the fitted quadratic model to estimate the number of grams of sulfamerazine per 100 pounds of fish to maximize the mean amount of hemoglobin in the blood of the brown trout.

5. **Bean-soaking experiment, continued**

Use residual plots to check the assumptions of the cubic regression model for the data of the bean-soaking experiment. (The data are in Table 8.6, p. 269).

6. **Bean-soaking experiment, continued**

Suppose the experimenter in the bean-soaking experiment of Sect. 8.8 had presumed that the quadratic regression model would be adequate for soaking times ranging from 12 to 30 h.

(a) Figure 8.8, p. 273, shows the fitted response curve and the standardized residuals each plotted against soaking time. Based on these plots, discuss model adequacy.

(b) Test the quadratic model for lack of fit.

7. **Orthogonal polynomials**

Consider an experiment in which an equal number of observations are collected for each of the treatment factor levels $x = 10, 20, 30, 40, 50$.

(a) Compute the corresponding values z_x for the linear orthogonal polynomial, and determine the rescaling factor by which the z_x differ from the coefficients of the linear trend contrast.

(b) Compute the values $z_x^{(2)}$ for the quadratic orthogonal polynomial, and determine the rescaling factor by which the $z_x^{(2)}$ differ from the coefficients of the quadratic trend contrast.

(c) Use the data of Table 8.1 and the orthogonal polynomial coefficients to test that the quadratic and linear trends are zero.

(d) Using the data of Table 8.1 and a statistical computing package, fit a quadratic model to the original values. Test the hypotheses

$$H_0^L : \{\beta_2 = 0\} \quad \text{and} \quad H_0 : \{\beta_1 = \beta_2 = 0\}$$

against their respective two-sided alternative hypotheses. Compare the results of these tests with those in (c).

8. **Orthogonal polynomials**

Consider use of the quadratic orthogonal polynomial regression model (8.7.20), p. 265, for the data at levels 18, 24, and 30 of the bean-soaking experiment—the data are in Table 8.6, p. 269.

(a) Compute the least squares estimates of the parameters.

(b) Why is it not possible to test for lack of fit of the quadratic model?

(c) Give an analysis of variance table and test the hypothesis that a linear model would provide an adequate representation of the data.

9. **Heart–lung pump experiment, continued**

   In Example 8.5.1, p. 259, we fitted a linear regression model to the data of the heart–lung pump experiment. We rejected the null hypothesis that the slope of the line is zero.

   (a) Show that the numerator sum of squares for testing $H_0 : \{\beta_1 = 0\}$ against the alternative hypothesis $H_A : \{\beta_1 \neq 0\}$ is the same as the sum of squares *ssc* that would be obtained for testing that the linear trend is zero in the analysis of variance model (the relevant calculations were done in Example 4.2.3, p. 73).
   (b) Obtain a 95% confidence band for the regression line.
   (c) Calculate a 99% prediction interval for the fluid flow rate at 100 revolutions per minute.
   (d) Estimate the intercept β_0. This is not zero, which suggests that the fluid flow rate is not zero at 0 rpm. Since this should not be the case, explain what is happening.

Analysis of Covariance

<div style="text-align: right">**9**</div>

9.1 Introduction

In Chaps. 3–7, we used completely randomized designs and analysis of variance to compare the effects of one or more treatment factors on a response variable. If nuisance factors are expected to be a major source of variation, they should be taken into account in the design and analysis of the experiment. If the values of the nuisance factors can be measured in advance of the experiment or controlled during the experiment, then they can be taken into account at the design stage using blocking factors, as discussed in Chap. 10. Analysis of covariance, which is the topic of this chapter, is a means of adjusting the analysis for nuisance factors that cannot be controlled and that sometimes cannot be measured until the experiment is conducted. The method is applicable if the nuisance factors are related to the response variable but are themselves unaffected by the treatment factors.

For example, suppose an investigator wants to compare the effects of several diets on the weights of month-old piglets. The response (weight at the end of the experimental period) is likely to be related to the weight at the beginning of the experimental period, and these weights will typically be somewhat variable. To control or adjust for this prior weight variability, one possibility is to use a block design, dividing the piglets into groups (or blocks) of comparable weight, then comparing the effects of diets within blocks. A second possibility is to use a completely randomized design with response being the weight gain over the experimental period. This loses information, however, since heavier piglets may experience higher weight gain than lighter piglets, or vice versa. It is preferable to include the prior weight in the model as a variable, called a *covariate*, that helps to explain the final weight.

The model for a completely randomized design includes the effects of the treatment factors of interest, together with the effects of any nuisance factors (covariates). *Analysis of covariance* is the comparison of treatment effects, adjusting for one or more covariates. Standard analysis of covariance models and assumptions are discussed in Sect. 9.2. Least squares estimates are derived in Sect. 9.3. Sections 9.4 and 9.5 cover analysis of covariance tests and confidence interval methods for the comparison of treatment effects. Analysis using software is illustrated using SAS and R software in Sects. 9.6 and 9.7, respectively.

9.2 Models

Consider an experiment conducted as a completely randomized design to compare the effects of the levels of v treatments on a response variable Y. Suppose that the response is also affected by a nuisance

Fig. 9.1 Linear and quadratic parallel response curves

(a) Linear response curves **(b)** Quadratic response curves
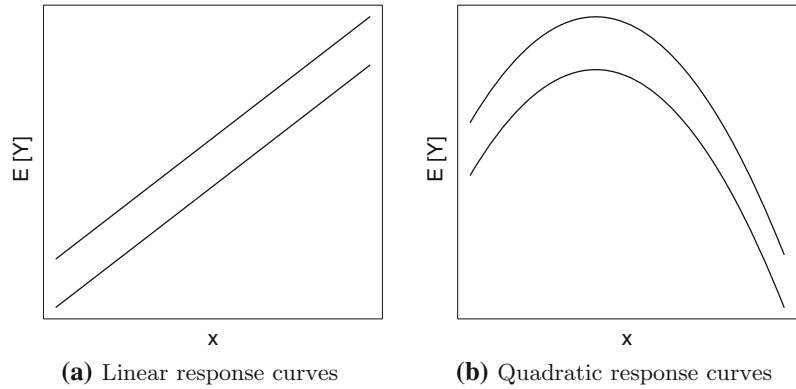
factor (covariate) whose value x can be measured during or prior to the experiment. Furthermore, suppose that there is a linear relationship between $E[Y]$ and x, with the same slope for each treatment. Then, if we plot $E[Y]$ versus x for each treatment separately, we would see parallel lines, as illustrated for two treatments in Fig. 9.1a. A comparison of the effects of the two treatments can be done by comparison of mean response at any value of x. The model that allows this type of analysis is the analysis of covariance model:

$$Y_{it} = \mu + \tau_i + \beta x_{it} + \epsilon_{it} \,, \tag{9.2.1}$$
$$\epsilon_{it} \sim N(0, \sigma^2) \,,$$
$$\epsilon_{it}{'}\text{s are mutually independent} \,,$$
$$t = 1, 2, \ldots, r_i; \quad i = 1, \ldots, v \,.$$

In this model, the effect of the ith treatment is modeled as τ_i, as usual. If there is more than one treatment factor, then τ_i represents the effect of the ith treatment combination and could be replaced by main-effect and interaction parameters. The value of the covariate on the tth time that treatment i is observed is written as x_{it}, and the linear relationship between the response and the covariate is modeled as βx_{it} as in a regression model. It is important for the analysis that follows that the value x_{it} of the covariate not be affected by the treatment—otherwise, comparison of treatment effects at a common x-value would not be meaningful.

A common alternative form of the analysis of covariance model is

$$Y_{it} = \mu^* + \tau_i + \beta(x_{it} - \overline{x}_{..}) + \epsilon_{it} \,, \tag{9.2.2}$$

in which the covariate values have been "centered." The two models are equivalent for comparison of treatment effects. The slope parameter β has the same interpretation in both models. In model (9.2.2), $\mu + \tau_i$ denotes the mean response when $x_{it} = \overline{x}_{..}$, whereas in model (9.2.1), $\mu^* + \tau_i$ denotes the mean response when $x_{it} = 0$, with the parameter relationship $\mu^* = \mu - \beta \overline{x}_{..}$. Model (9.2.2) is often used to reduce computational problems and is a little easier to work with in obtaining least squares estimates.

9.2.1   Checking Model Assumptions and Equality of Slopes

In addition to the usual assumptions on the error variables, the analysis of covariance model (9.2.2) assumes a linear relationship between the covariate and the mean response, with the same slope for each treatment, as illustrated in Fig. 9.1a. It is appropriate to start by checking for model lack of fit.

Lack of fit can be investigated by plotting the residuals versus the covariate for each treatment on the same scale. If the plot looks nonlinear for any treatment, then a linear relationship between the response and covariate may not be adequate. If each plot does look linear, one can assess whether the slopes are comparable. A formal test of equality of slopes can be conducted by comparing the fit of the analysis of covariance model (9.2.2) with the fit of the corresponding model that does not require equal slopes, for which

$$Y_{it} = \mu + \tau_i + \beta_i(x_{it} - \overline{x}_{..}) + \epsilon_{it} \ . \tag{9.2.3}$$

If there is no significant lack of fit of the model, then plots of the residuals versus run order, predicted values, and normal scores can be used as in Chap. 5 to assess the assumptions of independence, equal variances, and normality of the random error terms.

9.2.2   Model Extensions

The analysis of covariance model (9.2.1) can be generalized in various ways that we will mention here but not consider further.

If the effect of the covariate is not linear, then βx can be replaced with a higher-order polynomial function $\beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$ to adequately model the common shape of the response curves for each treatment, analogous to the polynomial response curve models of Chap. 8. For example, parallel quadratic response curves for two treatments are shown in Fig. 9.1b.

If there is more than one covariate, the single covariate term can be replaced by an appropriate polynomial function of all the covariates. For example, for two covariates x_1 and x_2, the second-order function

$$\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

might be used, analogous to the polynomial response surface models of Chap. 16. Centered forms of these functions can also be obtained (see Sect. 8.7).

9.3   Least Squares Estimates

We now obtain the least squares estimates for the parameters in the analysis of covariance model, and then illustrate the need to use adjusted means to compare treatment effects.

9.3.1   Normal Equations (Optional)

To obtain the least squares estimates of the parameters in model (9.2.2), we need to minimize the sum of squared errors,

$$\sum_{i=1}^{v}\sum_{t=1}^{r_i} e_{it}^2 = \sum_{i=1}^{v}\sum_{t=1}^{r_i} (y_{it} - \mu - \tau_i - \beta(x_{it} - \overline{x}_{..}))^2 \ .$$

Differentiating this with respect to each parameter in turn and setting the corresponding derivative equal to zero gives the normal equations as

$$y_{..} = n\hat{\mu} + \sum_{i=1}^{v} r_i \hat{\tau}_i , \tag{9.3.4}$$

$$y_{i.} = r_i(\hat{\mu} + \hat{\tau}_i) + \hat{\beta} \sum_{t=1}^{r_i} (x_{it} - \overline{x}_{..}), \quad i = 1, \ldots, v , \tag{9.3.5}$$

$$\sum_{i=1}^{v} \sum_{t=1}^{r_i} y_{it}(x_{it} - \overline{x}_{..}) = \sum_{i=1}^{v} \sum_{t=1}^{r_i} (\hat{\mu} + \hat{\tau}_i)(x_{it} - \overline{x}_{..}) \tag{9.3.6}$$
$$+ \sum_{i=1}^{v} \sum_{t=1}^{r_i} \hat{\beta}(x_{it} - \overline{x}_{..})^2 .$$

There are $v + 2$ normal equations and $v + 2$ unknown parameters. However, Eq. (9.3.4) is the sum of the v equations given in (9.3.5), since $\Sigma r_i = n$ and $\Sigma\Sigma(x_{it} - \overline{x}_{..}) = 0$. Thus, the normal equations are not linearly independent, and the equations do not have a unique solution. However, the $v + 1$ equations in (9.3.5) and (9.3.6) are linearly independent and can be solved to obtain the unique least squares estimates for $\beta, \mu + \tau_1, \ldots, \mu + \tau_v$ given in the next subsection.

9.3.2   Least Squares Estimates and Adjusted Treatment Means

Under model (9.2.2) the expected value

$$E[\overline{Y}_{i.}] = \mu + \tau_i + \beta(\overline{x}_{i.} - \overline{x}_{..})$$

is an estimate of the mean response of the ith treatment when the value of the covariate x_{it} is $\overline{x}_{i.}$. So, unless the covariate means $\overline{x}_{i.}$ all happen to be equal, the difference of response means $\overline{y}_{i.} - \overline{y}_{s.}$ does not estimate $\tau_i - \tau_s$ and cannot be used to compare treatment effects. The least squares estimates of the parameters in the model are obtained by solving the normal equations in optional Sect. 9.3.1 and are

$$\hat{\mu} + \hat{\tau}_i = \overline{y}_{i.} - \hat{\beta}(\overline{x}_{i.} - \overline{x}_{..}) , \quad i = 1, \ldots, v , \tag{9.3.7}$$

$$\hat{\beta} = sp_{xy}^* / ss_{xx}^*, \tag{9.3.8}$$

where

$$sp_{xy}^* = \sum_{i=1}^{v} \sum_{t=1}^{r_i} (x_{it} - \overline{x}_{i.})(y_{it} - \overline{y}_{i.}) \quad \text{and} \quad ss_{xx}^* = \sum_{i=1}^{v} \sum_{t=1}^{r_i} (x_{it} - \overline{x}_{i.})^2 .$$

In this notation, ss can be read as "sum of squares" and sp as "sum of products." In Exercise 2, the reader is asked to verify that $E[\hat{\beta}] = \beta$. Consequently,

$$E[\hat{\mu} + \hat{\tau}_i] = \mu + \tau_i .$$

Table 9.1   Hypothetical analysis of covariance data

i	x_{it}		y_{it}		$\overline{x}_{i.}$	$\overline{y}_{i.}$
1	20	44.29	39.51	42.87	35	49.11
	30	44.48	48.39	49.14		
	40	50.24	51.63	46.55		
	50	57.75	59.23	55.23		
2	70	48.67	56.79	52.03	85	61.68
	80	57.68	67.25	52.88		
	90	62.04	66.12	64.39		
	100	63.54	72.49	76.33		

The least squares estimators $\hat{\mu} + \hat{\tau}_i$ therefore estimate the mean response for the ith treatment at the value of the covariate equal to $\overline{x}_{..}$. We call the estimates $\hat{\mu} + \hat{\tau}_i$ the *adjusted means,* since they adjust the response mean $\overline{y}_{i.}$ by the amount $\hat{\beta}(\overline{x}_{i.} - \overline{x}_{..})$, which is equivalent to measuring the responses at the same point on the covariate scale. The need for this adjustment is illustrated in the following example.

Example 9.3.1 Adjusted versus unadjusted means

Table 9.1 contains hypothetical data arising from two treatments at various values of the covariate. Using Eqs. (9.3.7) and (9.3.8), one can show that the corresponding fitted model is

$$\hat{y}_{it} = \hat{\mu} + \hat{\tau}_i + 0.5372(x_{ij} - 60),$$

where

$$\hat{\mu} + \hat{\tau}_1 = 62.5416 \text{ and } \hat{\mu} + \hat{\tau}_2 = 48.2516.$$

The data and fitted model are plotted in Fig. 9.2. Observe that treatment one has the larger effect, and correspondingly the higher fitted line. However if the treatment effects were estimated as $\overline{y}_{1.}$ and $\overline{y}_{2.}$, it would appear that treatment two has the larger effect, since it has the larger unadjusted mean:

$$\overline{y}_{2.} = 61.68 > \overline{y}_{1.} = 49.11.$$

This bias in the treatment effect estimates is due to the relative values of $\overline{x}_{1.}$ and $\overline{x}_{2.}$.

These data provide an exaggerated illustration of the need for adjustment of treatment sample means in analysis of covariance.   □

9.4   Analysis of Covariance

For a completely randomized design and analysis of covariance model (9.2.2), a *one-way analysis of covariance* is used to test the null hypothesis $H_0 : \{\tau_1 = \tau_2 = \cdots = \tau_v\}$ against the alternative hypothesis H_A that at least two of the τ_i's differ. The test is based on the comparison of error sums of squares under the full and reduced models. If the null hypothesis is true with common treatment effect $\tau_i = \tau$, then the reduced model is

Fig. 9.2 Illustration of
bias if unadjusted means
$\overline{y}_{i.}$ are compared
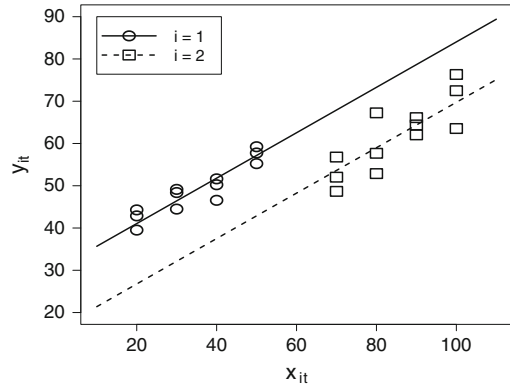
$$Y_{it} = \mu + \tau + \beta(x_{it} - \overline{x}_{..}) + \epsilon_{it}\,.$$

This is similar to a simple linear regression model, with constant $\beta_0 = \mu + \tau$, slope $\beta_1 = \beta$, and with
regressor x_{it} centered. Thus, if we replace x by $x_{it} - \overline{x}_{..}$ in the formula (8.5.6), p. 257, and the average
$\overline{x}_{..}$ in (8.5.6) by the averaged centered value 0, the least squares estimates under the reduced model
are

$$\hat{\mu} + \hat{\tau} = \overline{y}_{..} \quad \text{and} \quad \hat{\beta} = sp_{xy}/ss_{xx}\,,$$

where

$$sp_{xy} \;=\; \sum_{i=1}^{v}\sum_{t=1}^{r_i}(x_{it} - \overline{x}_{..})y_{it} \;=\; \sum_{i=1}^{v}\sum_{t=1}^{r_i}(x_{it} - \overline{x}_{..})(y_{it} - \overline{y}_{..})$$

and

$$ss_{xx} \;=\; \sum_{i=1}^{v}\sum_{t=1}^{r_i}(x_{it} - \overline{x}_{..})^2\,.$$

So,

$$\begin{aligned}
ssE_0 &= \sum_i \sum_t \left(y_{it} - \hat{\mu} - \hat{\tau} - \hat{\beta}(x_{it} - \overline{x}_{..})\right)^2 &\text{(9.4.9)}\\
&= \sum_i \sum_t \left(y_{it} - \overline{y}_{..} - sp_{xy}(x_{it} - \overline{x}_{..})/ss_{xx}\right)^2 \\
&= ss_{yy} - (sp_{xy})^2/ss_{xx}\,,
\end{aligned}$$

where $ss_{yy} = \sum_i \sum_t (y_{it} - \overline{y}_{..})^2$. The number of degrees of freedom for error is equal to the number
of observations minus a degree of freedom for the constant $\mu + \tau$ and one for the slope β; that is, $n - 2$.

Under the full analysis of covariance model (9.2.2), using the least squares estimates given in Eqs.
(9.3.7) and (9.3.8), the error sum of squares is

$$ssE = \sum_i \sum_t \left(y_{it} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}(x_{it} - \overline{x}_{..}) \right)^2$$

$$= \sum_i \sum_t \left(y_{it} - \overline{y}_{i.} + \hat{\beta}(\overline{x}_{i.} - \overline{x}_{..}) - \hat{\beta}(x_{it} - \overline{x}_{..}) \right)^2$$

$$= \sum_i \sum_t \left((y_{it} - \overline{y}_{i.}) - \hat{\beta}(x_{it} - \overline{x}_{i.}) \right)^2$$

$$= ss_{yy}^* - \hat{\beta}(sp_{xy}^*) \quad = \quad ss_{yy}^* - (sp_{xy}^*)^2/ss_{xx}^* \,, \tag{9.4.10}$$

where

$$ss_{xx}^* = \sum_i \sum_t (x_{it} - \overline{x}_{i.})^2 \,,$$

$$ss_{yy}^* = \sum_i \sum_t (y_{it} - \overline{y}_{i.})^2 \,,$$

$$sp_{xy}^* = \sum_i \sum_t (x_{it} - \overline{x}_{i.})(y_{it} - \overline{y}_{i.}) \,.$$

The values ss_{xx}^* and ss_{yy}^* can be obtained from a computer program as the values of ssE fitting the one-way analysis of variance models with x_{it} and y_{it} as the response variables, respectively. The number of error degrees of freedom is $n - v - 1$ (one less than the error degrees of freedom under the analysis of variance model due to the additional parameter β).

The sum of squares for treatments $ss(T|\beta)$ is the difference in the error sums of squares under the reduced and full models,

$$ss(T|\beta) = ssE_0 - ssE \tag{9.4.11}$$
$$= \left(ss_{yy} - (sp_{xy})^2/ss_{xx} \right) - \left(ss_{yy}^* - (sp_{xy}^*)^2/ss_{xx}^* \right) \,.$$

The difference in the error degrees of freedom for the reduced and full models is

$$(n - 2) - (n - v - 1) = v - 1 \,.$$

We denote the corresponding mean square by

$$ms(T|\beta) = ss(T|\beta)/(v - 1) \,.$$

If the null hypothesis is true, then

$$MS(T|\beta)/MSE \sim F_{v-1, n-v-1} \,,$$

so we can obtain a decision rule for testing $H_0 : \{\tau_1 = \tau_2 = \cdots = \tau_v\}$ against $H_A : \{\tau_i \text{ not all equal}\}$ as

$$\text{reject } H_0 \text{ if } ms(T|\beta)/msE > F_{v-1, n-v-1, \alpha}$$

at chosen significance level α. The information for testing equality of the treatment effects is typically summarized in an analysis of covariance table such as that shown in Table 9.2.

Table 9.2   Analysis of covariance for one linear covariate

Source of variation	Degrees of freedom	Sum of squares	Mean squares	Ratio
$T\|\beta$	$v-1$	$ss(T\|\beta)$	$\frac{ss(T\|\beta)}{v-1}$	$\frac{ms(T\|\beta)}{msE}$
$\beta\|T$	1	$ss(\beta\|T)$	$ss(\beta\|T)$	$\frac{ms(\beta\|T)}{msE}$
Error	$n-v-1$	ssE	msE	
Total	$n-1$	ss_{yy}		

Formulae

$ss(T|\beta) = \left(ss_{yy} - (sp_{xy})^2/ss_{xx}\right) - \left(ss^*_{yy} - (sp^*_{xy})^2/ss^*_{xx}\right)$

$ss(\beta|T) = (sp^*_{xy})^2/ss^*_{xx}$ $ssE = ss^*_{yy} - (sp^*_{xy})^2/ss^*_{xx}$

$ss_{xx} = \sum_i \sum_t (x_{it} - \overline{x}_{..})^2$ $ss_{yy} = \sum_i \sum_t (y_{it} - \overline{y}_{..})^2$

$sp_{xy} = \sum_i \sum_t (x_{it} - \overline{x}_{..})(y_{it} - \overline{y}_{..})$ $ss^*_{xx} = \sum_i \sum_t (x_{it} - \overline{x}_{i.})^2$

$sp^*_{xy} = \sum_i \sum_t (x_{it} - \overline{x}_{i.})(y_{it} - \overline{y}_{i.})$ $ss^*_{yy} = \sum_i \sum_t (y_{it} - \overline{y}_{i.})^2$

The table also includes information for testing the null hypothesis $H_0 : \{\beta = 0\}$ against the alternative hypothesis $H_A : \{\beta \neq 0\}$. The reduced model for this test is the one-way analysis of variance model (3.3.1), for which $Y_{it} = \mu + \tau_i + \epsilon_{it}$. From Chap. 3, the corresponding error sum of squares is

$$ssE_0 = \sum_i \sum_t (y_{it} - \overline{y}_{i.})^2 = ss^*_{yy},$$

and the number of error degrees of freedom is $n - v$. The error sum of squares for the full model is given in (9.4.10). Denoting the difference in error sums of squares by $ss(\beta|T)$, we have

$$ss(\beta|T) = ssE_0 - ssE = (sp^*_{xy})^2/ss^*_{xx} = \hat{\beta}^2 ss^*_{xx}.$$

The difference in the error degrees of freedom is $(n-v) - (n-v-1) = 1$, so the corresponding mean square, $ms(\beta|T)$, has the same value $ss(\beta|T)$. Under the assumptions of the analysis of covariance model (9.2.2), if $H_0 : \{\beta = 0\}$ is true, then

$$MS(\beta|T)/MSE \sim F_{1,n-v-1}.$$

Thus, the decision rule for testing $H_0 : \{\beta = 0\}$ against $H_A : \{\beta \neq 0\}$, at significance level α, is

$$\text{reject } H_0 \text{ if } ms(\beta|T)/msE > F_{1,n-v-1,\alpha}.$$

Example 9.4.1 Balloon experiment, continued

Consider the balloon experiment of Meily Lin, in which she compared the effects of four colors on balloon inflation time. In Example 5.5.1, p. 108, the standardized residuals were plotted against the run order of the observations. The plot, reproduced in Fig. 9.3, shows a clear linear decreasing trend in the residuals. This trend indicates a definite lack of independence in the error terms under the one-way analysis of variance model, but the trend can be eliminated by including the run order as a covariate in the model.

The analysis of covariance table for this experiment is shown in Table 9.3. Residual plots for checking the model assumptions will be discussed in Sects. 9.6 and 9.7 and reveal no anomalies.

Fig. 9.3 Residual plot for the balloon experiment
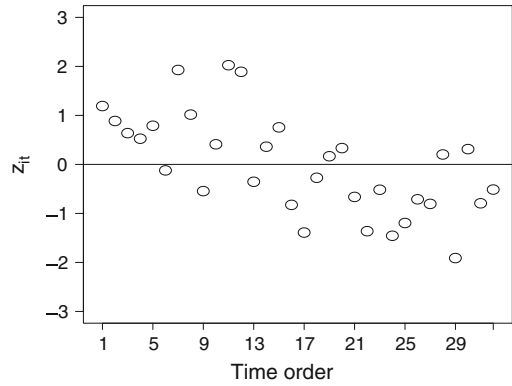
Table 9.3 Analysis of covariance for the balloon experiment

Source of variation	Degrees of freedom	Sum of squares	Mean squares	Ratio
$T\|\beta$	3	127.679	42.560	6.32
$\beta\|T$	1	120.835	120.835	17.95
Error	27	181.742	6.731	
Total	31	430.239		

The decision rule for testing equality of the treatment effects is to

$$\text{reject } H_0 : \{\tau_1 = \cdots = \tau_v\} \text{ if } ms(T|\beta)/msE = 6.32 \ > \ F_{3,27,\alpha} \,.$$

Since $F_{3,27,.01} = 4.60$, the null hypothesis is rejected at significance level $\alpha = 0.01$, and we can conclude that there is a difference in inflation times for the different colors of balloon.

Of secondary interest is the test of $H_0 : \{\beta = 0\}$ against $H_A : \{\beta \neq 0\}$. The decision rule is to reject the null hypothesis if $ms(\beta|T)/msE = 17.95 > F_{1,27,\alpha}$. Again, the null hypothesis is rejected at significance level $\alpha = 0.01$, since $F_{1,27,.01} = 7.68$. We may conclude that the apparent linear trend in the inflation times due to order is a real trend and not due to random error.

9.5   Treatment Contrasts and Confidence Intervals

9.5.1   Individual Confidence Intervals

Since $\mu + \tau_i$ is estimable under model (9.2.2), any treatment contrast $\sum_i c_i \tau_i$ ($\sum_i c_i = 0$) is also estimable. From (9.3.7), $\sum_i c_i \tau_i$ has least squares estimator

$$\sum_i c_i \hat{\tau}_i = \sum_i c_i (\hat{\mu} + \hat{\tau}_i) = \sum_i c_i \left(\overline{Y}_{i.} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) \right) = \sum_i c_i \left(\overline{Y}_{i.} - \hat{\beta}\bar{x}_{i.} \right) . \qquad (9.5.12)$$

(The term $\Sigma c_i \hat{\beta}\bar{x}_{..}$ is zero, since $\Sigma c_i = 0$.) Now, $\text{Var}\left(\overline{Y}_{i.}\right) = \sigma^2/r_i$, and it can be shown that $\text{Var}(\hat{\beta}) = \sigma^2/ss_{xx}^*$ and $\text{Cov}(\overline{Y}_{i.}, \hat{\beta}) = 0$. Using these results and (9.5.12), the variance of $\sum_i c_i \hat{\tau}_i$ is

$$\mathrm{Var}\left(\sum_i c_i \hat{\tau}_i\right) = \mathrm{Var}\left(\sum_i c_i \overline{Y}_{i.}\right) + \mathrm{Var}\left(\hat{\beta}\sum_i c_i \overline{x}_{i.}\right)$$

$$= \sigma^2\left(\sum_i \frac{c_i^2}{r_i}\right) + \left(\frac{\sigma^2}{ss_{xx}^*}\right)\left(\sum_i c_i \overline{x}_{i.}\right)^2 .$$

So, the estimated variance is

$$\widehat{\mathrm{Var}}(\Sigma c_i \hat{\tau}_i) = msE\left(\sum_i \frac{c_i^2}{r_i} + \frac{\left(\sum_i c_i \overline{x}_{i.}\right)^2}{ss_{xx}^*}\right). \tag{9.5.13}$$

From (9.5.12), the least squares estimator $\sum_i c_i \hat{\tau}_i$ is a function of $\overline{Y}_{i.}$ and $\hat{\beta}$. Since Y_{ij} has a normal distribution, both $\overline{Y}_{i.}$ and $\hat{\beta}$ are normally distributed. Consequently, $\sum_i c_i \hat{\tau}_i$ also has a normal distribution. Also, MSE/σ^2 has a chi-squared distribution with $n - v - 1$ degrees of freedom. Then, for any treatment contrast $\sum_i c_i \tau_i$, it follows that

$$\frac{\sum c_i \hat{\tau}_i - \sum c_i \tau_i}{\sqrt{\widehat{\mathrm{Var}}(\sum c_i \hat{\tau}_i)}} \sim t_{n-v-1} .$$

So, a $100(1 - \alpha)\%$ confidence interval for $\sum_i c_i \tau_i$ is

$$\sum_i c_i \tau_i \in \left(\sum_i c_i \hat{\tau}_i \pm t_{n-v-1,\alpha/2}\sqrt{\widehat{\mathrm{Var}}\left(\sum_i c_i \hat{\tau}_i\right)}\right). \tag{9.5.14}$$

9.5.2 Multiple Comparisons

The multiple comparison methods of Bonferroni and Scheffé are applicable in the analysis of covariance setting. However, since the adjusted treatment means $\hat{\mu} + \hat{\tau}_i = \overline{Y}_{i.} - \hat{\beta}(\overline{x}_{i.} - \overline{x}_{..})$ are not independent unless the $\overline{x}_{i.}$ are all equal, the methods of Tukey and Dunnett are not known to apply. It is believed that Tukey's method does still control the experimentwise confidence level in this case, but there is no known proof of this conjecture.

Confidence intervals are obtained as

$$\sum_i c_i \tau_i \in \left(\sum_i c_i \hat{\tau}_i \pm w\sqrt{\widehat{\mathrm{Var}}\left(\sum_i c_i \hat{\tau}_i\right)}\right), \tag{9.5.15}$$

where w is the appropriate critical coefficient. For the Bonferroni method and m predetermined treatment contrasts, $w = t_{n-v-1,\alpha/2m}$. For the Scheffé method for all treatment contrasts, $w = \sqrt{(v-1)F_{v-1,n-v-1,\alpha}}$. Formulae for the estimate $\Sigma c_i \hat{\tau}_i$ and the corresponding estimated variance are given in Eqs. (9.5.12) and (9.5.13).

Example 9.5.1 `Balloon experiment, continued`

We now illustrate the Scheffé method of multiple comparisons to obtain simultaneous 95% confidence intervals for all pairwise treatment comparisons for the balloon experiment of Example 9.4.1. (The data

Table 9.4   Scheffé pairwise comparisons for the balloon experiment; overall confidence level is 95%

i	s	$\hat{\tau}_i - \hat{\tau}_s$	$\sqrt{\widehat{\text{Var}}(\hat{\tau}_i - \hat{\tau}_s)}$	msd
1	2	-4.106	1.298	3.868
1	3	-3.801	1.299	3.871
1	4	0.071	1.297	3.865
2	3	0.304	1.301	3.877
2	4	4.176	1.298	3.868
3	4	3.872	1.298	3.868

are in Table 3.13, p. 68.) For pairwise comparisons, the confidence intervals are obtained from (9.5.15),

$$\tau_i - \tau_s \in \left(\hat{\tau}_i - \hat{\tau}_s \; \pm \; \sqrt{3F_{3,27,.05}} \sqrt{\widehat{\text{Var}}(\hat{\tau}_i - \hat{\tau}_s)} \right) ,$$

where $\hat{\tau}_i - \hat{\tau}_s = (\overline{y}_{i.} - \overline{y}_{s.}) - \hat{\beta}(\overline{x}_{i.} - \overline{x}_{s.})$. The treatment and covariate means are

$$\overline{y}_{1.} = 18.337, \quad \overline{y}_{2.} = 22.575, \quad \overline{y}_{3.} = 21.875, \quad \overline{y}_{4.} = 18.187,$$
$$\overline{x}_{1.} = 16.250, \quad \overline{x}_{2.} = 15.625, \quad \overline{x}_{3.} = 17.500, \quad \overline{x}_{4.} = 16.625,$$

and from (9.3.8), we obtain

$$\hat{\beta} = sp_{xy}^* / ss_{xx}^* = -572.59/2713.3 = -0.21103 .$$

Now, $msE = 6.731$ from Table 9.3, so

$$\begin{aligned}
\widehat{\text{Var}}(\hat{\tau}_i - \hat{\tau}_s) &= msE \left(\frac{1}{8} + \frac{1}{8} + \frac{(\overline{x}_{i.} - \overline{x}_{s.})^2}{ss_{xx}^*} \right) \\
&= (6.731) \left(0.25 + \frac{(\overline{x}_{i.} - \overline{x}_{s.})^2}{2713.3} \right) \\
&= 1.68275 + (0.00248)(\overline{x}_{i.} - \overline{x}_{s.})^2 .
\end{aligned}$$

Using the critical coefficient $w = \sqrt{3F_{3,27,.05}} = \sqrt{3 \times 2.96}$, one can obtain the confidence interval information given in Table 9.4. The estimated difference exceeds the minimum significant difference

$$msd = w\sqrt{\widehat{\text{Var}}(\hat{\tau}_i - \hat{\tau}_s)} \;\; \text{with} \;\; w = \sqrt{3F_{3,27,.05}}$$

for the first two and last two comparisons. One can conclude from the corresponding confidence intervals that the mean time to inflate balloons is longer for color 2 (yellow) than for colors 1 and 4 (pink and blue), and the mean inflation time for color 3 (orange) is longer than for color 4 (blue). At a slightly lower confidence level, we would also detect a difference in mean inflation times for colors 3 and 1 (orange and pink). The corresponding six intervals with overall confidence level 95% are

$$\tau_2 - \tau_1 \in (\ 0.238, 7.974), \quad \tau_3 - \tau_1 \in (-0.070, 7.672), \quad \tau_4 - \tau_1 \in (-3.936, 3.794),$$
$$\tau_2 - \tau_3 \in (-3.573, 4.181), \quad \tau_2 - \tau_4 \in (\ 0.308, 8.044), \quad \tau_3 - \tau_4 \in (\ 0.004, 7.740).$$

$\square$

Whenever the data are used to determine or modify the model, the confidence levels and error rates associated with any subsequent analyses of the same data will not be exactly as stated. Such is the case for the analyses presented in Example 9.5.1 for the balloon experiment, since the covariate "run order" was included in the model as a result of a trend observed in the residuals from the original analysis of variance model. Thus, although Scheffé's method was used, we cannot be certain that the overall level of the confidence intervals in Example 9.5.1 is exactly 95%.

9.6   Using SAS Software

Table 9.5 contains a sample SAS program for performing a one-way analysis of covariance involving a single covariate with a linear effect. The program uses the data from the balloon experiment discussed in Examples 9.4.1 and 9.5.1. The data are given in Table 3.13, p. 68. The experimenter was interested in comparing the effects of four colors (pink, yellow, orange, and blue) on the inflation time of balloons, and she collected eight observations per color. The balloons were inflated one after another by the same person. Residual analysis for the one-way analysis of variance model showed a linear trend in the residuals plotted against run order (Fig. 9.3, p. 293). Hence, run order is included in the model here as a linear covariate. To obtain the "centered" form of the model, as in model (9.2.2), a centered variable has been created immediately after the INPUT statement, using the SAS statement

```
X = RUNORDER - 16.5;
```

where 16.5 is the average value of RUNORDER.

In Table 9.5, PROC GLM is used to generate the analysis of covariance. The output is shown in Fig. 9.4. The treatment factor COLOR has been included in the CLASS statement to generate a parameter τ_i for each level of the treatment factor "color," while the covariate X has been excluded from the class statement so that it is included in the model as a regressor, or covariate, as in model (9.2.2). To obtain the "uncentered" form of the model, as in model (9.2.1), the variable RUNORDER would replace X throughout the program. The output in Fig. 9.4 would not change, since only the definition of the constant in the model has been altered.

The information for testing the null hypotheses $H_0^T : \{\tau_1 = \cdots = \tau_4\}$ against $H_A^T : \{H_0^T \text{ not true}\}$ and $H_0 : \{\beta = 0\}$ against $H_A : \{\beta \neq 0\}$ is in Fig. 9.4 under the heading Type III SS. Specifically, $ss(T|\beta) = 127.678829$ and $ss(\beta|T) = 120.835325$. The corresponding ratio statistics and p-values are listed under F Value and Pr > F, respectively. Since the p-values are very small, both null hypotheses would be rejected for any reasonable overall significance level. Thus, there are significant differences in the effects of the four colors on inflation time after adjusting for linear effects of run order. Also, there is a significant linear trend in mean inflation as a function of run order after adjusting for the treatment effects. The least squares estimate for β is negative ($\hat{\beta} = -0.211$), so the trend is decreasing, as we saw in Fig. 9.3.

The Type I and Type III sums of squares for color are similar but not quite equal, indicating that the treatment effects and the covariate effect are not independent. This is because the comparison of treatment effects is a comparison of the adjusted means, which do depend on β, since the covariate means $\overline{x}_{i.}$ are not all equal for these data.

Table 9.5 SAS program for analysis of covariance—Balloon experiment

```
DATA;
  INPUT RUNORDER COLOR INFTIME;
  X = RUNORDER - 16.5;
  LINES;
      1 1 22.0
      2 3 24.6
      3 1 20.3
      4 4 19.8
      : :   :
     30 1 19.3
     31 1 15.9
     32 3 20.3
;
PROC GLM;
  CLASS COLOR;
  MODEL INFTIME = COLOR X;
  ESTIMATE '1-2'  COLOR    1 -1  0  0;
  ESTIMATE '1-3'  COLOR    1  0 -1  0;
  ESTIMATE '1-4'  COLOR    1  0  0 -1;
  ESTIMATE '2-3'  COLOR    0  1 -1  0;
  ESTIMATE '2-4'  COLOR    0  1  0 -1;
  ESTIMATE '3-4'  COLOR    0  0  1 -1;
  ESTIMATE 'BETA'   X    1;
  OUTPUT OUT=B P=PRED R=Z;
PROC STANDARD STD=1;
  VAR Z;
PROC RANK NORMAL=BLOM OUT=C;
  VAR Z;
  RANKS NSCORE;
PROC SGPLOT;
  SCATTER Y = Z X = RUNORDER;
  YAXIS VALUES = (-2 TO 2 BY 1);
  XAXIS LABEL = "Run Order" VALUES = (0 TO 35 BY 5);
PROC SGPLOT; SCATTER Y = Z X = PRED;
PROC SGPLOT; SCATTER Y = Z X = COLOR;
PROC SGPLOT; SCATTER Y = Z X = NSCORE;
```

ESTIMATE statements under PROC GLM are used to generate the least squares estimate and estimated standard error for each pairwise comparison of treatment effects and for the coefficient β of the covariate. The standard errors of each $\hat{\tau}_i - \hat{\tau}_j$ are not quite equal but are all approximately 1.30. To compare all treatment effects pairwise using Scheffé's method and a simultaneous 95% confidence level, the calculations proceed as shown in Example 9.5.1.

The OUTPUT statement under PROC GLM and the procedures PROC STANDARD, PROC RANK, and PROC SGPLOT are used as they were in Chap. 5 to generate four residual plots. The resulting plots (not shown) show no problems with the model assumptions. Of special note, the plot of the residuals against run order in Fig. 9.5 no longer shows any trend, so the linear run order covariate has apparently adequately modeled any run order dependence in the observations.

Fig. 9.4 Output from SAS
PROC GLM

The SAS System

The GLM Procedure

Dependent Variable: INFTIME

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	248.4965749	62.1241437	9.23	<.0001
Error	27	181.7421751	6.7311917		
Corrected Total	31	430.2387500			

R-Square	Coeff Var	Root MSE	INFTIME Mean
0.577578	12.81607	2.594454	20.24375

Source	DF	Type I SS	Mean Square	F Value	Pr > F
COLOR	3	127.6612500	42.5537500	6.32	0.0022
X	1	120.8353249	120.8353249	17.95	0.0002

Source	DF	Type III SS	Mean Square	F Value	Pr > F
COLOR	3	127.6788293	42.5596098	6.32	0.0022
X	1	120.8353249	120.8353249	17.95	0.0002

Parameter	Estimate	Standard Error	t Value	Pr > |t|
1-2	-4.10560387	1.29760048	-3.16	0.0038
1-3	-3.80129227	1.29872024	-2.93	0.0069
1-4	0.07086232	1.29736147	0.05	0.9568
2-3	0.30431160	1.30058436	0.23	0.8168
2-4	4.17646618	1.29818288	3.22	0.0034
3-4	3.87215459	1.29795891	2.98	0.0060
BETA	-0.21103382	0.04980823	-4.24	0.0002

A test for equality of slopes as discussed in Sect. 9.2.1 can be generated using the SAS statements

```
PROC GLM;  CLASS COLOR;
  MODEL INFTIME = COLOR X COLOR*X;
```

The interaction term COLOR*X will be significantly different from zero if the linear run order trends
are not the same for each color.

Fig. 9.5 SAS plot of z_{it} against run order
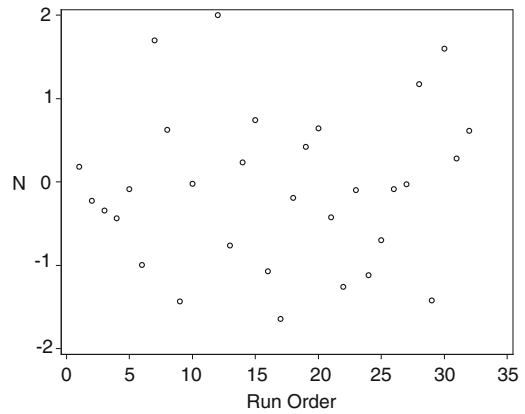
9.7   Using R Software

Table 9.6 contains a sample R program for performing a one-way analysis of covariance involving a single covariate with a linear effect. The program uses the data from the balloon experiment discussed in Examples 9.4.1 and 9.5.1. The data are given in Table 3.13, p. 68. The experimenter was interested in comparing the effects of four colors (pink, yellow, orange, and blue) on the inflation time of balloons, and she collected eight observations per color. The balloons were inflated one after another by the same person. Residual analysis for the one-way analysis of variance model showed a linear trend in the residuals plotted against run order (Fig. 9.3, p. 293). Hence, run order is included in the model here as a linear covariate. To obtain the "centered" form of the model, as in model (9.2.2), a centered variable x has been created after reading the data from file, using the R statement

```
x = Order - 16.5
```

within `balloon.data`, where 16.5 is the average value of `Order`.

In the second block of code in Table 9.6, the linear models function `lm` and related functions are used to generate the analysis of covariance. Selected output is shown in Table 9.7. The factor variable `fC` has been included in the model to generate a parameter τ_i for each level of the treatment factor "color," while the covariate `x`, because it is a numeric variable but not a factor variable, is included in the model as a regressor, or covariate, as in model (9.2.2). To obtain the "uncentered" form of the model, as in model (9.2.1), the variable `Order` would replace `x` throughout the program. The output in Table 9.7 would not change, since only the definition of the constant in the model would be altered.

The information for testing the null hypotheses $H_0^T : \{\tau_1 = \cdots = \tau_4\}$ against $H_A^T : \{H_0^T$ not true$\}$ and $H_0 : \{\beta = 0\}$ against $H_A : \{\beta \neq 0\}$ is in Table 9.7 under the `drop1` command that generates it. Specifically, the Type I sums of squares are $ss(T|\beta) \approx 128$ and $ss(\beta|T) \approx 121$. The corresponding ratio statistics and p-values are listed under `F value` and `PR(>F)`, respectively. Since the p-values are very small, both null hypotheses would be rejected for any reasonable overall significance level. Thus, there are significant differences in the effects of the four colors on inflation time after adjusting for linear effects of run order. Also, there is a significant linear trend in mean inflation as a function of run order after adjusting for the treatment effects. The least squares estimate for β is negative ($\hat{\beta} = -0.2110$), so the trend is decreasing, as we saw in Fig. 9.3.

The command `anova(model1)` generates type 3 sums of squares and the corresponding analysis of variance tables. The p-values for color for the Type I and Type III tests are similar but not identical,

Table 9.6   R program for analysis of covariance—Balloon experiment

```
balloon.data = read.table("data/balloon.txt", header=T)
head(balloon.data, 3)
balloon.data = within(balloon.data,
  {x = Order - 16.5; fC = factor(Color) })

options(contrasts = c("contr.sum", "contr.poly"))
model1 = lm(Time ~ fC + x, data=balloon.data)
summary(model1) # LSE etc. for covariate, model F-test
drop1(model1, ~., test="F") # Type 3 tests
anova(model1) # Type 1 tests

# Multiple comparisons: Scheffe's method
library(lsmeans)
lsmC = lsmeans(model1, ~ fC)
summary(contrast(lsmC, method="pairwise", adjust="Scheffe"),
        infer=c(T,T))

# Residual plots
balloon.data = within(balloon.data,
  {pred=fitted(model1); e=resid(model1); z=e/sd(e);
   n=length(e); q=rank(e); nscore=qnorm((q-0.375)/(n+0.25)) })
plot (z ~ Order, data=balloon.data); abline(h=0)
plot (z ~ pred, data=balloon.data); abline(h=0)
plot (z ~ Color, data=balloon.data); abline(h=0)
plot (z ~ nscore, data=balloon.data); qqline(balloon.data$z)
```

indicating that the corresponding Type I and Type III sums of squares are not quite equal, though both values have rounded to 128. This discrepancy, though minor, indicates that the treatment effects and the covariate effect are not independent. This is because the comparison of treatment effects is a comparison of the adjusted means, which do depend on β, since the covariate means $\overline{x}_{i.}$ are not all equal for these data.

The least squares means function $\mathtt{lsmeans}$ and the corresponding $\mathtt{summary}$ statement in the third block of code are used to compare all treatment effects pairwise using Scheffé's method and a simultaneous 95% confidence level (by default), generating the simultaneous 95% confidence intervals and related information shown at the bottom of Fig. 9.4. These results correspond to those of Example 9.5.1. The standard errors of each $\hat{\tau}_i - \hat{\tau}_j$ are not quite equal but are all approximately 1.30, so the widths of the confidence intervals obtained by Scheffé's method will be similar but not identical.

In the last block of code in Table 9.6, the saved predicted and residual values are used as they were in Chap. 5 to generate four residual plots. The resulting plots (not shown) show no problems with the model assumptions. Of special note, the plot of the residuals against run order (not shown here; see Fig. 9.5 for the same plot created in SAS) no longer shows any trend, so the linear run order covariate has apparently adequately modeled any run order dependence in the observations.

A test for equality of slopes as discussed in Sect. 9.2.1 can be generated using the R statements

```
model2 = lm(Time ~ fC + x + fC:x, data=balloon.data)
anova(model1, model2)
```

The interaction term $\mathtt{fC:x}$ will be significantly different from zero if the linear run order trends are not the same for each color.

Table 9.7 R selected output for analysis of variance and multiple comparisons

```
> summary(model1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
               :          :          :        :
x             -0.2110     0.0498    -4.24  0.00024

Residual standard error: 2.59 on 27 degrees of freedom
Multiple R-squared:  0.578,Adjusted R-squared:  0.515
F-statistic: 9.23 on 4 and 27 DF,  p-value: 0.000078

> drop1(model1, ~., test="F")

Single term deletions
       Df Sum of Sq RSS  AIC F value  Pr(>F)
<none>                182 65.6
fC      3       128 309 76.6    6.32 0.00217
x       1       121 303 79.9   17.95 0.00024

> anova(model1)

Analysis of Variance Table
Response: Time
          Df Sum Sq Mean Sq F value  Pr(>F)
fC         3    128    42.6    6.32 0.00218
x          1    121   120.8   17.95 0.00024
Residuals 27    182     6.7

> summary(contrast(lsmC, method="pairwise", adjust="Scheffe"),
+        infer=c(T,T))

 contrast  estimate      SE df   lower.CL  upper.CL t.ratio p.value
 1 - 2    -4.105604 1.2976 27 -7.9725957 -0.238612  -3.164  0.0341
 1 - 3    -3.801292 1.2987 27 -7.6716211  0.069037  -2.927  0.0557
 1 - 4     0.070862 1.2974 27 -3.7954172  3.937142   0.055  1.0000
 2 - 3     0.304312 1.3006 27 -3.5715725  4.180196   0.234  0.9966
 2 - 4     4.176466 1.2982 27  0.3077388  8.045194   3.217  0.0304
 3 - 4     3.872155 1.2980 27  0.0040946  7.740215   2.983  0.0497

Confidence level used: 0.95
Confidence-level adjustment: scheffe method for a family of 4 estimates
P value adjustment: scheffe method for a family of 4 tests
```

Exercises

1. Consider the hypothetical data of Example 9.3.1, in which two treatments are to be compared.

   (a) Fit the analysis of covariance model (9.2.1) or (9.2.2) to the data of Table 9.1, p. 289.
   (b) Plot the residuals against the covariate, the predicted values, and normal scores. Use the plots to evaluate the model assumptions.
   (c) Test for inequality of slopes, using a level of significance $\alpha = 0.05$.

Table 9.8 Bracket thickness x_{it} and plating thickness y_{it} in 10^{-5} inches for three vendors (Hicks 1965)

| | Vendor | | | | | |
| | 1 | | 2 | | 3 | |
t	x_{1t}	y_{1t}	x_{2t}	y_{2t}	x_{3t}	y_{3t}
1	110	40	60	25	62	27
2	75	38	75	32	90	24
3	93	30	38	13	45	20
4	97	47	140	35	59	13

Source Hicks (1965). Copyright © 1965 American Society for Quality. Reprinted with permission

(d) Test for equality of the treatment effects, using a significance level of $\alpha = 0.05$. Discuss the results.

(e) Construct a 95% confidence interval for the difference in treatment effects. Discuss the results.

2. (optional) Assume that the analysis of covariance model (9.2.2) holds, so that $Y_{it} = \mu + \tau_i + \beta(x_{it} - \overline{x}_{..}) + \epsilon_{it}$.

(a) Compute $E[Y_{it}]$.

(b) Verify that $sp^*_{xY} = \sum_i \sum_t (x_{it} - \overline{x}_{i.})Y_{it}$, given that $sp^*_{xY} = \sum_i \sum_t (x_{it} - \overline{x}_{i.})(Y_{it} - \overline{Y}_{i.})$.

(c) Show that $E[\hat{\beta}] = \beta$, where $\hat{\beta} = sp^*_{xY}/ss^*_{xx}$ and $ss^*_{xx} = \sum_i \sum_t (x_{it} - \overline{x}_{i.})^2$.

(d) Verify that $\mathrm{Var}(\hat{\beta}) = \sigma^2/ss^*_{xx}$ and $\mathrm{Cov}(\overline{Y}_{i.}, \hat{\beta}) = 0$.

(e) Verify that $E[\hat{\mu} + \hat{\tau}_i] = \mu + \tau_i$, where $\hat{\mu} + \hat{\tau}_i = \overline{Y}_{i.} - \hat{\beta}(\overline{x}_{i.} - \overline{x}_{..})$.

(f) Using the results of (c) and (e), argue that $\mu + \tau_i$ and β and all linear combinations of these are estimable.

3. **Zinc plating experiment**

The following experiment was used by C. R. Hicks (1965), *Industrial Quality Control*, to illustrate the possible bias caused by ignoring an important covariate. The experimental units consisted of 12 steel brackets. Four steel brackets were sent to each of three vendors to be zinc plated. The response variable was the thickness of the zinc plating, in hundred-thousandths of an inch. The thickness of each bracket before plating was measured as a covariate. The data are reproduced in Table 9.8.

(a) Plot y_{it} versus x_{it}, using the vendor index i as the plotting symbol. Discuss the relationship between plating thickness and bracket thickness before plating. Based on the plot, discuss appropriateness of the analysis of covariance model. Based on the plot, discuss whether there appears to be a vendor effect.

(b) Fit the analysis of covariance model (9.2.1) or (9.2.2) to the data.

(c) Plot the residuals against the covariate, predicted values, and normal scores. Use the plots to evaluate model assumptions.

(d) Test for equality of slopes, using a level of significance $\alpha = 0.05$.

(e) Test for equality of the vendor effects, using a significance level $\alpha = 0.05$.

(f) Fit the analysis of variance model to the data, ignoring the covariate.

(g) Using analysis of variance, ignoring the covariate, test for equality of the vendor effects using a significance level $\alpha = 0.05$.

Table 9.9 Data for the paper tower absorbancy experiment

Run	Treatment	AB	Drops	Time	Area	Rate	Absorbancy
1	2	12	89	50	121.00	1.780	0.7355
2	4	22	28	15	99.00	1.867	0.2828
3	2	12	47	22	121.00	2.136	0.3884
4	1	11	82	42	121.00	1.952	0.6777
5	5	31	54	30	123.75	1.800	0.4364
6	1	11	74	37	121.00	2.000	0.6116
7	4	22	29	14	99.00	2.071	0.2929
8	6	32	80	41	123.75	1.951	0.6465
9	3	21	25	11	99.00	2.272	0.2525
10	3	21	27	12	99.00	2.250	0.2727
11	6	32	83	40	123.75	2.075	0.6707
12	5	31	41	19	123.75	2.158	0.3313

(h) Compare and discuss the results of parts (e) and (g). For which model is *msE* smaller? Which model gives the greater evidence that vendor effects are not equal? What explanation can you offer for this?

4. **Paper towel absorbancy experiment**

   S. Bortnick, M. Hoffman, K.K. Lewis and C. Williams conducted a pilot experiment in 1996 to compare the effects of two treatment factors, brand and printing, on the absorbancy of paper towels. Three brands of paper towels were compared (factor A at 3 levels). For each brand, both white and printed towels were evaluated (factor B, 1=white, 2=printed). For each observation, water was dripped from above a towel, which was horizontally suspended between two pairs of books on a flat surface, until the water began leaking through to the surface below. The time to collect each observation was measured in seconds. Absorbancy was measured as the number of water drops absorbed per square inch of towel. The rate at which the water droplets fell to the towel was measured (in drops per second) as a covariate. The data are reproduced in Table 9.9.

   (a) Plot absorbancy versus rate, using the treatment level as the plotting symbol. Based on the plot, discuss appropriateness of the analysis of covariance model, and discuss whether there appear to be treatment effects.
   (b) Fit the one-way analysis of covariance model to the data.
   (c) Plot the residuals against the covariate, run order, predicted values, and normal scores. Use the plots to evaluate model assumptions.
   (d) Test for equality of slopes, using a level of significance $\alpha = 0.05$.
   (e) Test for equality of treatment effects, using a significance level $\alpha = 0.05$.
   (f) Conduct a two-way analysis of covariance. Test the main effects and interactions for significance.

5. **Catalyst experiment, continued**

   The catalyst experiment was described in Exercise 5 of Chap. 5. The data were given in Table 5.18, p. 134. There were twelve treatment combinations consisting of four levels of reagent, which we may recode as $A = 1$, $B = 2$, $C = 3$, $D = 4$, and three levels of catalyst, which we may recode as $X = 1$, $Y = 2$, $Z = 3$, giving the treatment combinations 11, 12, 13, 21, ..., 43.

   The order of observation of the treatment combinations is also given in Table 5.18.

(a) Fit a two-way complete model to the data and plot the residuals against the time order. If you are happy about the independence of the error variables, then check the other assumptions on the model and analyze the data. Otherwise, go to part (b).

(b) Recode the treatment combinations as $1, 2, \ldots, 12$. Fit an analysis of covariance model (9.2.1) or (9.2.2) to the data, where the covariate x_{it} denotes the time in the run order at which the tth observation on the ith treatment combination was made. Check all of the assumptions on your model, and if they appear to be satisfied, analyze the data.

(c) Plot the adjusted means of the twelve treatment combinations in such a way that you can investigate the interaction between the reagents and catalysts. Test the hypothesis that the interaction is negligible.

(d) Check the model for lack of fit; that is, investigate the treatment $\times$ time interaction. State your conclusions.

Complete Block Designs

10

10.1 Introduction

In step (b)(iii) of the checklist in Chap. 2, we raised the possibility that an experiment may involve one or more nuisance factors that, although not of interest to the experimenter, could have a major effect on the response. We classified these nuisance factors into three types: blocking factors, noise factors, and covariates. Different types of nuisance factors lead to different types of analyses, and the choice between these is revisited in Sect. 10.2.

The cotton-spinning experiment of Sect. 2.3, p. 13, illustrates some of the considerations that might lead an experimenter to include a blocking factor in the model and to adopt a block design. The most commonly used block designs are the complete block designs. These are defined in Sect. 10.3 and their randomization is illustrated. Models, multiple comparisons, and analysis of variance for randomized complete block designs in which each treatment is observed once in each block are given in Sect. 10.4 and those for more general complete block designs in Sect. 10.6. Model assumption checks are outlined briefly in Sect. 10.7. An analysis of the cotton-spinning experiment is described in Sect. 10.5 and, in Sect. 10.8, we illustrate the analysis of a complete block design with factorial treatment combinations. Analyses of complete block designs using the SAS and R computer packages are discussed in Sects. 10.9 and 10.10, respectively.

10.2 Blocks, Noise Factors or Covariates?

It is not always obvious whether to classify a nuisance factor as a blocking factor, a covariate, or a noise factor. The decision will often be governed by the goal of the experiment.

Nuisance factors are classified as *noise factors* if the objective of the experiment is to find settings of the treatment factors whose response is least affected by varying the levels of the nuisance factors. Settings of noise factors can usually be controlled during an experiment but are uncontrollable outside the laboratory. We will give some examples illustrating noise factors in Chap. 15.

Covariates are nuisance factors that cannot be controlled but can be measured prior to, or during, the experiment. Sometimes covariates are of interest in their own right, but when they are included in the model as nuisance variables, their effects are used to adjust the responses so that treatments can be compared as though all experimental units were identical (see Chap. 9).

A *block design* is appropriate when the goal of the experiment is to compare the effects of different treatments averaged over a range of different conditions. The experimental units are grouped into sets

in such a way that two experimental units in the same set are similar and can be measured under similar experimental conditions, but two experimental units in different sets are likely to give rise to quite different measurements even when assigned to the same treatment. The sets of similar experimental units are called *blocks*, and the conditions that vary from block to block form the levels of the *blocking factor*. The intent of blocking is to prevent large differences in the experimental units from masking differences between treatment effects, while at the same time allowing the treatments to be examined under different experimental conditions.

The levels of a blocking factor may be the values of a covariate that has been measured prior to the experiment and whose values are used to group the experimental units. More often, however, the levels of a blocking factor are groupings of characteristics that cannot be conveniently measured. For example, grouping the time slots in the same day into the same block, as was done for the cotton-spinning experiment in Sect. 2.3, ensures that environmental conditions within a block are fairly similar without the necessity of measuring them.

Since the levels of the blocking factor do not necessarily need to be measured, the block design is very popular. Agricultural experimenters may know that plots close together in a field are alike, while those far apart are not alike. Industrial experimenters may know that two items produced by one machine have similar characteristics, while those produced by two different machines are somewhat different. Medical experimenters may know that measurements taken on the same subject will be alike, while those taken on different subjects will not be alike. Consequently, blocks may be formed without actually knowing the precise levels of the blocking factor. Some more examples are given in the next section and throughout the chapter.

10.3   Design Issues

10.3.1   Block Sizes

Although it is perfectly possible for the numbers of experimental units in each block to be unequal, the most common setting, and the only one that we will examine here, is when the blocks are of the same size. We will use b to represent the number of blocks and k to represent the common block sizes.

Sometimes the block sizes are naturally defined, and sometimes they need to be specifically selected by the experimenter. In a bread-baking experiment, for example, the experimental units are the baking tins in different positions in the oven. If the temperature cannot be carefully controlled, there may be a temperature gradient from the top shelf to the bottom shelf of the oven, although the temperature at all positions within a shelf may be more or less constant. If the measured response is affected by temperature, then experimental units on the same shelf are alike, but those on different shelves are different. There is a natural grouping of experimental units into blocks defined by the shelf of the oven. Thus, the shelves are the blocks of experimental units and represent the levels of the blocking factor "temperature." The number b of blocks is the number of shelves in the oven. The block size k is the number of baking tins that can be accommodated on each shelf.

Block size is not always dictated by the experimental equipment. The size often needs to be determined by the judgment of the experimenter. For example, the data in Fig. 10.1 were gathered in a pilot experiment by Bob Belloto in the Department of Pharmacy at The Ohio State University. The data show the readings obtained by a breathalyzer for a given concentration of alcohol. Notice how the readings decrease over time. Likely causes for this decrease include changes in atmospheric conditions, evaporation of alcohol, and deterioration of the breathalyzer filters. In other experiments, such trends in the data can be caused by equipment heating over time, by variability of batches of raw material, by experimenter fatigue, etc.

Fig. 10.1 Pilot data for the breathalyzer experiment
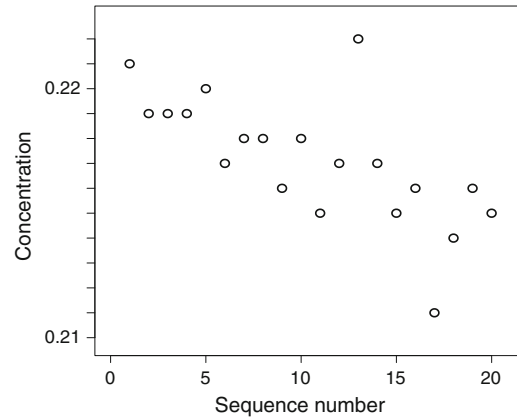
The block sizes for the breathalyzer experiment were chosen to be five, that is, the first five observations would be in one block, the next five in the next block, and so on. The reason for the choice was twofold. First, it can be seen from Fig. 10.1 that the observations in the pilot experiment seem to be fairly stable in groups of five. Secondly, the experiment was to be run by two different technicians, who alternated shifts, and five observations could be taken per shift. Thus the blocking factor was factorial in nature, and its levels represented combinations of time and technicians.

It is not uncommon in industry for an experiment to be automatically divided into blocks according to time of day as a precaution against changing experimental conditions. A pilot experiment using a single treatment such as that in the breathalyzer experiment is an ideal way of determining the necessity for blocking. If blocks were to be created when they are not needed, hypothesis tests would be less powerful and confidence intervals would be wider than those obtained via a completely randomized design.

10.3.2   Complete Block Design Definitions

Having decided on the block size and having grouped the experimental units into blocks of similar units, the next step is to assign the units to the levels of the treatment factors. The worst possible assignment of experimental units to treatments is to assign all the units within a block to one treatment, all units within another block to a second treatment, and so on. This assignment is bad because it does not allow the analysis to distinguish block differences from treatment differences. The effects of the treatment factors and the effects of the blocking factor are said to be *confounded*.

The best possible assignment is one that allocates to every treatment the same number of experimental units per block. This can be achieved only when the block size k is a multiple of v, the number of treatments. Such designs are called *complete block designs*, and in the special case of $k = v$, they have historically been called *randomized complete block designs* or, simply, *randomized block designs*. The historical name is unfortunate, since all block designs need to be randomized. Nevertheless, we will retain the name randomized complete block design for block size $k = v$ and use the name *general complete block design* for block size a larger multiple of v.

If the block size is not a multiple of v, then the block design is known as an *incomplete block design*. This term is sometimes reserved for the smaller designs where $k < v$, but we will find it convenient to classify all designs as either complete or incomplete. Incomplete block designs are more complicated to design and analyze than complete block designs, and we postpone their discussion to Chap. 11. For

complete block designs, every treatment is observed $s = v/k$ times in every block, and so is observed $r = bs$ times in the experiment.

10.3.3   The Randomized Complete Block Design

A *randomized complete block design* is a design with v treatments (which may be factorial treatment combinations) and with $n = bv$ experimental units grouped into b blocks of $k = v$ units in such a way that units within a block are alike and units in different blocks are substantially different. The $k = v$ experimental units within each block are randomly assigned to the v treatments so that each treatment is assigned one unit per block. Thus, each treatment appears once in every block ($s = 1$) and $r = b$ times in the design.

Example 10.3.1 `Bread-baking experiment`

An experimenter wishes to compare the shelf life of loaves made from $v = 4$ different bread doughs, coded 1, 2, 3, 4. An oven with three shelves will be used, and each shelf is large enough to take four baking tins. A temperature difference is anticipated between the different shelves but not in different positions within a shelf. The oven will be used twice, giving a total of six blocks defined by shelf/run of the oven, and the block size is $k = 4$ defined by the four positions on each shelf: FL, FR, BL, BR (front left, front right, back left, back right).

Since the block size is the same as the number of treatments, a randomized complete block design can be used. The experimental units (positions) in each block (shelf/run) are assigned at random to the four levels of the treatment factor (doughs) using the procedure described in Sect. 3.2, p. 31, for each block separately. For example, suppose we obtain the four 2-digit random numbers 74, 11, 39, 68 from a computer random number generator, or from Table A.1, and associate them in this order with the four treatments to be observed once each in block 1. If we now sort the random numbers into ascending order, the treatment codes are sorted into the order 2, 3, 4, 1. We can then allocate the experimental units in the order FL, FR, BL, BR to the randomly sorted treatments, and we obtain the randomized block shown in the first row of Table 10.1. The other randomized blocks in Table 10.1 are obtained in a similar fashion.

Notice that the randomization that we have obtained in Table 10.1 has allowed bread dough 1 to be observed four times in the back right position, and that dough 2 is never observed in this position. If a temperature difference in positions is present, then this could cause problems in estimating treatment differences, and the randomized complete block design is not the correct design to use. Instead, a row–column design (Chap. 12) should be used within each run of the oven.   □

Table 10.1 Example of a randomized complete block design

Block	Run	Shelf	FL	FR	BL	BR
1	1	1	2	3	4	1
2		2	1	2	3	4
3		3	4	3	2	1
4	2	1	2	4	3	1
5		2	2	4	1	3
6		3	3	2	4	1

10.3.4   The General Complete Block Design

A *general complete block design* is a design with v treatments (which may be factorial treatment combinations) and with $n = bvs$ experimental units grouped into b blocks of $k = vs$ units in such a way that units within a block are alike and units in different blocks are substantially different. The $k = vs$ experimental units within each block are randomly assigned to the v treatments so that each treatment is assigned s units per block. Thus, each treatment appears s times in every block and $r = bs$ times in the design.

Example 10.3.2 `DCIS` experiment

An experiment was run by Matthew Darr, David Holman, Nasser Kashou, and Angela Wendel in 2006 to examine the variability of a Dynamic Inline Conveyor Scale (DCIS). The DCIS is an automated system for weighing individual pieces of large fruit (for example, watermelons) while they are conveyed from one location to another. The device uses an optical switch to trigger the weighing system, a load cell to perform the weighing operation and a computer-based data recording system. The objective of the experiment was to reduce the variability associated with the recorded weight of each piece of fruit. The researchers decided to examine the effects of two treatment factors. Treatment factor A was the length of time during which the weight was recorded (with three levels: 50, 75, 100, milliseconds; coded 1, 2, 3). Treatment factor B was the position of the optical switch (with two levels; 1 inch and 2 inches from the end of the scale plate; coded 1, 2). Thus there were six treatments (treatment combinations), coded as follows:

$$
\begin{aligned}
(50\,\text{millisec}, 1\,\text{inch}) &= 1, & (50\,\text{millisec}, 2\,\text{inch}) &= 2, \\
(75\,\text{millisec}, 1\,\text{inch}) &= 3, & (75\,\text{millisec}, 2\,\text{inch}) &= 4, \\
(100\,\text{millisec}, 1\,\text{inch}) &= 5, & (100\,\text{millisec}, 2\,\text{inch}) &= 6,
\end{aligned}
$$

The currently employed setting was treatment 4 (time 75 milliseconds, switch at 2 inches from the scale plate). The experimenters wished to see if there was a better setting, while taking into account a range of possible lubrication levels of the conveyor system. Changing the lubrication levels was difficult and time consuming, so the experiment was run in two blocks at the extreme levels of lubrication. In block 1, the conveyor pan was completely saturated with oil lubricant and, in block 2, all lubricant was removed from the conveyor pan. In each block, there were $s = 2$ observations on each treatment. A random assignment of the $k = sv = 12$ experimental units (time slots) to the treatments within each block gave the following observation order:

$$
\begin{aligned}
\text{Block1}: &\ \ 1\ 5\ 2\ 4\ 1\ 5\ 2\ 6\ 4\ 3\ 3\ 6 \\
\text{Block2}: &\ \ 3\ 1\ 2\ 4\ 2\ 5\ 6\ 1\ 5\ 4\ 3\ 6
\end{aligned}
$$

For each observation on a specified treatment in a block, the response was a function, called "uncertainty", of the standard deviation of 30 weighings of a watermelon. The data are shown in Table 10.8 and discussed in Example 10.6.1.   □

10.3.5   How Many Observations?

If the block size $k = sv$ is pre-determined, we can calculate the number of blocks b that are required to achieve a confidence interval of given length, or a hypothesis test of desired power, in much the same way as we calculated sample sizes in Chap. 6. If the number of blocks b is fixed, but the block

sizes can be large, then the same techniques can be used to calculate s for a general complete block design. A calculation of the required number of blocks using confidence intervals is illustrated for a randomized complete block design in Sect. 10.5.2, and a calculation of the required block size using the power of a test is done in Sect. 10.6.3 for a general complete block design.

10.4   Analysis of Randomized Complete Block Designs

10.4.1   Model and Analysis of Variance

The standard model for a randomized complete block design (with $s = 1$ observation on each treatment in each block) is

$$
\begin{aligned}
Y_{hi} &= \mu + \theta_h + \tau_i + \epsilon_{hi} \,, \\
\epsilon_{hi} &\sim N(0, \sigma^2) \,, \\
\epsilon_{hi}'\text{s are} &\text{ mutually independent} \,, \\
h = 1, \ldots, b; &\quad i = 1, \ldots, v \,,
\end{aligned}
\tag{10.4.1}
$$

where μ is a constant, θ_h is the effect of the hth block, τ_i is the effect of the ith treatment, Y_{hi} is the random variable representing the measurement on treatment i observed in block h, and ϵ_{hi} is the associated random error. We will call this standard model the *block–treatment model*.

Notice that the block–treatment model does not include a term for the interaction between blocks and treatments. If interaction effects were to be included in the model, there would be no degrees of freedom for error with which to estimate the error variance (cf. Sect. 6.7). In many blocked experiments, absence of block×treatment interaction is a reasonable assumption. However, if interaction is suspected in a given experiment, then the block size must be increased to allow its estimation (as in Sect. 10.6).

The block–treatment model (10.4.1) looks similar to the two-way main-effects model (6.2.3) for two treatment factors in a completely randomized design with one observation per cell. Not surprisingly, then, the analysis of variance table in Table 10.2 for the randomized complete block design looks similar to the two-way analysis of variance table in Table 6.7, p. 170, for two treatment factors and one observation per treatment combination. There is, however, an important difference. In a completely randomized design, the experimental units are randomly assigned to the treatment combinations, and

Table 10.2   Analysis of variance: randomized complete block design

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio
Block	$b - 1$	$ss\theta$	$ms\theta = \frac{ss\theta}{b-1}$	–
Treatment	$v - 1$	ssT	$msT = \frac{ssT}{v-1}$	$\frac{msT}{msE}$
Error	$bv - b - v + 1$	ssE	$msE = \frac{ssE}{bv-b-v+1}$	
Total	$bv - 1$	$sstot$		
Computational formulae				

$$ss\theta = v \sum_h \overline{y}_{h.}^2 - bv\overline{y}_{..}^2 \qquad\qquad sstot = \sum_h \sum_i y_{hi}^2 - bv\overline{y}_{..}^2$$

$$ssT = b \sum_i \overline{y}_{.i}^2 - bv\overline{y}_{..}^2 \qquad\qquad ssE = sstot - ss\theta - ssT$$

so to the levels of *both* factors. On the other hand, in a block design, although observations are taken on all combinations of treatments and blocks, the experimental units are randomly assigned to the levels of the treatment factor only. The levels of the block factor represent intentional groupings of the experimental units. This leads to some controversy as to whether or not a test of equality of block effects is valid. However, when blocks represent nuisance sources of variation, we do not need to know much about the block effects since it is very unlikely that we can use the identical blocks again. So, rather than testing for equality of block effects, we will merely compare the block mean square $ms\theta$ with the error mean square msE to determine whether or not blocking was beneficial in the experiment at hand.

If $ms\theta$ is considerably larger than msE, this suggests that the creation of blocks was worthwhile in the sense of having reduced the size of the error mean square. If $ms\theta$ is less than msE, then the creation of blocks was not helpful and, in fact, has lowered the power of hypothesis tests and increased the lengths of confidence intervals for treatment contrasts. The comparison of $ms\theta$ and msE is not a significance test. There is no statistical conclusion about the equality of block effects. The comparison is merely an assessment of the usefulness of having created blocks in this particular experiment and does provide some information for the planning of future similar experiments. Of course, if $ms\theta$ is less than msE, it is *not* valid to pretend that the experiment was designed as a completely randomized design and to remove the block effects from the model—the randomization is not correct for a completely randomized design.

For testing hypotheses about treatment effects, we can use the analogy with the two-way main-effects model. The decision rule for testing the null hypothesis $H_0 : \{\tau_1 = \tau_2 = \cdots = \tau_v\}$, that the treatments have the same effect on the response, against the alternative hypothesis $H_A : \{$at least two of the τ_i differ$\}$ is

$$\text{reject } H_0 \text{ if } msT/msE > F_{v-1, bv-b-v+1, \alpha} \tag{10.4.2}$$

for some chosen significance level α, where msT and msE are defined in Table 10.2.

Example 10.4.1 `Resting metabolic rate experiment`

In the 1993 issue of *Annals of Nutrition and Metabolism*, R. C. Bullough and C. L. Melby describe an experiment that was run to compare the effects of inpatient and outpatient protocols on the in-laboratory measurement of resting metabolic rate (RMR) in humans. A previous study had indicated measurements of RMR on elderly individuals to be 8% higher using an outpatient protocol than with an inpatient protocol. If the measurements depend on the protocol, then comparison of the results of studies conducted by different laboratories using different protocols would be difficult. The experimenters hoped to conclude that the effect on RMR of different protocols was negligible.

The experimental treatments consisted of three protocols: (1) an inpatient protocol in which meals were controlled—the patient was fed the evening meal and spent the night in the laboratory, then RMR was measured in the morning; (2) an outpatient protocol in which meals were controlled—the patient was fed the same evening meal at the laboratory but spent the night at home, then RMR was measured in the morning; and (3) an outpatient protocol in which meals were not strictly controlled—the patient was instructed to fast for 12 hours prior to measurement of RMR in the morning. The three protocols formed the $v = 3$ treatments in the experiment.

Since subjects tend to differ substantially from each other, error variability can be reduced by using the subjects as blocks and measuring the effects of all treatments for each subject. In this experiment, there were nine subjects (healthy, adult males of similar age) and they formed the $b = 9$ levels of a blocking factor "subject." Every subject was measured under all three treatments (in a random order), so the blocks were of size $k = 3 = v$. RMR readings were taken over a one-hour period shortly after the subject arrived in the laboratory. The data collected during the second 30 minutes of testing are

Table 10.3 Data for the resting metabolic rate experiment

Subject	Protocol		
	1	2	3
1	7131	6846	7095
2	8062	8573	8685
3	6921	7287	7132
4	7249	7554	7471
5	9551	8866	8840
6	7046	7681	6939
7	7715	7535	7831
8	9862	10087	9711
9	7812	7708	8179

Source Bullough and Melby (1993). Copyright © 1993 Karger, Basel. Reprinted with permission

Fig. 10.2 Resting metabolic rates by protocol and subject
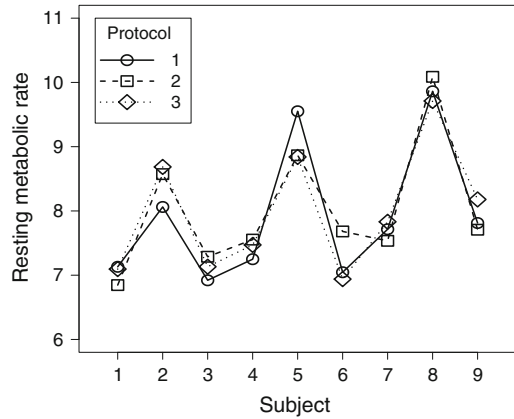
given in Table 10.3 and are plotted in Fig. 10.2. The figure clearly suggests large subject differences, but no consistent treatment differences.

The analysis of variance is shown in Table 10.4. The value of $ms\theta$ is 37 times larger than msE, indicating that blocking by subject has greatly reduced the error variance estimate. So a block design was a good choice for this experiment.

The null hypothesis of no difference in the protocols cannot be rejected at any reasonable selection of α, since $msT/msE = 0.23$. The ratio tells us that the average variability of the measurements from one protocol to another was four times smaller than the measurement error variability. This is unusual, since measurements from one protocol to another must include measurement error. The p-value is 0.7950, indicating that there is only a 20% chance that we would see a value this small or smaller when there is no difference whatsoever in the effects of the protocols. Thus, we should ask how well the model fits the data—perhaps treatment–block interaction is missing from the model and has been included incorrectly in the error variability. Even if this were the case, however, there is still no indication that protocols 2 and 3 provide higher RMR readings than protocol 1—in fact, for six of the nine subjects, one or both of these outpatient protocols resulted in lower readings than the inpatient protocol.

It is not possible to check the model assumptions of equal error variances for each cell because of the small amount of data. But we can check the equal-variance assumptions for the different levels of the treatment factor. We find that the variances of the unstandardized residuals are very similar for

Table 10.4 Analysis of variance for the resting metabolic rate experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Subject	8	23,117,462.30	2,889,682.79	–	–
Protocol	2	35,948.74	17,974.37	0.23	0.7950
Error	16	1,235,483.26	77,217.70		
Total	26	24,388,894.30			

the three protocols. The normality assumption seems to be reasonable. The only possible outlier is the observation for protocol 1, subject 5, but its removal does not change the above conclusions.

In their article, the experimenters discuss possible reasons for the fact that their conclusions differ from those of previous studies. Reasons included the different age of the subjects (27–29 years rather than 64–67 years) and the fact that they provided transport to the laboratory for the outpatients, whereas previous studies had not.   □

10.4.2   Multiple Comparisons

Since the block–treatment model (10.4.1) for the randomized complete block design is similar to the two-way main-effects model (6.2.3) for an experiment with two treatment factors and one observation per cell, the least squares estimator for each $\mu + \theta_h + \tau_i$ $(h = 1, \ldots, b;\ i = 1, \ldots, v)$ is similar to the estimator for each $\mu + \alpha_i + \beta_j$ $(i = 1, \ldots, a;\ j = 1, \ldots, b)$ in (6.5.26), p. 161, without the third subscript; that is,

$$\hat{\mu} + \hat{\theta}_h + \hat{\tau}_i = \overline{Y}_{h.} + \overline{Y}_{.i} - \overline{Y}_{..} . \tag{10.4.3}$$

It follows that any contrast $\Sigma c_i \tau_i$ (with $\Sigma c_i = 0$) in the treatment effects is estimable in the randomized complete block design and has least squares estimator

$$\Sigma c_i \hat{\tau}_i \ = \ \Sigma c_i \overline{Y}_{.i} ,$$

and least squares estimate $\Sigma c_i \overline{y}_{.i}$ with corresponding variance $\sigma^2 (\Sigma c_i^2 / b)$. As for the two-way main-effects model, all of the multiple comparison procedures of Chap. 4 are valid for treatment contrasts in the randomized complete block design. The formulae, adapted from (6.5.39), p. 166, are

$$\sum c_i \tau_i \ \in \ \left(\sum c_i \overline{y}_{.i} \pm w \sqrt{msE \sum c_i^2 / b} \right), \tag{10.4.4}$$

where the critical coefficients for the Bonferroni, Scheffé, Tukey, and Dunnett methods are, respectively,

$$w_B = t_{bv-b-v+1, \alpha/2m} ;\quad w_S = \sqrt{(v-1) F_{v-1, bv-b-v+1, \alpha}} ;$$
$$w_T = q_{v, bv-b-v+1, \alpha} / \sqrt{2} ;\quad w_{D2} = |t|^{(0.5)}_{v-1, bv-b-v+1, \alpha} . \tag{10.4.5}$$

Example 10.4.2 `Resting metabolic rate experiment, continued`

In the resting metabolic rate experiment, described in Example 10.4.1, p. 311, all three pairwise comparisons in the $v = 3$ protocol effects were of interest prior to the experiment, together with the contrast that compares the inpatient protocol with the two outpatient protocols. This latter contrast

has coefficient list $[1, -\frac{1}{2}, -\frac{1}{2}]$. Suppose that the experimenters had wished to calculate simultaneous 95% confidence intervals for these four contrasts. The formula is given in (10.4.4) and there are several possible choices for the critical coefficient. The Bonferroni method could be used for these specific four contrasts, whereas the Scheffé method would allow for any number of contrasts to be examined. From (10.4.5), the critical coefficients are

$$w_S = \sqrt{2F_{2,16,.05}} = \sqrt{2(3.63)} = 2.694 \text{ and } w_B = t_{16,.05/(2m)} = t_{16,.00625} \approx 2.783 \,.$$

Hence, in this particular situation, the Scheffé method gives slightly tighter intervals as well as being more flexible. An alternative is to divide $\alpha = 0.05$ between a t interval for the fourth contrast and Tukey intervals for the pairwise comparisons. For example, a 99% t interval and 96% Tukey intervals would have critical coefficients

$$w_t = t_{16,.01/2} = 2.9208 \text{ and } w_T = q_{3,16,.04}/\sqrt{2} = 3.8117/\sqrt{2} = 2.6953 \,,$$

and again the Scheffé method is prefable in this example.

For each pairwise comparison $\tau_i - \tau_p$, we have $\sum c_i^2 = 2$, so using the Scheffé method of multiple comparisons and $msE = 77217.7$ from Table 10.4, the interval becomes

$$\tau_i - \tau_p \in \left(\overline{y}_{.i} - \overline{y}_{.p} \pm 2.694\sqrt{(77217.7)(2)/9} \right) = \left(\overline{y}_{.i} - \overline{y}_{.p} \pm 352.89 \right) \,.$$

The treatment sample means are obtained from the data in Table 10.3 as

$$\overline{y}_{.1} = 7927.7, \quad \overline{y}_{.2} = 8015.2, \quad \overline{y}_{.3} = 7987.0 \,,$$

the biggest difference being $\overline{y}_{.2} - \overline{y}_{.1} = 87.5$. Since all three intervals contain zero, we can assert with 95% confidence that no two protocols differ significantly in their effects on the resting metabolic rate.

Similarly, the Scheffé confidence interval for $\tau_1 - \frac{1}{2}(\tau_2 + \tau_3)$ is

$$\tau_1 - \frac{1}{2}(\tau_2 + \tau_3) \in \left(\overline{y}_{.1} - \frac{1}{2}(\overline{y}_{.2} + \overline{y}_{.3}) \right) \pm 2.694\sqrt{(77217.7)(1.5)/9}$$
$$= (-73.44 \pm 305.62) \,,$$

and again the interval contains zero. These results are expected in light of the failure in Example 10.4.1 to reject equality of treatment effects in the analysis of variance.   □

10.5   A Real Experiment—Cotton-Spinning Experiment

10.5.1   Design Details

The checklist for the cotton-spinning experiment was given in Sect. 2.3, p. 13. After considering several different possible designs, the experimenters settled on a randomized complete block design. Each experimental unit was the production of one full set of bobbins on a single machine with a single operator. A block consisted of a group of experimental units with the same machine, the same operator, and observed in the same week. Thus, the different levels of the blocking factor represented differences due to combinations of machines, operators, environmental conditions, and raw material. The block

size was chosen to be six, as this was equal to the number of treatment combinations and also to the number of observations that could be taken on one machine in one week.

The treatment combinations were combinations of levels of two treatment factors, "flyer" and "degree of twist." Flyer had two levels, "ordinary" and "special." Twist had four levels, 1.63, 1.69, 1.78, and 1.90. For practical reasons, the combinations of flyer and twist equal to (ordinary, 1.63) and (special, 1.90) were not observed. We will recode the six treatment combinations that were observed as follows:

$$(\text{ordinary}, 1.69) = 1, \ (\text{ordinary}, 1.78) = 2, \ (\text{ordinary}, 1.90) = 3,$$
$$(\text{special}, 1.63) = 4, \quad (\text{special}, 1.69) = 5, \quad (\text{special}, 1.78) = 6.$$

The goal of the experiment was to investigate the effects of the flyers and degrees of twist on the breakage rate of cotton.

10.5.2 Sample-Size Calculation

Since the experimenters were interested in all pairwise comparisons of the effects of the treatment combinations, as well as some other special treatment contrasts, we will apply the Scheffé method of multiple comparisons at overall confidence level 95%. The experimenters initially wanted a confidence interval to indicate a difference in the effects of a pair of treatment combinations if the true difference was at least 2 breaks per 100 pounds of material. We will calculate the number of blocks that are needed to obtain a minimum significant difference of at most 2 for the Scheffé simultaneous confidence intervals for pairwise comparisons. Using (10.4.4) with $v = 6$, $\alpha = 0.05$, and $\Sigma c_i^2 = 2$, we need to find b such that

$$\sqrt{5 F_{5,5b-5,0.05}} \ \sqrt{msE\,(2/b)} \leq 2 \,.$$

The error variability σ^2 was expected to be about 7 breaks^2, so we need to find the smallest value of b satisfying

$$F_{5,5b-5,0.05} \leq \frac{4 \times b}{5 \times 7 \times 2} = \frac{2b}{35} \,.$$

Trial and error shows that $b = 40$ will suffice.

Each block took a week to complete, and it was not clear how many machines would be available at any one time, so the experimenters decided that they would analyze the data after the first 13 blocks had been observed. With $b = 13$, $v = 6$, and a value of σ^2 expected to be about 7 breaks^2, the Scheffé 95% confidence intervals for pairwise comparisons have minimum significant difference equal to

$$msd = \sqrt{5 F_{5,5(13-1),0.05}} \ \sqrt{7 \times (2/13)} \ \ = \ \ 3.57\,,$$

nearly twice the target length. Thus, with msE = 7 and 13 blocks, a difference in treatment combinations i and p will be indicated if their observed average difference is more than 3.57 breaks per 100 pounds (with a probability of 0.95 of no false indications) rather than 2 breaks per 100 pounds.

10.5.3 Analysis of the Cotton-Spinning Experiment

The data for the first $b = 13$ blocks observed in the experiment were shown in Table 2.3 (p. 16) and some of the data were plotted in Fig. 2.1. There is an indication of block differences over time. The low number of breaks tend to be in block 1, and the high number of breaks in blocks 11, 12, and 13. This suggests that blocking was worthwhile. This is also corroborated by the fact that $ms\theta$ is nearly three times as large as msE (see Table 10.5).

Table 10.5   Analysis of variance for the cotton-spinning experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Block	12	177.155	14.763	–	
Treatment	5	231.034	46.207	9.05	0.0001
Error	60	306.446	5.107		
Total	77	714.635			

The error assumptions for the block–treatment model (10.4.1) are satisfied apart from two outlying observations for treatment 1 (from blocks 5 and 10). The two outliers cause the variances of the unstandardized residuals to be unequal. Also, the normality assumption appears to be not quite satisfied. Since the experiment was run a long time ago, we are not able to investigate possible causes of the outliers. The best we can do is to run the analysis both with and without them. Here, we will continue the analysis including the outliers, and in Exercise 17, we ask the reader to verify that the model assumptions are approximately satisfied when the outliers are removed and that similar conclusions can be drawn.

The analysis of variance table is shown in Table 10.5. Luckily, the error variance is smaller than expected (the observed msE is 5.1), and consequently, the confidence intervals will not be as wide as feared. The null hypothesis of equality of the treatment effects is rejected at significance level $\alpha = 0.01$, since the p-value is less than 0.01; equivalently,

$$msT/msE = 9.05 > F_{5,60,.01} = 3.34 \,.$$

The treatment sample means are

i :	1	2	3	4	5	6
$\overline{y}_{.i}$:	10.8000	9.2769	7.1846	6.7538	7.0846	5.6538

With $b = 13$ blocks and $msE = 5.107$, the minimum significant difference for a set of Scheffé's simultaneous 95% confidence intervals is

$$msd = \sqrt{5F_{5,60,0.05}}\sqrt{msE\,\Sigma c_i^2/13} = \sqrt{5(2.37)}\sqrt{5.107\,\Sigma c_i^2/13}$$

$$= 3.442 \times 0.6268\sqrt{\Sigma c_i^2} = 2.158\sqrt{\Sigma c_i^2} \,. \tag{10.5.6}$$

For pairwise comparisons we have $\Sigma_i c_i^2 = 2$, so $msd = 3.052$. Comparing this value with differences in treatment sample means, we see that treatment 1 (ordinary flyer, 1.69 twist) yields significantly more breaks on average than all other treatment combinations except treatment 2 (ordinary flyer, 1.78 twist), and 2 is significantly worse on average than 6 (special flyer, 1.78 twist). This might lead one to suspect that the special flyer might be better than the ordinary flyer.

The contrast $\frac{1}{2}(\tau_1 + \tau_2) - \frac{1}{2}(\tau_5 + \tau_6)$ compares the two flyers, averaging over the *common levels* (1.69 and 1.78) of twist. The corresponding confidence interval (still using Scheffé's method at an overall 95% confidence level and msd (10.5.6)) is

$$\left(\frac{1}{2}(\overline{y}_{.1} + \overline{y}_{.2}) - \frac{1}{2}(\overline{y}_{.5} + \overline{y}_{.6}) \pm 2.158\sqrt{\Sigma c_i^2}\right) = (3.670 \pm (2.158 \times 1.0))$$

$$= (1.512, 5.828) \,.$$

Fig. 10.3 Mean number of breaks per 100 pounds for the cotton-spinning experiment
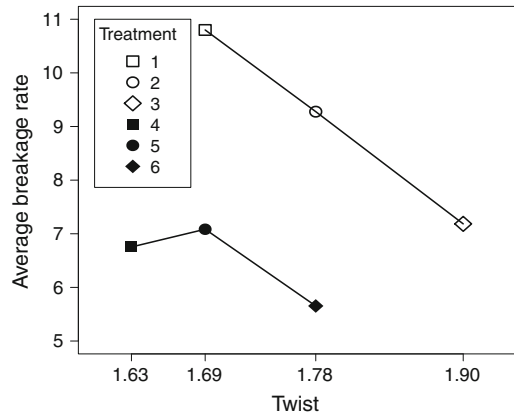
This confidence interval suggests that averaged over the middle two levels of twist, the ordinary flyer is worse than the special flyer, producing on average between about 1.5 and 5.8 more breaks per 100 pounds.

In Fig. 10.3, the treatment sample means $\overline{y}_{.i}$ are plotted against the uncoded twist levels, with the open symbols (labels 1, 2, 3) indicating those treatments with the ordinary flyer, and the black symbols (labels 4, 5, 6) indicating the special flyer. This plot reveals informative patterns in the treatment means. In particular, it appears as if the mean number of breaks per 100 pounds decreases almost linearly as the amount of twist increases for the ordinary flyer (treatments 1, 2, 3), yielding consistently smaller means for each amount of twist. Notice that the levels of twist are not equally spaced, so we cannot use the contrast coefficients in Appendix A.2 to measure trends in the breakage rate due to increasing twist. We could use the formula (4.2.4) p. 73 to obtain the linear trend coefficients, but since a different three of the four levels of twist are observed for the two flyers, the analysis of a linear trend would be more useful if done for each flyer separately. For example, for the three levels of the ordinary flyer, the coefficients for the linear trend would be calculated as

$$13 \times (x_i - 1.79) = -1.300, -0.130, 1.430, \text{ respectively, where} 1.79 = \overline{x} = \sum_{i=1}^{3} 13x_i/39,$$

and these become -10, -1, 11, respectively, when multiplied by the choice of constant 100/13. Using these integer coefficients, the estimate of the linear trend in the breakage rate due to increasing twist for the ordinary flyer is $\sum_{i=1}^{3} c_i \overline{y}_i = -38.246$ with corresponding estimated standard deviation

$$\sqrt{msE \sum_{i=1}^{3} c_i^2/r_i} = \sqrt{5.107 \times 17.0767} = 9.3387,$$

giving a ratio of $-38.246/9.3387 = -4.0954$. If we test, at level $\alpha = 0.01$, the hypothesis of no linear trend in the breakages due to increasing amounts of twist using the ordinary flyer against the alternative hypothesis that there is a decreasing linear trend, we would reject the null hypothesis in favor of the alternative since $-4.0954 < -t_{60, 0.01} = -2.390$.

Sections 10.9 and 10.10 illustrate the use of the SAS and R software for obtaining the analyses presented in this section, together with corresponding analyses using either the factorial main effects model or the analogous model treating twist as a linear regressor, and also the lack-of-fit test of the latter model.

10.6   Analysis of General Complete Block Designs

10.6.1   Model and Analysis of Variance

In this section we discuss general complete block designs with $s > 1$ observations on each treatment in each block. Having every level of the treatment factor observed more than once per block gives sufficient degrees of freedom to be able to measure a block×treatment interaction if one is anticipated. Therefore, there are two standard models for the general complete block design, the *block–treatment model* (without interaction)

$$Y_{hit} = \mu + \theta_h + \tau_i + \epsilon_{hit} \tag{10.6.7}$$

and the *block–treatment interaction model*, which includes the effect of block–treatment interaction:

$$Y_{hit} = \mu + \theta_h + \tau_i + (\theta\tau)_{hi} + \epsilon_{hit} \,. \tag{10.6.8}$$

In each case, the model includes the error assumptions

$$\epsilon_{hit} \sim N(0, \sigma^2) \,,$$
$$\epsilon_{hit}\text{'s are mutually independent}\,,$$
$$t = 1, \ldots, s \,;\ h = 1, \ldots, b \,;\ i = 1, \ldots, v.$$

The assumptions on these two models should be checked for any given experiment (see Sect. 10.7).

The block–treatment model (10.6.7) for a general complete block design is similar to the two-way main-effects model (6.2.3), and the block–treatment interaction model (10.6.8) is like the two-way complete model (6.2.2) for two treatment factors in a completely randomized design, each with s observations per cell. Analogously, the analysis of variance tables (Tables 10.6 and 10.7) for the block–treatment models, with and without interaction, look similar to those for the two-way main-effects and two-way complete models (Tables 6.4 and 6.7, pp. 159 and 170).

The decision rule for testing the null hypothesis $H_0^T : \{\tau_1 = \tau_2 = \cdots = \tau_v\}$ that the treatment effects are equal against the alternative hypothesis H_A^T that at least two of the treatment effects differ is given by the decision rule

$$\text{reject } H_0^T \text{ if } msT/msE > F_{v-1,df,\alpha} \,, \tag{10.6.9}$$

where α is the chosen significance level, and where msT, msE, and the error degrees of freedom, df, are obtained from Tables 10.6 or 10.7 as appropriate.

Table 10.6 Analysis of variance for the general complete block design with negligible block×treatment interaction and block size $k = vs$

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio
Block	$b - 1$	$ss\theta$	$-$	$-$
Treatment	$v - 1$	ssT	$msT = \frac{ssT}{v-1}$	$\frac{msT}{msE}$
Error	$bvs - b - v + 1$	ssE	$msE = \frac{ssE}{bvs-b-v+1}$	
Total	$bvs - 1$	$sstot$		
Computational formulae				
$ss\theta = vs \sum_h \overline{y}_{h..}^2 - bvs\,\overline{y}_{...}^2$		$sstot = \sum_h \sum_i \sum_t y_{hit}^2 - bvs\,\overline{y}_{...}^2$		
$ssT = bs \sum_i \overline{y}_{.i.}^2 - bvs\,\overline{y}_{...}^2$		$ssE = sstot - ss\theta - ssT$		

Table 10.7 Analysis of variance for the general complete block design with block×treatment interaction and block size $k = vs$

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio
Block	$b - 1$	$ss\theta$	–	–
Treatment	$v - 1$	ssT	$msT = \frac{ssT}{(v-1)}$	$\frac{msT}{msE}$
Interaction	$(b-1)(v-1)$	$ss\theta T$	$ms\theta T = \frac{ss\theta T}{(b-1)(v-1)}$	$\frac{ms\theta T}{msE}$
Error	$bv(s-1)$	ssE	$msE = \frac{ssE}{bv(s-1)}$	
Total	$bvs - 1$	$sstot$		

Computational formulae

$ss\theta = vs \sum_h \bar{y}_{h..}^2 - bvs\,\bar{y}_{...}^2$

$ssT = bs \sum_i \bar{y}_{.i.}^2 - bvs\,\bar{y}_{...}^2$

$ssE = sstot - ss\theta - ssT - ss\theta T$

$ss\theta T = s \sum_h \sum_i \bar{y}_{hi.}^2 - bs \sum_i \bar{y}_{.i.}^2$
$\qquad\quad - vs \sum_h \bar{y}_{h..}^2 + bvs\,\bar{y}_{...}^2$

$sstot = \sum_h \sum_i \sum_t y_{hit}^2 - bvs\,\bar{y}_{...}^2$

Table 10.8 Data for the DCIS weighing system

Block	Time, position (treatment)					
	11 (1)	12 (2)	21 (3)	22 (4)	31 (5)	32 (6)
1	0.637	0.174	0.886	0.378	0.396	0.386
	0.645	0.238	0.655	0.459	0.415	0.453
2	0.675	0.187	0.528	0.270	0.594	0.799
	0.480	0.183	0.701	0.426	0.545	0.413

If the block×treatment interaction term is included in the model, a test of the hypothesis $H_0^{\theta T}$: $\{(\theta\tau)_{hi} - \overline{(\theta\tau)}_{h.} - \overline{(\theta\tau)}_{.i} + \overline{(\theta\tau)}_{..} = 0$ for all $h, i\}$ against the alternative hypothesis $H_A^{\theta T}$ that at least one interaction contrast is nonzero is given by

$$\text{reject } H_0^{\theta T} \text{ if } ms\theta T / msE > F_{(b-1)(v-1),\,bv(s-1),\,\alpha} \tag{10.6.10}$$

for some chosen significance level α, where $ms\theta T$ and msE are obtained from Table 10.7. As usual, if the interaction is significantly different from zero, a test of equality of the treatment effects may not be of interest (unless done within each block separately). An evaluation of the usefulness of blocking in the experiment at hand can be made by comparing $ms\theta$ with msE as in Sect. 10.4.1.

Example 10.6.1 DCIS experiment, continued

The objective of the DCIS experiment, described in Example 10.3.2, was to reduce the variability in the DCIS weighing system. The setting currently employed was treatment 4 (time 75 milliseconds, and switch at 2 inches from the scale plate). The experiment was run as a general block design with $s = 2$ observations per treatment per block (the blocking factor levels were the amounts of conveyor pan lubrication). The responses ("uncertainty" calculated as a function of the standard deviation of 30 repeated weighings of a watermelon) are shown in Table 10.8 and plotted in Fig. 10.4.

A block-treatment interaction model (10.6.8) was fitted, so the block-treatment interaction is examined first. Figure 10.4 suggests that there might be a small interaction between block and treatment combination.

The analysis of variance table is shown in Table 10.9, and we see that $ms\theta T = 0.019$ and $msE = 0.013$. Using (10.6.10), the hypothesis $H_0^{\theta T}$ of negligible interaction would be rejected
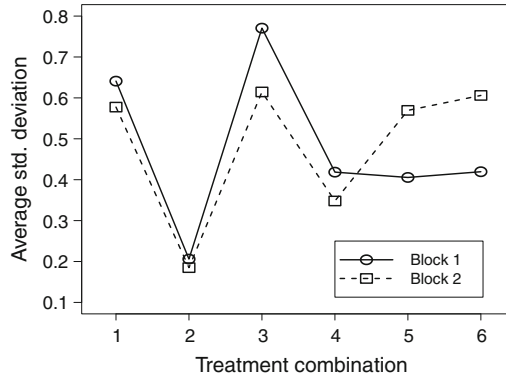
Fig. 10.4   Response for the DCIS experiment

Table 10.9   Analysis of variance for the DCIS experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Block	1	0.0003	–	–	
Treatment	5	0.6132	0.1226	9.41	0.0008
Block $\times$ Treatment	5	0.0952	0.0190	1.46	0.2726
Error	12	0.1563	0.0130		
Total	23	0.8649			

if $ms\theta T/msE = 1.46$ is larger than $F_{5,12,\alpha}$ for some chosen significance level α. However, $F_{5,12,.01} = 9.89$, so there is not sufficient evidence to reject $H_0^{\theta T}$ at level $\alpha = 0.01$. Notice that the p-value for the test is 0.2726, so in fact, no reasonable choice of α would lead to rejection of $H_0^{\theta T}$. So the block×treatment interaction that appears in Fig. 10.4 could be due to error variability. From Table 10.9, $msT/msE = 9.41 > F_{5,12,0.01} = 5.06$, and we conclude that, averaged over the two blocks (levels of pan lubrication), there is a significant difference between the treatment combinations at level $\alpha = 0.01$. Figure 10.4 suggests that treatment 2 might be the best treatment in both blocks (and better than the current treatment 4). The overall significance level of the two tests is at most 0.02.   □

10.6.2   Multiple Comparisons for the General Complete Block Design

No Interaction Term in the Model

The Bonferroni, Scheffé, Tukey, and Dunnett methods described in Sect. 4.4 can all be used for obtaining simultaneous confidence intervals for sets of treatment contrasts in a general complete block design. Since the block–treatment model (10.6.7), without interaction, is similar to the two-way main-effects model (6.2.3) with s observations per cell, formulae for multiple comparisons are similar to those given in (6.5.39), p. 166, with a replaced by v and r replaced by s. Thus, a set of $100(1-\alpha)\%$ simultaneous confidence intervals for treatment contrast $\Sigma c_i \tau_i$ is of the form

$$\sum c_i \tau_i \quad \in \quad \left(\sum c_i \bar{y}_{.i.} \pm w \sqrt{msE \sum c_i^2/bs} \right), \qquad (10.6.11)$$

where the critical coefficients for the four methods are, respectively,

$$w_B = t_{df,\alpha/2m} \; ; \;\; w_S = \sqrt{(v-1)F_{v-1,df,\alpha}} \; ;$$
$$w_T = q_{v,df,\alpha}/\sqrt{2} \; ; \;\; w_{D2} = |t|_{v-1,df,\alpha}^{(0.5)} \; ,$$

where $n = bvs$ and $df = n - b - v + 1$.

Interaction Term Included in the Model

The block–treatment interaction model (10.6.8) for the general complete block design is similar to the two-way complete model (6.2.2), p. 142, for two treatment factors with s observations per cell. Consequently, formulae for confidence intervals for treatment comparisons, averaging over the block×treatment interaction, are similar to those given in (6.4.18), p. 152, with a replaced by v and r replaced by s. The general formula for a set of $100(1 - \alpha)\%$ simultaneous confidence intervals for treatment contrasts is given by (10.6.11) above, but where the number of error degrees of freedom df in each of the critical coefficients is

$$df = (vbs - 1) - (b-1)(v-1) - (b-1) - (v-1) = bv(s-1) = n - bv \, .$$

Treatment comparisons may not be of interest if treatments do interact with blocks, in which case within-block comparisons are likely to be preferred instead. These are similar to the simple contrasts of Sect. 6.3.1 and are most easily calculated via a cell-means representation of the model. If we write

$$\eta_{hi} = \theta_h + \tau_i + (\theta\tau)_{hi} \, , \tag{10.6.12}$$

then the comparison of treatments i and p in block h is the contrast

$$\eta_{hi} - \eta_{hp} = (\theta_h + \tau_i + (\theta\tau)_{hi}) - (\theta_h + \tau_p + (\theta\tau)_{hp}) \, .$$

In a general complete block design, there are equal numbers of observations on each treatment in each block so, for block h, the least squares estimate of $\sum_i c_{hi}\eta_{hi}$ is $\sum_i c_{hi}\overline{y}_{hi.}$ with corresponding variance $msE \sum_i c_{hi}^2/s$, giving confidence intervals for $\sum_i c_{hi}\eta_{hi}$ of the form

$$\sum_{i=1}^{v} c_{hi}\eta_{hi} \in \left(\sum_{i=1}^{v} c_{hi}\overline{y}_{hi.} \pm \sqrt{msE \sum_{i=1}^{v} c_{hi}^2/s} \right) \, .$$

The critical coefficients are as for (10.6.11) but again with $df = n - bv$ error degrees of freedom.

Example 10.6.2 `DCIS experiment, continued`

In the analysis of the DCIS experiment in Example 10.6.1, the block-treatment interaction appeared to be negligible, and so the experimenters had the choice of examining the treatment effects on the weight "uncertainty" averaged over blocks or examining the difference in the treatment effects for each block separately. The former would be of most interest if it is not possible to control the level of conveyor pan lubrication in an industrial setting, while the latter would be of interest if the amount of lubrication could be fixed. Here, as an example, we investigate the latter. Writing the effect of treatment i in block h as in (10.6.12), we look at the difference between the effects of treatments 2 (apparent best) and 4 (currently used) for each block separately, which compares times 50 and 75 milliseconds at position 2

for each pan lubrication level. We also look at the differences in the effect of switch position averaged over weighing time, again for each block separately; that is,

$$\eta_{h2} - \eta_{h4} \quad \text{and} \quad \frac{1}{3}(\eta_{h1} + \eta_{h3} + \eta_{h5}) - \frac{1}{3}(\eta_{h2} + \eta_{h4} + \eta_{h6}) \quad \text{for} \quad h = 1, 2.$$

These have least squares estimates and standard errors as follows:

contrast	$\Sigma c_{hi}\bar{y}_{hi.}$	$\sqrt{msE \ \Sigma c_{hi}^2/s}$
block 1, treatment 4-2	0.2125	0.1141
block 2, treatment 4-2	0.1630	0.1141
block 1, switch 1in vs 2in	0.2577	0.0659
block 2, switch 1in vs 2in	0.2075	0.0659

Using Bonferroni's method at overall level at least 96%, the four confidence intervals are

$$\eta_{14} - \eta_{12} \in (0.2125 \pm 0.1141 t_{.005,12}) = (0.2125 \pm 0.1141 \times 3.055) = (-.136, .561)$$
$$\eta_{24} - \eta_{22} \in (0.1630 \pm 0.1141 t_{.005,12}) = (0.1630 \pm 0.1141 \times 3.055) = (-.186, .512)$$
$$\frac{1}{3}(\eta_{11} + \eta_{13} + \eta_{15}) - \frac{1}{3}(\eta_{12} + \eta_{14} + \eta_{16}) \in (0.2577 \pm 0.0659 \times 3.055) = (.056, .459)$$
$$\frac{1}{3}(\eta_{21} + \eta_{23} + \eta_{25}) - \frac{1}{3}(\eta_{22} + \eta_{24} + \eta_{26}) \in (0.2075 \pm 0.0659 \times 3.055) = (.006, .409)$$

Thus at overall confidence level of at least 96%, we cannot detect a difference between the effects of treatment 2 and the current treatment 4 in either block. However, switch position 2 inches from the end of the scale plate (averaged over weighing times) reduces "uncertainty" as compared with switch position 1 inch by up to .4 in both blocks, and hence seems to be the better position. □

10.6.3 Sample-Size Calculations

A complete block design has $n = bvs$ experimental units divided into b blocks of size $k = vs$. The block size k and the number of blocks b must be chosen to accommodate the experimental conditions, the budget constraints, and the requirements on the lengths of confidence intervals or powers of hypothesis tests in the usual way. If the number of blocks is fixed, one can calculate the number of observations s per treatment per block needed to achieve a prescribed power of a test of no treatment differences. Analogous to the sample-size calculation for testing main effects of a factor in a two-way layout, s must satisfy

$$s \geq \frac{2v\sigma^2\phi^2}{b\Delta^2}, \tag{10.6.13}$$

where Δ is the minimum difference between the treatment effects that is to be detected. The tables for power $\pi(\Delta)$ as a function of ϕ are in Appendix Table A.7. In (10.6.13), we can switch the role of s and b, and instead, calculate the number of blocks needed if the block size is fixed.

For calculation of the sample size needed to achieve confidence intervals of specified length, we specify the maximum allowed msd in (10.6.11) and solve for s or b.

An example of the calculation of b to achieve confidence intervals of given length was given for the randomized complete block design in Sect. 10.5.2. In Example 10.6.3, we calculate, for a general complete block design, the block size required to achieve a given power of a hypothesis test.

Example 10.6.3 `Colorfastness experiment`

The colorfastness experiment was planned by D-Y Duan, H. Rhee, and C. Song in 1990 to investigate the effects of the number of washes on the color change of a denim fabric. The experiment was to be carried out according to the guidelines of the American Association of Textile Chemists and Colorists Test 61-1980. The levels of the treatment factor were the number of times of laundering, and these were selected to be 1, 2, 3, 4, and 5.

The experimenters anticipated that there would be systematic differences in the way they made their color determinations, and consequently, they grouped the denim fabric swatches into blocks according to which experimenter was to make the determination. Thus the levels of the blocking factor denoted the experimenter, and there were $b = 3$ blocks. They decided to use a general complete block design and allowed the block size to be $k = vs = 5s$, where s could be chosen. Rightly or wrongly, they did not believe that experimenter fatigue would have a large effect on the results, and they were happy for the block sizes to be large.

They planned to use a block–treatment interaction model (10.6.8), and they wanted to test the null hypothesis of no treatment differences whether or not there was block×treatment interaction. Suppose the test was to be carried out at significance level 0.05, and suppose the experimenters wanted to reject the null hypothesis with probability 0.99 if there was a true difference of $\Delta = 0.5$ or more in the effect of the number of washes on color rating. They expected σ to be no larger than about 0.4.

We need to find the minimum value of s that satisfies equation (10.6.13); that is,

$$ s \geq \frac{2v\sigma^2\phi^2}{b\Delta^2} = \frac{(2)(5)(0.4)^2\phi^2}{(3)(0.5)^2} = 2.13\phi^2 \,. $$

The denominator (error) degrees of freedom for the block–treatment interaction model is $\nu_2 = bv(s-1) = 15(s-1)$. First we locate that portion of Appendix Table A.7 corresponding to numerator degrees of freedom $\nu_1 = v - 1 = 4$ and $\alpha = 0.05$. Then to achieve power $\pi = 0.99$, trial and error starting with $s = 100$ gives

s	$15(s-1)$	ϕ	$s = 2.13\phi^2$	Action
100	1485	2.25	10.78	Round up to $s = 11$
11	150 (use 120)	2.325	11.51	Round up to $s = 12$
12	165 (use 120)	2.325	11.51	Stop, and use $s = 12$

So about $s = 12$ observations per treatment per block should be taken.

Instead, suppose that s had been fixed at $s = 4$, so that blocks were to be of size $k = vs = 20$, then the roles of b and s in (10.6.13) would have been reversed, so that

$$ b \geq \frac{(2)(5)(.4)^2\phi^2}{(4)(.5)^2} = 1.6\phi^2 \,, $$

with $\nu_2 = (5)(3)b$. Then trial and error would lead to approximately $b = 9$.   □

10.7   Checking Model Assumptions

The assumptions on the block–treatment models (10.4.1) and (10.6.7) and on the block–treatment interaction model (10.6.8) for complete block designs need to be checked as usual. The assumptions on the error variables are that they have equal variances, are independent, and have a normal distribution. The form of the model must also be checked.

Table 10.10   Checking error assumptions for a complete block design

To check for:	Plot residuals against:
Independence	Order of observations (in space or time)
Equal variance, and Outliers	Predicted values $\hat{y}_{hit}$, levels of treatment factor, levels of block factor
Normality	Normal scores (also plot separately for each treatment if r is large and for each block if k is large)

A visual check of an assumption of no block×treatment interaction can be made by plotting $\overline{y}_{hi.}$ against the treatment factor levels i for each block h in turn. If the lines plotted for each block are parallel (as in plots (a)–(d) of Fig. 6.1, p. 140), then block×treatment interaction is likely to be absent, and error variability is small. If the lines are not parallel, then either block×treatment interaction is present or error variability is large.

For the block–treatment model (10.4.1) for the randomized complete block design, the (hi)th residual is

$$\hat{e}_{hi} = y_{hi} - \hat{y}_{hi} \;\; = \;\; y_{hi} - \overline{y}_{h.} - \overline{y}_{.i} + \overline{y}_{..} \, .$$

For the block–treatment model (10.6.7) for the general complete block design, the (hit)th residual is similar; that is,

$$\hat{e}_{hit} = y_{hit} - \hat{y}_{hit} \;\; = \;\; y_{hit} - \overline{y}_{h..} - \overline{y}_{.i.} + \overline{y}_{...} \, .$$

For the block–treatment interaction model (10.6.8), the (hit)th residual is

$$\hat{e}_{hit} = y_{hit} - \hat{y}_{hit} \;\; = \;\; y_{hit} - \overline{y}_{hi.} \, .$$

The error assumptions are checked by residual plots, as summarized in Table 10.10 and described in Chap. 5.

10.8   Factorial Experiments

When the treatments are factorial in nature, the treatment parameter τ_i in the complete block design models (10.4.1), (10.6.7), and (10.6.8) can be replaced by main-effect and interaction parameters. Suppose, for example, we have an experiment with two treatment factors that is designed as a randomized complete block design—a situation similar to that of the cotton-spinning experiment of Sect. 10.5. In order not to confuse the number b of blocks with the number of levels of a treatment factor, we will label the two treatment factors as C and D with c and d levels respectively. If we retain the two digit codes for the treatment combinations, then the block–treatment model is

$$Y_{hijt} = \mu + \theta_h + \tau_{ij} + \epsilon_{hijt}$$

with the usual assumptions on the error variables. We can then express τ_{ij}, the effect of treatment combination ij, in terms of γ_i (the effect of C at level i), δ_j (the effect of D at level j), and $(\gamma\delta)_{ij}$ (the effect of their interaction when C is at level i and D is at level j); that is,

$$Y_{hijt} = \mu + \theta_h + \gamma_i + \delta_j + (\gamma\delta)_{ij} + \epsilon_{hijt} \, . \tag{10.8.14}$$

In a general complete block design with $s > 1$ observations per treatment combination per block, we may include in the model some or all of the block×treatment interactions. For example, with two treatment factors, the block–treatment interaction model can be expressed as

$$Y_{hijt} = \mu + \theta_h + \gamma_i + \delta_j + (\gamma\delta)_{ij} + (\theta\gamma)_{hi} \qquad (10.8.15)$$
$$+ (\theta\delta)_{hj} + (\theta\gamma\delta)_{hij} + \epsilon_{hijt},$$

In both (10.8.14) and (10.8.15), the model assumptions are

$$\epsilon_{hijt} \sim N(0, \sigma^2),$$
$$\epsilon_{hijt}\text{'s are mutually independent},$$
$$t = 1, \ldots, s ; \ h = 1, \ldots, b ; \ i = 1, \ldots, c ; \ j = 1, \ldots, d.$$

If there are more than two factors, the additional main effects and interactions can be added to the model in the obvious way.

Example 10.8.1 `Banana experiment`

The objectives section of the report of an experiment run in 1995 by K. Collins, D. Marriott, P. Kobrin, G. Kennedy, and S. Kini reads as follows:

> Recently a banana hanging device has been introduced in stores with the purpose of providing a place where bananas can be stored in order to slow the ripening process, thereby allowing a longer time over which the consumer has to ingest them. Commercially, bananas are picked from trees while they are fully developed but quite green and are artificially ripened prior to transport. Once they are purchased and brought into the consumer's home, they are typically placed on a counter top and left there until they are either eaten or turn black, after which they can be thrown away or made into banana bread. Considering that the devices currently being marketed to hang bananas cost some money and take up counter space, it is of interest to us to determine whether or not they retard the ripening process.

> While there exist many ways to measure the degree of banana ripening, perhaps the simplest method is via visual inspection. The banana undergoes a predictable transition from the unripened green color to yellow then to yellow speckled with black and finally to fully black. The percentage of black color can be quantified through computer analysis of photographs of the skins of the bananas.

> The major objective of our experiment, then, is to determine whether or not any differences in the percentage of black skin exist between bananas that are treated conventionally, i.e., placed on a counter, and bananas that are hung up. As a minor objective, we would like to determine whether or not any difference exists in the percentage of black skin between bananas allowed to ripen in a normal day/night cycle versus those ripening in the dark such as might occur if placed in a pantry.

The unripened bananas were bought as a single batch from a single store. They were assigned at random to four treatment combinations, consisting of combinations of two 2-level factors. Factor C was Lighting conditions (1 = day/night cycle, 2 = dark closet). Factor D was Storage method (1 = hanging, 2 = counter-top). Twelve bananas were assigned at random to each treatment combination. After five days, the bananas were peeled and the skin photographed. The images from the photographic slides were traced by hand, and the percentage of blackened skin was calculated using an image analyzer on a computer. Three of the experimenters prepared the images for the image analyzer and, since they were unskilled, they decided to regard themselves as blocks in order to remove experimenter differences from the comparisons of the treatment combinations. They selected a general complete block design and assigned the treated bananas in such a way that $s = 4$ observations on each treatment combination were obtained by each experimenter. The treatment combinations were observed in a random order, and the resulting data are shown in Table 10.11.

Since the experimenters did not anticipate a block×treatment interaction, they selected block–treatment model (10.8.14) to represent the data. The decision rule for testing the hypothesis H_0^{CD} of no interaction between the treatment factors Light and Storage (averaged over blocks), using a Type I

Table 10.11 Percentage blackened banana skin

Experimenter (Block)	Light C	Storage D	y_{hijt}, percentage of blackened skin			
I	1	1	30	30	17	43
	1	2	43	35	36	64
	2	1	37	38	23	53
	2	2	22	35	30	38
II	1	1	49	60	41	61
	1	2	57	46	31	34
	2	1	20	63	64	34
	2	2	40	47	62	42
III	1	1	21	45	38	39
	1	2	42	13	21	26
	2	1	41	74	24	51
	2	2	38	22	31	55

Table 10.12 Analysis of variance for the banana experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Block (Experimenter)	2	1255.79	627.89	–	
C (Light)	1	80.08	80.08	0.42	0.5218
D (Storage)	1	154.08	154.08	0.80	0.3754
CD	1	24.08	24.08	0.13	0.7250
Error	42	8061.88	191.95		
Total	47	9575.92			

error probability of $\alpha = 0.01$, is

$$\text{reject } H_0^{CD} \text{ if } \frac{ms(CD)}{msE} > F_{(c-1)(d-1), df, 0.01} \,,$$

where $ms(CD) = ss(CD)/(c-1)(d-1)$ and both $ss(CD)$ and the number of error degrees of freedom df are given in Table 10.12. Since there are equal numbers of observations per cell, these values can be obtained from rule 4 of Chap. 7, p. 209; that is,

$$ss(CD) = bs \sum_i \sum_j \overline{y}_{.ij.}^2 - bds \sum_i \overline{y}_{.i..}^2 - bcs \sum_j \overline{y}_{..j.}^2 + bcds\, \overline{y}_{....}^2 = 24.0833,$$

and

$$df = (bcds - 1) - (b - 1) - (c - 1) - (d - 1) - (c - 1)(d - 1)$$
$$= 47 - 2 - 1 - 1 - 1 = 42 \,.$$

Other sums of squares are obtained similarly. From Table 10.12, we can see that the mean square for blocks is much larger than the error mean square, so it was worthwhile designing this experiment as a block design. We also see that the mean square for the Light×Storage interaction is a lot smaller than the error mean square. As mentioned in the context of the resting metabolic rate experiment (Example 10.4.1, p. 311), this is unusual when the model fits well since the Light and Storage
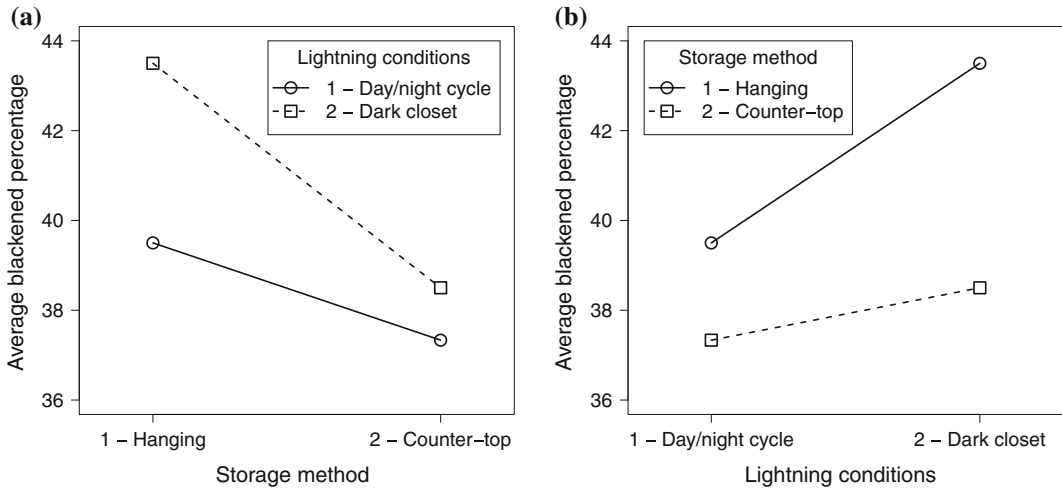
Fig. 10.5 Interaction plot for the banana experiment

measurements include the error measurement. It suggests that the error mean square may have been inflated by some other source of variability, such as block×treatment interaction, that has been omitted from the model.

Interaction plots of the two factors Light and Storage (averaged over blocks) are shown in Fig. 10.5. There is no indication that hanging bananas (Storage level 1) might retard the ripening process. In fact, Storage level 1 seems to have given a higher percentage of blackened skin on average than Storage level 2. However, this apparent difference may be due to chance, as the treatment effects are not significantly different from each other. The experimenters commented that it was difficult to select the correct threshold levels for the image analysis and also that the bananas themselves seemed extremely variable. The experimenters felt that rather than draw firm conclusions at this stage, it might be worthwhile working to improve the experimental procedure to reduce variability and then to repeat the experiment.   □

10.9 Using SAS Software

The analysis of variance table for a complete block design can be obtained from any computer package that has an analysis of variance routine or a regression routine. It is good practice to enter the block term into the model before the terms for the treatment factors. Although the order does not matter for complete block designs, it does matter for the incomplete block designs in the next chapter.

Computer programs do not distinguish between block and treatment factors, so a test for the hypothesis of no block effects will generally be listed in the output. We suggest that the latter be ignored, and that blocking be considered to have been effective for the current experiment if $ms\theta$ exceeds msE (see Sect. 10.4.1).

Table 10.13 contains a SAS program illustrating analysis of a complete block design, using the data of the cotton-spinning experiment (Sect. 2.3, p. 13). Following input of the data, the first call of PROC GLM fits a block–treatment model to the data. Selected output is shown in Fig. 10.6. The TYPE I sums of squares (not shown) are equal to the TYPE III sums of squares since there is an equal number ($s = 1$) of observations per block–treatment combination. If the block×treatment interaction term had been included in the model, this would have been entered in the usual way as BLOCK*TRTMT. The first LSMEANS statement in Table 10.13 applies Tukey's method for comparing pairs of treatment

Table 10.13 A SAS program for analysis of the cotton-spinning experiment

```
DATA COTTON;
  INPUT BLOCK TRTMT FLYER TWIST BREAK;
  LINES;
   1  12 1 1.69  6.0
   2  12 1 1.69  9.7
   :   : :  :      :
  13  23 2 1.78  6.4
;
* Block-treatment model for a complete block design;
PROC GLM;
  CLASS BLOCK TRTMT;
  MODEL BREAK = BLOCK TRTMT;
  LSMEANS TRTMT / PDIFF = ALL CL ADJUST = TUKEY ALPHA = 0.05;
  ESTIMATE 'FLYER 1-2 COMN TWST' TRTMT 0.5 0.5 0 0 -0.5 -0.5;
  ESTIMATE 'LIN TWIST ORD FLYER' TRTMT -10 -1 11 0 0 0;
* Factorial main effects model plus blocks;
PROC GLM;
  CLASS BLOCK FLYER TWIST;
  MODEL BREAK = BLOCK FLYER TWIST;
  LSMEANS FLYER / PDIFF CL ALPHA=0.05;
  LSMEANS TWIST / PDIFF = ALL CL ADJUST = BON ALPHA = 0.05;
* Model with twist as a linear regressor variable;
PROC GLM;
  CLASS BLOCK FLYER;
  MODEL BREAK = BLOCK FLYER TWIST / SOLUTION;
  ESTIMATE 'FLYER 1-2'  FLYER 1 -1;
* Testing lack of fit of reduced model;
PROC GLM;
  CLASS BLOCK FLYER TRTMT;
  MODEL BREAK = BLOCK FLYER TWIST TRTMT;
```

Fig. 10.6 SAS selected output for the block–treatment model—cotton-spinning experiment

The GLM Procedure
Dependent Variable: BREAK

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	408.1892308	24.0111312	4.70	<.0001
Error	60	306.4461538	5.1074359		
Corrected Total	77	714.6353846			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BLOCK	12	177.1553846	14.7629487	2.89	0.0033
TRTMT	5	231.0338462	46.2067692	9.05	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
FLYER 1-2 COMN TWST	3.6692308	0.62680115	5.85	<.0001
LIN TWIST ORD FLYER	-38.2461538	9.33912683	-4.10	0.0001

Fig. 10.7 SAS selected output for the factorial main-effects model—cotton-spinning experiment

effects (output not shown). The two ESTIMATE statements reproduce the calculations in Sect. 10.5 for comparing the two flyers at common levels of twist, and evaluating the linear trend in twist for the ordinary flyer.

The second call of PROC GLM in Table 10.13 replaces the treatment combination factor TRTMT with main-effects of FLYER and TWIST; this removes their interaction effect from the model. Selected output is shown in Fig. 10.7. Since not every combination of FLYER and TWIST was observed, the TYPE I and TYPE III sums of squares for the individual factors are not equal. The TYPE III sums of squares are used for hypothesis testing, and LSMEANS statements are used for confidence intervals. The reader may verify that the Bonferroni method provides tighter simultaneous 95% confidence intervals for TWIST pairwise comparisons than Scheffé's method (results not shown).

The plot of the mean response against twist for both flyer types in Fig. 10.3, p. 317, suggested the possibility that the number of breaks per 100 pounds could be modeled by a flyer effect and a linear twist effect. This can be evaluated by comparing the fit of the block–treatment model,

$$Y_{hi} = \mu + \theta_h + \tau_i + \epsilon_{hi}$$

$(i = 1, \ldots, 6; \ h = 1, \ldots, 13)$, with the fit of the reduced model,

$$Y_{hjx} = \mu + \theta_h + \alpha_j + \gamma x + \epsilon_{hjx},$$

where α_j is the effect of flyer j ($j = 1, 2$) and x is the *uncoded* amount of twist (cf. Chaps. 8 and 9). The fit of these models can easily be compared via the SAS software, either using two calls of the GLM procedure, one for each model, or using one call that sequentially includes both models. The first call of PROC GLM in Table 10.13 fitted the full block–treatment model, and the third call of PROC GLM fits the reduced model that includes the flyer effect and a linear regression in the levels of twist. Notice the similarity of the third call with the factorial main-effects model in the second call. The difference is that when TWIST is to be regarded as a linear regressor, it is omitted from the CLASS statement. The reduced model fits parallel linear regression lines, with intercepts adjusted for block and flyer effects. Selected output for the reduced model is shown in Fig. 10.8. Again, the TYPE I and TYPE III sums of squares are unequal, indicating that FLYER and TWIST cannot be estimated independently. The fourth call of PROC GLM sequentially includes both models and will be discussed shortly.

In the third call of PROC GLM in Table 10.13, the SOLUTION option requests that the solution to the normal equations is printed. The NOTE at the bottom of the SAS output in Fig. 10.8 alerts us to

Fig. 10.8 SAS program output for the reduced model—cotton-spinning experiment
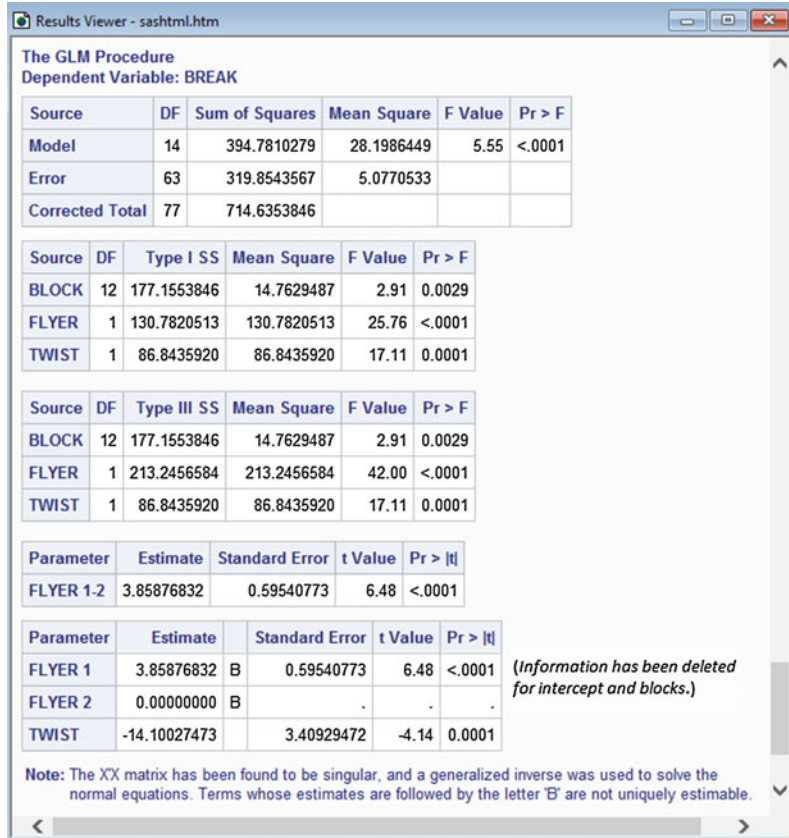
the fact that the individual flyer effect parameters are not estimable, and the numbers given just above the note and labeled B are nonunique solutions to the normal equations. The contrast representing the difference in the effects of the two flyers *is* estimable, and we can obtain its unique least squares estimate by taking the difference in the two values given for the individual flyers. This gives 3.8587, which matches the value obtained from the ESTIMATE statement. The difference in the effects of the two flyers is declared to be significantly different from zero, since the corresponding *p*-value is at most 0.0001. The slope coefficient of TWIST is estimated to be -14.1003, which, being negative, suggests that the breakages decrease as the twist increases. This slope is declared to be significantly different from zero, since the test of $H_0 : \gamma = 0$ versus $H_A : \gamma \neq 0$ has *p*-value at most 0.0001.

To test for lack of fit of the reduced model, the difference in the error sum of squares for the full and reduced models divided by the difference in the error degrees of freedom is the mean square for lack of fit, *msLF* (see Chap. 8). It provides the numerator of the test statistic for testing the null hypothesis H_0^R : {the reduced model is adequate} against the alternative hypothesis that the reduced model is not adequate. The decision rule is

$$\text{reject } H_0^R \text{ if } msLF/msE > F_{3,60,\alpha},$$

where

$$msLF = [ssE(\text{reduced}) - ssE(\text{full})] / [df(\text{reduced}) - df(\text{full})] .$$

For the cotton-spinning experiment,

$$msLF = (319.854 - 306.446)/(63 - 60) = 4.469.$$

Since $msLF/msE = 4.469/5.10744 = 0.88 < 1$, we cannot reject H_0^R for any reasonable significance level α. Hence, the reduced model appears to provide an adequate fit to the data, making interpretation of the parameters in the reduced model meaningful.

One can obtain these calculations from SAS software directly. The model

```
MODEL BREAK = BLOCK FLYER TWIST TRTMT;
```

in the fourth call of GLM in Table 10.13 sequentially includes both models, as the first three terms of the model give the reduced model, then the full model is obtained by inclusion of TRTMT. Since TRTMT has been entered into the model last, its Type I and III sums of squares will be the same, and the corresponding F-test

```
Source    DF    Type III SS    Mean Square   F Value   Pr > F
TRTMT      3     13.4082028      4.4694009      0.88    0.4591
```

is the test for lack of fit conducted above.

We note that this has been an exercise in model-building, and we can use the model to predict the breakage rates with either flyer over a range of values of twist. However, the model may not fit well for flyer 2 below a twist of 1.69 (see Fig. 10.3).

10.10   Using R Software

The analysis of variance table for a complete block design can be obtained from any computer package that has an analysis of variance routine or a regression routine. It is good practice to enter the block term into the model before the terms for the treatment factors. Although the order does not matter for complete block designs, it does matter for the incomplete block designs in the next chapter.

Computer programs do not distinguish between block and treatment factors, so a test for the hypothesis of no block effects will generally be listed in the output. We suggest that the latter be ignored, and that blocking be considered to have been effective for the current experiment if $ms\theta$ exceeds msE (see Sect. 10.4.1).

Tables 10.14, 10.15, 10.16 and 10.17 contain statements of an R program and selected output illustrating analysis of a complete block design, using the data of the cotton-spinning experiment (Sect. 2.3, p. 13). In Table 10.14, following input of the data from the file cotton.spinning.txt and the creation of factor variables fBlock, fTrtmt, fFlyer, and fTwist, the call of the linear models function lm fits a block–treatment model to the data. If the block×treatment interaction term had been included in the model, this would have been entered in the usual way as fBlock:fTrtmt. The anova statement generates the analysis of variance table shown. A drop1 statement (see Sects. 6.9 and 7.7) would have produced the same results since there is one observation per block–treatment combination. The lsmeans function and first corresponding summary(contrast…) statement applies Tukey's method for comparing the treatment effects pairwise (results not shown). The second corresponding summary(contrast…) statement reproduces the calculations in Sect. 10.5 for comparing

Table 10.14 An R program and selected output for analysis of the cotton-spinning experiment: data input and analysis of block–treatment model

```
> cotton.data = read.table("data/cotton.spinning.txt", header=T)
> head(cotton.data, 3)
  Block Trtmt Flyer Twist Break
1    1    12     1  1.69   6.0
2    2    12     1  1.69   9.7
3    3    12     1  1.69   7.4
> tail(cotton.data, 3)
  Block Trtmt Flyer Twist Break
78   13    23     2  1.78   6.4

> cotton.data = within(cotton.data,
+                      {fBlock = factor(Block); fTrtmt = factor(Trtmt);
+                       fFlyer = factor(Flyer); fTwist = factor(Twist) })

> # Analysis for randomized complete block design
> model1 = lm(Break ~ fBlock + fTrtmt, data=cotton.data)
> anova(model1)

Analysis of Variance Table

Response: Break
          Df Sum Sq Mean Sq F value    Pr(>F)
fBlock    12    177    14.8    2.89    0.0033
fTrtmt     5    231    46.2    9.05 0.0000019
Residuals 60    306     5.1

> library(lsmeans)
> lsmTrtmt = lsmeans(model1, ~ fTrtmt)
> summary(contrast(lsmTrtmt, method="pairwise", infer=c(T,T)))
> summary(contrast(lsmTrtmt,
+                   list(Flyer1m2AtComnTwst=c(0.5, 0.5, 0, 0, -0.5, -0.5),
+                        LinTwistOrdFlyer=c(-10, -1, 11, 0, 0, 0))),
+           infer=c(T,T), level=0.95, side="two-sided")
 contrast            estimate      SE df lower.CL upper.CL t.ratio p.value
 Flyer1m2AtComnTwst    3.6692  0.6268 60   2.4154    4.923   5.854  <.0001
 LinTwistOrdFlyer    -38.2462  9.3391 60 -56.9272  -19.565  -4.095  0.0001

Results are averaged over the levels of: fBlock
Confidence level used: 0.95
```

the two flyers at common levels of twist, and evaluating the linear trend in twist for the ordinary flyer.

An additional call of `lm`, shown in Table 10.15, replaces the treatment combination factor variable `fTrtmt` with main-effect factor variables `fFlyer` and `fTwist`; this removes their interaction effect from the model. Since not every combination of flyer and twist was observed, the `drop1` command and resulting type 3 sums of squares are used for hypothesis testing. The `lsmeans` function and corresponding `summary(contrast...)` statements again generate multiple comparisons results. The reader may verify that the Bonferroni method provides tighter simultaneous 95% confidence intervals for twist pairwise comparisons than Scheffé's method (results not shown).

Table 10.15 An R program and selected output, continued, for analysis of the cotton-spinning experiment: analysis of factorial main-effects model

```
> # Analysis with additive main effects
> model2 = lm(Break ~ fBlock + fFlyer + fTwist, data=cotton.data)
> drop1(model2, ~., test="F")

Single term deletions

Model:
Break ~ fBlock + fFlyer + fTwist
       Df Sum of Sq RSS AIC F value  Pr(>F)
<none>              306 141
fBlock 12      177 484 152    2.94 0.00281
fFlyer  1      175 481 174   34.84 1.7e-07
fTwist  3      100 407 157    6.65 0.00059

> lsmFlyer = lsmeans(model2, ~ fFlyer)
> summary(contrast(lsmFlyer, method="pairwise"), infer=c(T,T))
> lsmTwist = lsmeans(model2, ~ fTwist)
> summary(contrast(lsmTwist, method="pairwise", adjust="bonf"),
+          infer=c(T,T))
```

The plot of the mean response against twist for both flyer types in Fig. 10.3, p. 317, suggested the possibility that the number of breaks per 100 pounds could be modeled by a flyer effect and a linear twist effect. This can be evaluated by comparing the fit of the block–treatment model,

$$Y_{hi} = \mu + \theta_h + \tau_i + \epsilon_{hi}$$

($i = 1, \ldots, 6; \ h = 1, \ldots, 13$), with the fit of the reduced model,

$$Y_{hjx} = \mu + \theta_h + \alpha_j + \gamma x + \epsilon_{hjx},$$

where α_j is the effect of flyer j ($j = 1, 2$) and x is the *uncoded* amount of twist (cf. Chaps. 8 and 9). Comparing the fit of these models can be done easily in the R software, either using two calls of the `lm` function, one for each model, or using one call that sequentially includes both models. The call of `lm` in the center of Table 10.14 fitted the full block–treatment model, and the call of `lm` at the top of Table 10.16 fits the reduced model that includes the flyer effect and a linear regression in the levels of twist. This second `lm` call is similar to that in Table 10.15 for the the factorial main-effects model. The difference is that when twist is to be regarded as a linear regressor, it is entered as a numeric variable `Twist`—not as the factor variable `fTwist`. The reduced model fits parallel linear regression lines, with intercepts adjusted for block and flyer effects. Output for the reduced model is shown in Table 10.16. The type 1 and type 3 sums of squares generated by the respective `anova` and `drop1` commands are unequal, indicating that effects of `fFlyer` and `fTwist` cannot be estimated independently.

The `summary` command in Table 10.16 causes display of the least squares estimates of the model parameters. We only show one. The slope coefficient of `Twist` is estimated to be -14.100, which, being negative, suggests that the breakages decrease as the twist increases. This slope is declared to

Table 10.16 An R program and selected output, continued, for analysis of the cotton-spinning experiment: analysis of reduced model

```
> # Analysis of covariance: Twist as covariate
> model3 = lm(Break ~ fBlock + fFlyer + Twist, data=cotton.data)
> anova(model3) # to get sse for LOF test

Analysis of Variance Table
Response: Break
          Df Sum Sq Mean Sq F value    Pr(>F)
fBlock    12    177    14.8    2.91   0.00294
fFlyer     1    131   130.8   25.76 0.0000037
Twist      1     87    86.8   17.11   0.00011
Residuals 63    320     5.1

> drop1(model3, ~., test="F")

Single term deletions
Model:
Break ~ fBlock + fFlyer + Twist
       Df Sum of Sq RSS AIC F value  Pr(>F)
<none>                320 140
fBlock 12     177.2 497 150    2.91 0.00294
fFlyer  1     213.2 533 178   42.00 1.6e-08
Twist   1      86.8 407 157   17.11 0.00011

> summary(model3)

Call:
lm(formula = Break ~ fBlock + fFlyer + Twist, data = cotton.data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
                :          :        :       :
Twist        -14.100      3.409   -4.14  0.00011

> lsmFlyer = lsmeans(model3, ~ fFlyer)
> summary(contrast(lsmFlyer, method="pairwise"), infer=c(T,T))

 contrast estimate      SE df lower.CL upper.CL t.ratio p.value
 1 - 2      3.8588 0.59541 63   2.6689   5.0486   6.481  <.0001

Results are averaged over the levels of: fBlock
Confidence level used: 0.95
```

be significantly different from zero, since the t-test of $H_0 : \gamma = 0$ versus $H_A : \gamma \neq 0$ has p-value 0.00011.

The lsmeans function and corresponding summary(contrast…) command compute and display the least squares estimate of the pairwise contrast in the flyer effects and related statistics. The difference in the effects of the two flyers is declared to be significantly different from zero, since the p-value for the t test is less than 0.0001. Likewise, the 95% confidence interval also excludes zero.

The difference in the error sum of squares for the full and reduced models divided by the difference in the error degrees of freedom is the mean square for lack of fit, *msLF* (cf. Chap. 8). It provides the

Table 10.17 An R program and selected output, continued, for analysis of the cotton-spinning experiment: testing for lack of fit

```
> # Testing LOF of ancova model: 2 ways
> anova(model3, model1)

Analysis of Variance Table

Model 1: Break ~ fBlock + fFlyer + Twist
Model 2: Break ~ fBlock + fTrtmt
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1     63 320
2     60 306  3      13.4 0.88   0.46

> model4 = lm(Break ~ fBlock + fFlyer + Twist + fTrtmt, data=cotton.data)
> anova(model4)

Analysis of Variance Table

Response: Break
          Df Sum Sq Mean Sq F value    Pr(>F)
           :      :       :       :         :
fTrtmt     3   13.4     4.5    0.88   0.45914
```

numerator of the test statistic for testing the null hypothesis H_0^R : {the reduced model is adequate} against the alternative hypothesis that the reduced model is not adequate. The decision rule is

$$\text{reject } H_0^R \text{ if } msLF/msE > F_{3,60,\alpha},$$

where
$$msLF = [ssE(\text{reduced}) - ssE(\text{full})] \, / \, [df(\text{reduced}) - \, df(\text{full})] \, .$$

For the cotton-spinning experiment,

$$msLF = (320 - 306)/(63 - 60) = 4.667,$$

although one would obtain the value 4.533 if less rounding of ssE(reduced) and ssE(full) were used. Since $msLF/msE = 4.533/5.1 = 0.88 < 1$, we cannot reject H_0^R for any reasonable significance level α. Hence, the reduced model appears to provide an adequate fit to the data, making interpretation of the parameters in the reduced model meaningful.

One can obtain these calculations from R directly in a couple of ways, as illustrated in Table 10.17. The most direct is the statement

   anova(model3, model1)

which provides just the F-test for lack of fit, comparing the fit of the reduced and full models. Alternatively, the statement

```
model4 = lm(Break ~ fBlock + fFlyer + Twist + fTrtmt, data=cotton.data)
```

Table 10.18 Respiratory exchange ratio data

Subject	Protocol		
	1	2	3
1	0.79	0.80	0.83
2	0.84	0.84	0.81
3	0.84	0.93	0.88
4	0.83	0.85	0.79
5	0.84	0.78	0.88
6	0.83	0.75	0.86
7	0.77	0.76	0.71
8	0.83	0.85	0.78
9	0.81	0.77	0.72

Source Bullough and Melby (1993). Copyright © 1993 Karger, Basel. Reprinted with permission

sequentially includes both models, as the first three terms of the model give the reduced model, then the full model is obtained by inclusion of `fTrtmt`. Since `fTrtmt` has been entered into the model last, its type 1 and 3 sums of squares will be the same, and the corresponding F-test is the desired lack-of-fit test. The command `anova(model4)` generates the type 1 sums of squares, including that for `fTrtmt` which is the test for lack of fit conducted above.

We note that this has been an exercise in model-building, and we can use the model to predict the breakage rates with either flyer over a range of values of twist. However, the model may not fit well for flyer 2 below a twist of 1.69 (see Fig. 10.3).

Exercises

1. **Randomization**

   Conduct a randomization for a randomized complete block design with $v = 4$ treatments observed once ($s = 1$) in each of $b = 5$ blocks.

2. **Randomization**

   Conduct a randomization for a general complete block design for $v = 3$ treatments each observed twice ($s = 2$) in each of $b = 4$ blocks.

3. **DCIS experiment randomization**

   Suppose the DCIS experiment of Example 10.3.2, p. 309, had been designed as a randomized complete block designs with $b = 4$ blocks of size $k = v = 6$, so that each treatment is observed $s = 1$ time per block. Conduct a randomization of the treatments within each block and present the final design.

4. **Respiratory exchange ratio experiment**

   In the resting metabolic rate experiment introduced in Example 10.4.1, p. 311, the experimenters also measured respiratory exchange ratio, which is another measure of energy expenditure. The data for the second 30 minutes of testing are given in Table 10.18.

   (a) Evaluate the assumptions of the block–treatment model (10.4.1) for these data.
   (b) Construct an analysis of variance table and test for equality of the effects of the protocols on respiratory exchange ratio.

Table 10.19 Resistances for the light bulb experiment. Low resistance implies high illumination. (Order of observations is shown in parentheses.)

Block	Treatments					
	1	2	3	4	5	6
I	314 (12)	285 (3)	350 (6)	523 (2)	460 (1)	482 (7)
(60 watt)	300 (13)	296 (9)	339 (8)	497 (4)	470 (5)	498 (11)
	310 (15)	301 (10)	360 (14)	520 (18)	488 (17)	505 (19)
	290 (22)	292 (24)	333 (16)	510 (20)	468 (21)	490 (23)
II	214 (28)	196 (27)	235 (42)	303 (26)	341 (32)	342 (25)
(100 watt)	205 (31)	201 (29)	247 (44)	319 (30)	350 (38)	347 (33)
	197 (35)	197 (39)	233 (46)	305 (34)	323 (41)	352 (37)
	204 (47)	215 (40)	244 (48)	316 (36)	343 (45)	323 (43)

(c) Evaluate the usefulness of blocking.

(d) Use the Scheffé method of multiple comparisons to construct simultaneous 99% confidence intervals for all pairwise comparisons of the protocols as well as the inpatient versus outpatient protocols corresponding to the contrast coefficient list $[\ 1, -\frac{1}{2}, -\frac{1}{2}\ ]$.

5 Light bulb experiment

P. Bist, G. Deshpande, T.-W. Kung, R. Laifa, and C.-H. Wang ran an experiment in 1995 to compare the light intensities of three different brands of light bulbs (coded 1, 2, 3), together with the effect of the percentage capacity (100% and 50%) of the bulb, the latter being controlled by a dimmer switch. Thus, there were $v = 6$ treatment combinations in total:

$$(100\%, \text{Brand } 1) = 1, \ (100\%, \text{Brand } 2) = 2, \ (100\%, \text{Brand } 3) = 3,$$
$$(50\%, \text{Brand } 1) = 4, \ \ (50\%, \text{Brand } 2) = 5, \ \ (50\%, \text{Brand } 3) = 6.$$

Two blocks were used (one per day), with all 60 watt bulbs observed in one block, and all 100 watt bulbs in the other. Four observations were taken on each of the $v = 6$ treatment combinations in each block. The response variable was the observed resistance of a photoresistor connected to the bulb, where high illumination corresponds to low resistance. The data (resistances) are shown in Table 10.19.

(a) Fit a block-treatment-interaction model to the data and calculate an analysis of variance table.

(b) Show that the null hypothesis of no block×treatment interaction would be rejected at level $\alpha = 0.005$.

(c) Calculate a set of confidence intervals for pairwise differences in the treatments for each block separately. Specify the method that you are using and the overall confidence level.

(d) In terms of the six treatment parameters, write down two contrasts representing the interaction between brand and percentage capacity (averaged over blocks). Test whether these contrasts are significantly different from zero.

(e) Suppose that the purpose of the experiment was to determine the best brand (in terms of illumination) for each percentage capacity and for 100 watt bulbs. Which brand(s) would you recommend and why?

Table 10.20 Time in seconds for the water heating experiment; order of observation within each block in parentheses

Treatment combination	Block			
	1	2	3	4
111	261.0 (1)	279.0 (12)	296.7 (6)	282.8 (5)
112	259.4 (12)	249.4 (5)	280.7 (10)	259.9 (4)
121	300.0 (3)	331.8 (10)	308.3 (12)	314.2 (11)
122	286.6 (10)	281.3 (1)	287.7 (4)	276.3 (7)
211	255.7 (2)	304.4 (7)	286.8 (1)	276.4 (12)
212	245.6 (7)	254.7 (9)	249.1 (2)	263.6 (10)
221	266.1 (4)	291.5 (11)	285.7 (8)	294.5 (2)
222	256.4 (6)	262.2 (8)	259.2 (3)	264.0 (8)
311	162.2 (8)	168.1 (4)	147.8 (7)	132.2 (1)
312	137.0 (9)	168.1 (3)	151.9 (9)	169.9 (9)
321	109.6 (5)	109.3 (2)	109.3 (5)	294.5 (3)
322	108.2 (11)	135.3 (6)	111.2 (11)	110.0 (6)

6. **Water heating experiment**

   The purpose of the experiment run by M. Weber, R. Zielinski, J. Y. Lee, S. Xia, and Y. Guo in 2010 was to determine the best way to heat 3 cups of water (for preparation of boxed meals) to 90^oF on a kitchen stove as qsuickly as possible. In this experiment, only one stove was used, and the three treatment factors were

$$\text{C: diameter of pot (5.5, 6.25 and 8.625 inches; coded 1, 2, 3)}$$

$$\text{D: burner size (small, large; coded 1, 2)}$$

$$\text{E: cover (no, yes; coded 1, 2).}$$

   (a) Write out a checklist for such an experiment. Be careful to think about all sources of variation and how to control for them.
   (b) A pilot experiment suggested that the error variance σ^2 would be no larger than 318.9 sec^2. The experimenters wanted to be able to test the hypothesis of no differences in the effects of heating time due to the 12 treatments, with a probability of 0.9 of rejecting the hypothesis if the true difference was $\Delta = 60$ secs. The test was to be done at level $\alpha = 0.05$. Calculate the number of observations that should be taken on each of the 12 treatments.
   (c) The experimenters ultimately decided that they would use a randomized complete block design with $b = 4$ blocks for the experiment, where each block was defined by experimenter and day. The data are shown in Table 10.20. Using the block–treatment model (10.4.1), p. 310, for a randomized complete block design, check the assumptions on the model and test the hypothesis of no effects on the heating time due to treatments.
   (d) Using the factorial form of the block–treatment model similar to (10.8.15), p. 325, but with three treatment factors, test the hypotheses of no interactions between pairs of treatment factors, each test done at level 0.01.
   (e) Taking into account any interactions discovered in part (d), list the contrasts that are of interest to you and, using the Scheffé method, calculate a set of 95% confidence intervals for the contrasts of interest.

Table 10.21 Number of characters recalled for the memory recall experiment

Block		C 1						C 2					
		D 1		D 2		D 3		D 1		D 2		D 3	
		E 1	E 2	E 1	E 2	E 1	E 2	E 1	E 2	E 1	E 2	E 1	E 2
I	F 1	5	33	30	34	2	9	7	18	20	1	35	32
	F 2	22	16	11	10	3	29	21	17	24	23	15	12
	F 3	27	14	28	13	25	8	36	19	4	31	6	26
II	F 1	35	3	16	25	6	31	19	14	10	23	13	27
	F 2	24	5	33	11	8	20	18	34	4	36	9	12
	F 3	32	30	28	7	15	29	17	1	26	21	22	2
III	F 1	10	34	18	33	3	4	19	7	6	15	28	24
	F 2	27	32	26	16	14	22	23	21	11	25	12	36
	F 3	29	5	2	20	35	7	8	31	17	13	1	30
IV	F 1	24	21	16	36	9	18	23	26	28	31	30	11
	F 2	13	10	29	34	15	1	35	12	4	19	14	33
	F 3	5	7	22	2	8	17	27	20	32	7	25	3
V	F 1	20	12	13	21	34	7	9	4	14	23	24	36
	F 2	10	28	18	25	29	31	26	4	16	11	6	22
	F 3	33	2	15	5	30	27	19	8	35	17	32	1

7. **Memory recall experiment**

The memory recall experiment was run in 2007 by C. Lucas, A. Moczdlowski, M. Salwan, X. Wang, and J. Williams to investigate effects of various factors that might possibly be important for reaching consumers through print advertisements. The experiment was run as a randomized complete block design and involved $b = 5$ subjects, each of whom formed a block of the design. For each treatment combination, each subject was shown a grid of 25 characters. After studying the grid for a specified length of time, each subject was asked to recall the placement of characters on the grid. The number of characters correctly recalled in the correct location on the grid formed the response. The treatment factors were:

C: Paper (neon green or white, coded 1, 2)

D: Time for studying the grid (30, 60, 90 seconds, coded 1, 2, 3)

E: Background music (no music or classical music, coded 1, 2)

F: Character type (letters, numbers, or combination of both, coded 1, 2, 3)

The grids were randomly generated with characters according to the level of factor F. The data are shown in Table 10.21.

(a) Fit a block-treatment model to the data.
(b) Since the data are counts, one should be concerned whether the normality and equal variance assumptions on the errors are approximately satisfied. Also there may be a time order effect as the subject tires or gets better at the task (–the time orders are given in the data set on the website, p. 54, for the book). By examining some relevant residual plots, show that the assumptions on your model are approximately satisfied.
(c) Calculate an analysis of variance table and explain what conclusions you can draw. Give your reasons, including explicit hypotheses being tested and your Type I error rates.

Table 10.22 Data for the hypothetical chemical experiment

Lab (Block)	Treatment combinations							
	111	112	121	122	211	212	221	222
1	7.3	9.5	13.8	15.4	16.0	18.7	11.3	14.5
2	8.8	11.3	15.3	17.7	17.9	20.8	12.0	15.4
3	11.7	14.1	17.2	22.3	22.6	24.8	16.9	18.5
4	6.2	8.3	11.2	15.4	16.8	17.4	8.2	12.5

(d) At a significance level of 0.01, test the hypothesis that there is no quadratic trend in the number of characters recalled as the length of the study time increases.

(e) Suppose that you (pre-)planned to calculate a set of 99% simultaneous confidence intervals for the pairwise comparisons of the three times of study. Explain whether you would use Bonferroni, Tukey, Scheffé, or Dunnett's method.

(f) Calculate a 90% upper bound for the true value of σ^2.

(g) Write down the formula for a 95% confidence interval which compares the effect of no music versus classical music on the number of symbols recalled (averaged over the other factors).

(h) Suppose now that you wished to design a larger experiment for the future, but with the same treatment combinations. The design will be a randomized complete block design with each of subjects seeing all 36 treatment combinations in a random order. If you require the confidence interval in part (g) to be no wider than 2.0, how many subjects would you recommend?

8. **Hypothetical chemical experiment**

An experiment to examine the yield of a certain chemical was conducted in $b = 4$ different laboratories. The treatment factors of interest were

A : acid strength (80% and 90%, coded 1, 2)
B : time allowed for reaction (15 and 30 min, coded 1, 2)
C : temperature (50° and 75 °C, coded 1, 2)

The experiment was run as a randomized complete block design with the laboratories as the levels of the blocking factor. The resulting data (yields in grams) are shown in Table 10.22. The goal of the experiment was to find the treatment combination(s) that give(s) the highest average yield.

(a) Plot the data and comment on your chosen plots.

(b) Fit a block–treatment model to these data and show that the assumptions on the model are approximately satisfied.

(c) Suppose that the pre-plan was to calculate a 99% set of pairwise comparisons between the treatment combinations using Tukey's method, to calculate 99.5% intervals for the comparisons between the levels of A, B and C if these were not involved in interactions and a set of 99% intervals using Scheffé's method for any other contrasts that look interesting. The overall confidence level would then be at least 96.5%. List any contrasts that you would like to examine further after looking at the plots in part (a).

(d) Calculate an analysis of variance table and test any hypotheses of interest, each at level 0.01. State your conclusions clearly.

(e) Calculate confidence intervals for the contrasts specified in part (c), and state your conclusions.

(f) The objective of the experiment was to find the combination that gives the highest yield. Using all the information that you have gathered, which treatment combination would you recommend?

9. **Reaction time experiment, continued, (sample size)**

The reaction time pilot experiment was described in Exercise 4, p. 100, and analyzed in Examples 6.4.3 and 6.4.5, pp. 153 and 160. The experiment was run to compare the speed of response of a human subject to audio and visual stimuli. The two treatment factors were "Cue Stimulus" at two levels "auditory" and "visual" (Factor A, coded 1, 2), and "Cue Time" at three levels 5, 10, and 15 seconds between cue and stimulus (Factor B, coded 1, 2, 3), giving a total of $v = 6$ treatment combinations. The pilot experiment used only one subject, for whom $msE = 0.00029$ seconds^2 based on 12 degrees of freedom. An upper 95% confidence bound for the error variance was calculated in Example 6.4.2, p. 151, as $\sigma^2 \leq 0.000664$ seconds^2. To be able to draw conclusions about these six treatment combinations, it is important for the main experiment to use a random sample of subjects from the population.

(a) Consider using a randomized complete block design with b subjects representing blocks for the main experiment. Let the block sizes be $k = 6$, so that each treatment combination can be observed once for each subject. How many subjects are needed if the widths of simultaneous 99% confidence intervals for the pairwise comparisons of the treatment combinations need to be less than 0.01 seconds to be useful (that is, we require $msd < 0.005$ seconds)?
(b) If $b = 4$ subjects were available, and a general complete block design were used with block size $k = 6s$, how many observations would be needed on each treatment in each block to satisfy the requirements of the confidence intervals in part (a)?

10. **Length perception experiment**

The experiment was run by B. Millen, R. Shankar, K. Christoffersen, and P. Nevathia in 1996 to explore subjects' ability to reproduce accurately a straight line of given length. A 5 cm line (1.9685 inches) was drawn horizontally on an 11×8.5 in sheet of plain white paper. The sheet was affixed at eye level to a white projection screen located four feet in front of a table at which the subject was asked to sit. The subject was asked to reproduce the line on a sheet of white paper on which a border had been drawn. Subjects were selected from a population of university students, both male and female, between 20 and 30 years of age. The subjects were all right-handed and had technical backgrounds.

There were six different borders representing the combinations of three shapes—square, circle, equilateral triangle (levels of factor C, coded 1, 2, 3) and two areas—16 in^2 and 9 in^2 (levels of factor D, coded 1, 2). The purpose of the experiment was not to see how close to the 5 cm that subjects could draw, but rather to compare the effects of the shape and area of the border on the length of the lines drawn. The subjects were all able to draw reasonably straight lines by hand, and one of the experimenters measured, to the nearest half millimeter, the distance between the two endpoints of each line drawn. Data from 14 of the subjects are shown as deviations from the target 5 cm in Table 10.23.

(a) Fit a block–treatment model to the data using subjects as blocks and with six treatments representing the shape–area combinations. Check the error assumptions on your model.
(b) Draw at least one graph and examine the data.
(c) Write down contrasts in the six treatment combinations representing the following comparisons:

  (i) differences in the effects of area for each shape separately,
  (ii) average difference in the effects of area,
  (iii) average difference in the effects of shape.

Table 10.23 Data for the length perception experiment

Subject	Treatment combinations (shape, area)					
	11	12	21	22	31	32
1	0.20	−0.25	0.85	−0.50	0.40	0.05
2	1.70	0.30	1.80	0.40	1.40	1.80
3	−0.60	−0.90	−0.90	−0.50	−0.70	−0.50
4	0.60	0.10	0.70	0.20	0.70	0.60
5	0.50	0.40	0.30	0.70	0.50	0.60
6	0.20	−0.60	0.00	−1.40	−0.60	−1.20
7	1.30	−0.10	−0.40	0.50	−0.15	0.30
8	−0.85	−1.30	−0.40	−1.55	−0.85	−1.30
9	0.80	0.05	0.55	1.25	1.30	0.20
10	0.10	−0.10	−0.30	0.95	0.30	−0.95
11	−0.20	−0.40	−0.50	−0.30	−0.40	−0.40
12	0.05	−0.20	0.55	0.60	0.10	0.10
13	0.80	−0.60	0.20	−0.60	−0.60	−0.30
14	−0.25	−0.70	0.00	−0.70	−0.10	−0.95

(d) Give a set of 97% simultaneous confidence intervals for the contrasts in (c)(i). State your conclusions.

(e) Under what conditions would the contrasts in (c)(ii) and (iii) be of interest? Do these conditions hold for this experiment?

11. **Load-carrying experiment**

The purpose of the experiment run by M. Flannery, C. Lee, E. Nelson, and P. Sparto in 1993 was to investigate the load-carrying capability of the human arm. Subjects were selected from a population of healthy males. The maximum torque generated at the elbow joint was measured (in newtons) using a dynamometer for each subject in a 5 min exertion for nine different arm positions (in a random order). The nine arm positions were represented by $v = 9$ treatment combinations consisting of levels of the two factors "flex" with levels 0°, 45°, 90° of elbow flexion, coded 0, 1, 2, and "rotation" with levels 0°, 45°, 90° of shoulder rotation, coded 0, 1, 2.

The experiment was run as a randomized complete block design with four blocks, each block being defined by a different subject. The subjects were selected from the populations of male students in the 20–30 year range in a statistics class.

(a) Identify your proposed analysis, including your model, any hypotheses you would propose testing, and any confidence intervals you would propose calculating. What are your overall significance level and confidence level?

(b) The experimenters decided that they required Scheffé's 95% confidence intervals for any normalized contrast in the main effects of each factor separately to be no wider than 10 newtons. How many subjects would have been needed to satisfy this requirement if the error variance is similar to the value $msE = 670$ newtons^2 that was obtained in the pilot experiment?

(c) The data are shown in the order collected in Table 10.24. Plot the data. Are there any other contrasts that you would like to examine in addition to any pre-planned contrasts that you identified in (a)?

Table 10.24 Data and order of collection for the load-carrying experiment

Order	1	2	3	4	5	6	7	8	9
Treat. Comb.	31	21	23	11	22	13	12	32	33
Subj 1	250	230	170	160	240	160	150	200	180
Treat. Comb.	11	22	31	32	23	21	12	33	13
Subj 2	230	260	260	220	250	270	230	190	210
Treat. Comb.	21	13	31	32	11	22	12	33	23
Subj 3	230	180	210	190	150	190	140	160	180
Treat. Comb.	31	11	33	22	13	32	21	12	23
Subj 4	360	200	380	290	240	310	280	350	210

Table 10.25 Data for the biscuit experiment (percentage of original height)

Block	Treatment combination								
	11	12	13	21	22	23	31	32	33
1	350.0	375.0	362.5	237.5	237.5	256.3	191.7	216.7	208.3
	300.0	362.5	312.5	231.3	231.3	243.8	200.0	212.5	225.0
2	362.8	350.0	367.5	250.0	262.5	250.0	245.8	212.5	241.7
	412.5	350.0	387.5	268.8	231.3	237.5	225.0	250.0	225.0
3	350.0	387.5	425.0	300.0	275.0	231.3	204.4	187.5	187.5
	337.5	362.5	400.0	262.5	206.3	262.5	204.2	204.2	208.3
4	375.0	362.5	400.0	318.8	250.0	243.8	200.0	216.7	212.5
	350.0	337.5	350.0	256.3	243.8	250.0	150.0	183.3	187.5

(d) Are the assumptions on block–treatment model (10.4.1) approximately satisfied for these data? Pay particular attention to outliers. If the assumptions are satisfied, analyze the experiment. If they are not satisfied, what information can you gather from the data?

(e) Do your conclusions apply to the whole human population? Explain.

12. **Biscuit experiment**

The biscuit experiment was run in 1994 by N. Buurma, K. Davis, M. Gross, M. Kresja, and K. Zitoun in order to determine how to make fluffy biscuits. The two treatment factors of interest were "height of uncooked biscuit" (0.25, 0.50, or 0.75 inches, coded 1, 2, and 3) and "kneading time" (number of times: 7, 14, or 21, coded 1, 2, and 3). The design used was a general complete block design with $b = 4$ blocks, consisting of the four runs of an oven. The experimental units consisted of $k = 18$ positions on a baking pan. Each of the $v = 9$ treatment combinations was observed $s = 2$ times per block. The resulting observations are "percentage of original height" (so, for example, 362.5 means the height of the cooked biscuit is 3.625 times the height of the uncooked biscuit). The data are shown in Table 10.25.

(a) State a suitable model for this experiment and check that the assumptions on your model hold for these data.

(b) Evaluate whether blocking was worthwhile in this experiment.

(c) Use an appropriate multiple comparisons procedure to evaluate which treatment combination yields the largest percentage increase in height.

(d) Write down the contrast that measures the linear trend in the response as the kneading time increases. Test whether this contrast is significantly different from zero.

Table 10.26 Dissolving time for the effervescent experiment (Order of observation in parentheses)

Block	Treatment combination					
	11	12	13	21	22	23
I	75.525 (8)	68.125 (3)	44.825 (7)	78.350 (1)	40.575 (2)	27.450(18)
	70.325 (9)	47.525 (4)	36.200(10)	76.050(12)	40.000 (5)	26.600(19)
	69.925(17)	61.475 (6)	39.350(11)	78.425(15)	39.500(20)	24.950(21)
	69.800(23)	58.625(14)	37.425(13)	71.525(16)	40.400(22)	26.325(24)
II	83.475(34)	71.759(36)	51.975(28)	92.725(31)	42.275(30)	25.400(25)
	86.800(41)	70.825(37)	50.100(29)	77.957(35)	44.425(32)	26.333(26)
	83.750(44)	73.925(42)	51.225(33)	85.425(39)	42.475(38)	25.875(27)
	79.575(46)	71.550(48)	53.700(47)	87.333(45)	44.300(42)	26.650(40)

13. **Effervescent experiment**

The effervescent experiment was run by B. Bailey, J. Lewis, J. Speiser, Z. Thomas, and S. White in 2011 to compare dissolving times of two different brands (name brand, store brand, coded 1, 2) of cold medicine tablets in three different equally spaced water temperatures (6^oC, 23^oC, 40^oC, coded 1, 2, 3). A complete block design with $b = 2$ blocks was selected with $s = 4$ observations on each of the $v = 6$ treatment combinations in each block. In Block I, the liquid was stirred using a magnetic sirring plate at 350 revolutions per minute. In Block II, the liquid was not stirred.

The dissolving time was measured from the time a tablet was dropped (from a fixed height) into 60 mL of water to the time the tablet was completely dissolved. The recorded observation was taken as an average of the times as measured by four experimenters and the data are shown in Table 10.26.

(a) Plot the data and comment on any interesting features.
(b) Fit the block-treatment model to the data and check the assumptions on the model, paying particular attention to outliers and equal variances.
(c) Investigate the effects on the dissolving time due to the different temperatures, including pairwise comparisons, and linear and quadratic trends (assuming that these investigations are pre-planned).

14. **Colorfastness experiment, continued**

The colorfastness experiment was described in Example 10.6.3, p. 323. There were 5 levels of the treatment factor "number of times of laundering" and three blocks formed by the experimenters. The ideal number of observations was calculated in Example 10.6.3 to be $s = 12$ observations per treatment per block so a total of $k = vs = 60$ swatches of material were evaluated for color by each experimenter.
The experiment was carried out according to the guidelines of the American Association of Textile Chemists and Colorists Test 61-1980. The measurements that are given in Table 10.27 were made using the Gray Scale for Color Change. This scale is measured using the integers 1–5, where 5 represents no change from the original color. Using their own continuous version of the Gray Scale, each of the $b = 3$ experimenters made color determinations on $s = 12$ swatches of fabric for each of the $v = 5$ treatments (numbers of washes) in a random order and without knowledge of which treatment was being evaluated —a "blind study".

Table 10.27 Data for the colorfastness experiment

Block (Experimenter)	Number of washes	y_{hit} (Measurement on the gray scale)
1	1	3.8, 4.0, 4.0, 3.9, 3.8, 3.7, 3.9, 4.0, 4.0, 4.0, 3.9, 4.0
	2	3.0, 3.7, 3.8, 3.0, 3.7, 4.0, 2.9, 3.5, 3.2, 3.5, 4.0, 3.5
	3	3.7, 3.3, 3.5, 3.6, 3.1, 3.0, 3.2, 3.7, 3.8, 3.7, 3.6, 3.6
	4	3.0, 3.6, 3.9, 3.8, 3.8, 3.1, 3.6, 3.4, 4.0, 3.2, 3.0, 3.8
	5	3.6, 3.1, 3.8, 3.4, 3.9, 3.4, 3.5, 4.0, 3.4, 3.9, 3.0, 3.3
2	1	4.5, 3.8, 3.5, 3.5, 3.6, 3.8, 4.6, 3.9, 4.0, 3.9, 3.8, 4.2
	2	3.7, 3.6, 3.8, 3.5, 3.8, 4.0, 3.6, 3.6, 3.4, 3.7, 3.4, 3.3
	3	3.0, 3.7, 2.8, 3.0, 3.6, 3.4, 3.8, 3.6, 3.4, 3.7, 3.9, 3.8
	4	4.2, 3.8, 3.1, 2.8, 3.2, 3.0, 3.7, 3.0, 3.7, 3.5, 3.2, 3.9
	5	3.2, 3.5, 3.1, 3.3, 2.8, 3.5, 3.5, 3.2, 3.6, 3.7, 3.2, 3.2
3	1	4.0, 4.2, 3.8, 3.8, 4.2, 4.2, 3.8, 4.2, 4.2, 3.8, 4.2, 3.9
	2	3.2, 2.8, 2.8, 4.0, 3.0, 3.2, 3.8, 3.5, 4.0, 3.2, 3.5, 3.4
	3	3.8, 4.0, 3.8, 3.4, 4.2, 3.4, 4.0, 3.8, 4.2, 3.9, 3.9, 3.1
	4	4.2, 3.8, 3.5, 3.4, 4.2, 2.9, 3.5, 3.2, 3.5, 4.0, 3.2, 3.9
	5	3.5, 3.8, 2.8, 4.2, 4.0, 3.8, 3.9, 2.9, 3.9, 3.2, 3.5, 3.5

(a) Plot the treatment averages for each block. Comment on a possible interaction between experimenter and number of washes, and also on any surprising features of the data.
(b) Fit a block–treatment-interaction model (10.6.8) to these data. Check the assumptions of normality, equal variance, and independence of the error variables.
(c) Using only the data from experimenters 1 and 2, repeat part (b). Under what circumstances could you justify ignoring the results of experimenter 3?
(d) Investigate the linear and quadratic trends in the effect on color of the number of washes. If necessary, use Satterthwaite's approximation for unequal variances.

15. **Insole cushion experiment**

The insole cushion experiment was run in the Gait Laboratory at The Ohio State University by V. Agresti, S. Decker, T. Karakostas, E. Patterson, S. Schwartz, 1995. The objective of the experiment was to compare the effect on the force with which the foot hits the ground of a regular shoe insole cushion and a heel cushion (factor C, coded 1, 2, respectively) available both as brand name and a store name (factor D, coded 1, 2, respectively).

Only one subject (and one pair of shoes) was used. A pilot experiment indicated that fatigue would not be a factor. The natural walking pace of the subject was measured before the experiment. This same pace was maintained throughout the experiment by use of a metronome.

The experiment was divided into two days (blocks). On one day, measurements were taken on the subject's dominant leg (kicking leg) and the nondominant leg was examined on the second day. Each of the $v = 4$ treatment combinations were measured $s = 5$ times per block in a randomized order. For each treatment combination, the subject was instructed to walk naturally along the walkway of the laboratory without looking down. As the foot hit the "force plate," an analog signal was sent to a computer, which then converted the signal to a digital form. The response variable, shown in Table 10.28, is the maximum deceleration of the vertical component of the ground reaction force measured in newtons.

Table 10.28 Data for the insole cushion experiment

		Block I (Right leg)				
C	*D*	Response in Newtons (order)				
1	1	899.99 (3)	910.81 (5)	927.79 (10)	888.77 (11)	911.93 (16)
1	2	924.92 (2)	900.10 (6)	923.55 (12)	891.56 (17)	885.73 (20)
2	1	888.09 (4)	954.11 (7)	937.41 (9)	911.85 (14)	908.41 (18)
2	2	884.01 (1)	918.36 (8)	880.23 (13)	891.16 (15)	917.16 (19)
		Block II (Left Leg)				
C	*D*	Response in Newtons (order)				
1	1	852.94 (22)	866.28 (27)	886.65 (28)	851.14 (33)	869.80 (34)
1	2	882.95 (21)	865.58 (24)	868.15 (25)	893.82 (37)	875.98 (38)
2	1	920.93 (26)	880.26 (31)	897.10 (35)	893.78 (39)	885.80 (40)
2	2	872.50 (23)	892.76 (29)	895.93 (30)	899.44 (32)	912.00 (36)

(a) Fit a model that includes a block×treatment interaction. Prepare an analysis of variance table. What can you conclude?

(b) Draw interaction plots for the *CD* interaction, *C*× block interaction, *D*× block interaction, and Treatment Combination×block interaction. Which contrasts would be of interest to examine?

(c) Calculate confidence intervals for any means or contrasts that you identified in part (b), after having looked at the data.

(d) Check the assumptions on the model. There are two possible outliers. Re-examine the data without either or both of these observations. Do any of your conclusions change? Which analysis would you report?

16. **Yeast experiment**

The investigators (K. Blenk, M. Chen, G. Evans, J. Chen Ibinson, J. Lamack, and E. Scott, 2000) planned an experiment to investigate how rapid rise yeast and regular yeast differ in terms of their rate of rising. They were also interested in finding out whether temperature had significant effect on the rising rate. For each observation, 0.3 gm of yeast and 0.45 gm sugar were mixed together and added to a test tube, together with 6ml of water. The test tube was placed into a water bath of specified temperature. The level (height) of the mixture in the test tube was recorded immediately and then again after 15 minutes. Each response is the percentage gain in height of the mixture in the test tube after the 15 minutes. There were three treatment factors:

Factor C: Initial temperature of water mixed with the yeast and flour
(3 levels: 100°F, 115°F, 130°F)
Factor D: Type of yeast (2 levels: Rapid rise, Regular)
Factor E: Temperature of water bath (2 levels: 70°F, 85°F)

There were $b = 3$ blocks with each experimenter-team forming a block. Each team took $s = 2$ observations on each of the $v = 12$ treatment combinations. The factorial form of the block–treatment model similar to (10.8.15), but with three treatment factors, was selected.

(a) Explain in at most two sentences why the treatment combinations should be randomly ordered in each block before they are observed.

Table 10.29 Data for the yeast experiment (percentage rise). Treatment combinations are the levels of (water temperature, yeast, bath temperature)

	Block 1		Block 2		Block 3	
water/yeast/bath	y_{hijk1}	y_{hijk2}	y_{hijk1}	y_{hijk2}	y_{hijk1}	y_{hijk2}
111	8.2	2.7	10.9	1.8	11.4	4.8
112	30.0	39.8	31.8	36.0	42.4	20.0
121	12.6	17.7	3.5	3.4	3.4	8.5
122	64.7	73.9	42.0	32.6	34.5	30.0
211	18.1	5.4	12.3	18.3	8.5	8.0
212	63.5	66.3	23.7	57.5	30.6	45.3
221	4.2	12.2	7.7	8.3	6.0	8.2
222	96.8	71.1	34.1	40.9	49.3	46.0
311	44.4	16.4	5.0	4.8	8.5	3.3
312	58.2	63.3	29.2	27.8	37.5	18.2
321	19.8	9.4	4.8	6.7	6.4	12.9
322	99.7	92.3	53.2	58.9	43.9	73.7

(b) Why would you check for normality of the residuals in general, and in this experiment in particular?

(c) Before the experiment, the experimenters had planned to calculate a set of 99% confidence intervals for the pairwise comparisons of the treatment combinations (averaged over blocks). A pilot experiment indicated that the normality assumption is approximately satisfied and that the error variance would be about 9 percent^2. They wished to have confidence intervals of half-width at most 5 percent. How many observations per treatment combination per block would this have required?

(d) Due to time considerations, the experimenters were only able to take 2 observations per treatment combination per block as shown in Table 10.29. Calculate an analysis of variance table and explain what conclusions you can draw from it. Be careful about interpreting main effects in the presence of interactions, and be careful about your levels of significance.

(e) Illustrate the main points about the treatment factor effects (averaged over the blocks) that you mentioned in part (d) by sketching two plots. Choose these plots carefully and explain why you chose them.

(f) Test at significance level 0.01 whether or not there is a significant linear trend in the "percent-gain in height" as the level of initial water temperature increases (averaged over all the other variables). State your conclusions.

17. **Cotton-spinning experiment**

In the cotton-spinning experiment of Sect. 10.5, p. 314, the two observations on treatment 1 (ordinary flier, 1.69 twist) arising from blocks 5 and 10 appear to be outliers.

(a) Using a computer package, repeat the analysis of Sect. 10.5 without these two observations.

(b) Show that the assumptions on the block–treatment model (10.4.1) are approximately satisfied when these two outliers are omitted.

(c) Draw conclusions about the fliers and degrees of twist from your analysis. Do any of your conclusions contradict those drawn when the outliers were included?

(d) Which analysis do you prefer and why?

Incomplete Block Designs

<div align="right">

11

</div>

11.1 Introduction

When an experiment needs to be run in blocks but, for practical reasons, the block size cannot be a multiple of the number of treatments, then a complete block design (Chap. 10) cannot be used—an *incomplete block design* needs to be used instead. The incomplete block designs discussed in this chapter have block size smaller than the number of treatments, but larger block sizes can be obtained by adding one or more complete set of treatments to every block.

In Sect. 11.2, we discuss basic design issues of block size, randomization and estimability. Then, in Sect. 11.3, three useful and efficient types of incomplete block designs (balanced incomplete block designs, group divisible designs, and cyclic designs) are introduced. Analysis of incomplete block designs is described in Sect. 11.4, including some specific formulae for balanced incomplete block designs and group divisible designs. In Sect. 11.5, we describe and analyze an experiment that was designed as a cyclic group divisible design. Sample-size calculations are discussed in Sect. 11.6, and factorial experiments in incomplete block designs are considered in Sect. 11.7. In general, since not every combination of treatment and block is observed, incomplete block designs are most easily analyzed using computer software. Illustrations are given in Sect. 11.8 by SAS software and in Sect. 11.9 by R.

11.2 Design Issues

11.2.1 Block Sizes

Block sizes are dictated by the availability of groups of similar experimental units. For example, in the breathalyzer experiment examined in Sect. 10.3.1, p. 306, the block size was chosen to be $k = 5$. This choice was made because the pilot experiment indicated that experimental conditions were fairly stable over a time span of five observations taken close together, and also because five observations could be taken by a single technician in a shift. In other experiments, the block size may be limited by the capacity of the experimental equipment, the availability of similar raw material, the length of time that a subject will agree to remain in the study, the number of observations that can be taken by an experimenter before fatigue becomes a problem, and so on. Such restrictions on the block size may result in the blocks being too small for every treatment to be observed the same number of times in every block. The breathalyzer experiment required the comparison of $v = 36$ treatment combinations (twelve different

Table 11.1   An incomplete block design with $b = 8$, $k = 3$, $v = 8$, $r = 3$

Block				Block			
I	1	3	8	V	5	7	4
II	2	4	1	VI	6	8	5
III	3	5	2	VII	7	1	6
IV	4	6	3	VIII	8	2	7

alcohol concentrations combined with three air-intake ports) of which only five would be observed per block. Skill was then needed in selecting the best design that would still allow all treatment contrasts to be estimable with high precision.

11.2.2   Design Plans and Randomization

All the designs that we discuss in this chapter are equireplicate; that is, every treatment (or treatment combination) is observed r times in the experiment. These tend to be the most commonly used designs, although nonequireplicate designs are occasionally used in practice.

   We use the symbol n_{ih} to denote the number of times that treatment i is observed in block h. In general, it is better to observe as many different treatments as possible in a block, since this tends to decrease the average variance of the contrast estimators. Therefore, when the block size is smaller than the number of treatments, each treatment should usually be observed either once or not at all in a block. Such block designs are called *binary,* and every n_{ih} is either 0 or 1. For most purposes, the best binary designs are those in which pairs of treatments occur together in the same number (or nearly the same number) of blocks. These designs give rise to equal (or nearly equal) lengths of confidence intervals for pairwise comparisons of treatment effects.

   There are three stages in designing an experiment with incomplete blocks. The first stage is to obtain as even a distribution as possible of treatment labels within the blocks. This results in an *experimental plan*. The plan in Table 11.1, for example, shows a design with $b = 8$ blocks (labeled I, II, ..., VIII) each of size $k = 3$, which can be used for an experiment with $v = 8$ treatments (labeled 1, ..., 8) each observed $r = 3$ times. The treatment labels are evenly distributed in the sense that no label appears more than once per block and pairs of labels appear together in a block either once or not at all, which is "as equal as possible".

   The experimental plan is often called the "design," even though it is not ready for use until the random assignments have been made. There are three steps to the randomization procedure, as follows.

  (i) Randomly assign the block labels in the plan to the levels of the blocking factor(s).
 (ii) Randomly assign the experimental units in a block to those treatment labels allocated to that block.
(iii) Randomly assign the treatment labels in the plan to the actual levels of the treatment factor.

   The randomization procedure is illustrated in the following example.

Example 11.2.1   `Metal alloy experiment`

Suppose an experiment is to be run to compare $v = 7$ compositions of a metal alloy in terms of tensile strength. Further, suppose that only three observations can be taken per day, and that the experiment must be completed within seven days. It may be thought advisable to divide the experiment into blocks, with each day representing a block, since different technicians may work on the experiment on different

Table 11.2 Randomization of an incomplete block design

Block label	Unrandomized design	Block label	Day	Design after step (i)	Design after step (ii)
I	1 2 4	VI	1	6 7 2	2 7 6
II	2 3 5	II	2	2 3 5	5 3 2
III	3 4 6	III	3	3 4 6	4 3 6
IV	4 5 7	I	4	1 2 4	2 1 4
V	5 6 1	V	5	5 6 1	1 5 6
VI	6 7 2	IV	6	4 5 7	4 5 7
VII	7 1 3	VII	7	7 1 3	7 3 1

days and the laboratory temperature may vary from day to day. Thus, an incomplete block design with $b = 7$ blocks of size $k = 3$ and with $v = 7$ treatment labels is needed. The plan shown in the first two columns of Table 11.2 is of the correct size. It is binary, with every treatment appearing 0 or 1 times per block and $r = 3$ times in total. Also, all pairs of treatments occur together in a block exactly once, so the treatment labels are evenly distributed over the blocks. Randomization now proceeds in three steps.

Step (i): The block labels need to be randomly assigned to the 7 days. Suppose we obtain the following pairs of random digits from a random number generator or from Table A.1 and associate them with the blocks:

Random digits: 71 36 65 93 92 02 97
Block labels: I   II III IV V VI VII

Then, sorting the random numbers into ascending order, the blocks of the plan are assigned to the seven days as in columns 3–5 of Table 11.2.

Step (ii): Now we randomly assign time slots within each day to the treatment labels. Again, using pairs of random digits either from a random number generator or from where we left off in Table A.1, we associate the random digits with the treatment labels as we illustrate here for the first three days:

Day: Day 1   Day 2   Day 3
Block: (Block VI) (Block II) (Block III)
Random digits: 50 29 03 65 34 30 74 56 88
Treatment labels: 6   7   2   2   3   5   3   4   6

Sorting the random numbers into ascending order for each day separately gives the treatment label order 2, 7, 6 for day 1, and 5, 3, 2 for day 2, and 4, 3, 6 for day 3, and so on. The design after step (ii) is shown in the last column in Table 11.2. A third set of random digits is now required to associate the treatment numbers in the plan with the 7 compositions of metal alloy.   □

11.2.3 Estimation of Contrasts

The importance of selecting an experimental plan with an even distribution of treatment labels within the blocks, such as those in Tables 11.1 and 11.2, is to ensure that all treatment contrasts are estimable and that pairwise comparison estimators have similar variances. The plan shown in Table 11.3 is poor. Although all treatments appear $r = 3$ times in the design, some pairs of treatment labels (such as 1 and 3) occur together in two blocks, while other pairs (such as 1 and 2) never appear together. Worse

Table 11.3 A disconnected incomplete block design with $b = 8, k = 3, v = 8, r = 3$

Block				Block			
I	1	3	5	V	5	7	1
II	2	4	6	VI	6	8	2
III	3	5	7	VII	7	1	3
IV	4	6	8	VIII	8	2	4

Fig. 11.1 Connectivity graphs to check connectedness of designs
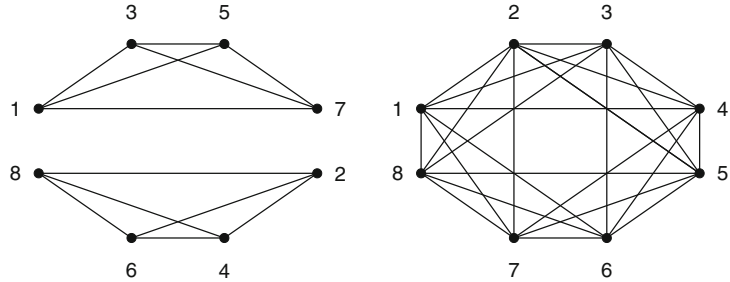
(a) Disconnected design of Table 11.3   **(b)** Connected design in Table 11.1

still, is that some blocks contain all the even-numbered treatment labels, and the other blocks contain all the odd-numbered labels. The result is that every pairwise comparison between an even-numbered and an odd-numbered treatment is not estimable. The design is said to be *disconnected*.

Disconnectedness can be illustrated through a *connectivity graph* as follows. Draw a point for each treatment and then draw a line between every two treatments that occur together in any block of the design. The connectivity graph for the disconnected design in Table 11.3 is shown in Fig. 11.1a. Notice that the graph falls into two pieces. There is no line between any of the odd-labeled treatments and the even-labeled treatments.

A design is *connected* if every treatment can be reached from every other treatment via lines in the connectivity graph. The connectivity graph for the connected design in Table 11.1 is shown in Fig. 11.1b and it can be verified that there is a path between every pair of treatments. For example, although treatments 1 and 5 never occur together in a block and so are not connected by a line, there is nevertheless a path from 1 to 4 to 5. *All contrasts in the treatment effects are estimable in a design if and only if the design is connected*. The connectivity graph therefore provides a simple means of checking estimability.

Although disconnected designs will be useful in Chap. 13 for single-replicate ($r = 1$) factorial experiments arranged in blocks, they need never be used for experiments with at least two observations per treatment. All balanced incomplete block designs are connected, and so are most group divisible designs and cyclic designs. These three types of design are described next.

11.3   Some Special Incomplete Block Designs

11.3.1 Balanced Incomplete Block Designs

A *balanced incomplete block design* is a design with v treatment labels, each occurring r times, and with bk experimental units grouped into b blocks of size $k < v$ in such a way that the units within a

Table 11.4   A balanced incomplete block design with $v = 8, r = 7, b = 14, k = 4, \lambda = 3$

Block	Treatments				Block	Treatments			
I	1	2	3	4	VIII	2	3	5	8
II	1	2	5	6	IX	2	3	6	7
III	1	2	7	8	X	2	4	5	7
IV	1	3	5	7	XI	2	4	6	8
V	1	3	6	8	XII	3	4	5	6
VI	1	4	5	8	XIII	3	4	7	8
VII	1	4	6	7	XIV	5	6	7	8

block are alike and units in different blocks are substantially different. The plan of the design satisfies the following conditions:

 (i) Each treatment label appears either once or not at all in a block (that is, the design is binary).
(ii) Each pair of labels appears together in λ blocks, where λ is a fixed integer.

Block design randomization is carried out as illustrated in Sect. 11.2.2.

All balanced incomplete block designs have the desirable properties that all treatment contrasts are estimable and all pairwise comparisons of treatment effects are estimated with the same variance so that their confidence intervals are all the same length. Balanced incomplete block designs also tend to give the shortest confidence intervals on the average for any large number of contrasts. For these reasons, the balanced incomplete block design is a popular choice among experimenters. The main drawback is that such designs exist only for some choices of v, k, b, and r.

The design in Table 11.2 is a balanced incomplete block design for $v = 7$ treatments and $b = 7$ blocks of size $k = 3$. It can be seen that conditions (i) and (ii) are satisfied, with every pair of labels appearing together in exactly $\lambda = 1$ block. A second example of a balanced incomplete block design, prior to randomization, is shown in Table 11.4 for $v = 8$ treatments in $b = 14$ blocks of size $k = 4$. Again, conditions (i) and (ii) are satisfied, this time with $\lambda = 3$.

We can verify that the design in Table 11.1 (p. 350) with $v = b = 8, r = k = 3$ is not a balanced incomplete block design. Label 2, for example, appears in one block with each of labels 1, 3, 4, 5, 7, and 8 but never with label 6. The following simple argument shows that no balanced incomplete block design can possibly exist for this size of experiment. In a balanced incomplete block design with $v = b = 8, r = k = 3$, label 2, for example, must appear in $r = 3$ blocks in the design, and in each block there are $k - 1 = 2$ other labels. So label 2 must appear in a block with a total of $r(k-1) = 6$ other treatment labels. Consequently, if label 2 were to appear λ times with each of the other $v - 1 = 7$ labels, then 7λ would have to be equal to $r(k - 1) = 6$. This would require that $\lambda = 6/7 = r(k - 1)/(v - 1)$. Since λ is not an integer, a balanced incomplete block design of this size cannot exist. However, for the size of design in Table 11.4, λ is an integer since $\lambda = r(k - 1)/(v - 1) = 7(3)/(7) = 3$.

There are three necessary conditions for the existence of a balanced incomplete block design, all of which are easy to check. These are

$$vr = bk\,,$$
$$r(k - 1) = \lambda(v - 1)\,, \tag{11.3.1}$$
$$b \geq v\,.$$

The first condition is satisfied by all block designs with equal replication and equal block sizes. The second condition is obtained by the argument above, and the third condition is called Fisher's inequality. Although the three necessary conditions can be used to verify that a balanced incomplete block design of a given size may exist, they do not guarantee its existence. Lists of balanced incomplete block designs can be found in Cochran and Cox (1957, chapter 11) and Fisher and Yates (1973), or can be obtained by some computer packages (see, for example, PROC OPTEX in the SAS software, described in Sect. 11.8.1 and the ibd package in R, Sect. 11.9.1).

11.3.2 Group Divisible Designs

A *group divisible design* is a design with $v = g\ell$ treatment labels (for some integers $g > 1$ and $\ell > 1$), each occurring r times, and bk experimental units grouped into b blocks of size $k < v$ in such a way that the units within a block are alike and units in different blocks are substantially different. The plan of the design satisfies the following conditions:

(i) The $v = g\ell$ treatment labels are divided into g groups of ℓ labels—any two labels within a group are called *first associates* and any two labels in different groups are called *second associates*.
(ii) Each treatment label appears either once or not at all in a block (that is, the design is binary).
(iii) Each pair of first associates appears together in λ_1 blocks.
(iv) Each pair of second associates appears together in λ_2 blocks.

Block design randomization is carried out as in Sect. 11.2.2.

It will be seen in Sect. 11.4.5 that the values of λ_1 and λ_2 govern the lengths of confidence intervals for treatment contrasts. Generally, it is preferable to have λ_1 and λ_2 as close as possible, which ensures that the confidence intervals of pairwise comparisons are of similar lengths. Group divisible designs with λ_1 and λ_2 differing by one are usually regarded as the best choice of incomplete block design when no balanced incomplete block design exists.

An example of a group divisible design (prior to randomization) is the experimental plan shown in Table 11.1, p. 350. It has the following $g = 4$ groups of $\ell = 2$ labels:

$$(1, 5), \quad (2, 6), \quad (3, 7), \quad (4, 8).$$

Labels in the same group (first associates) never appear together in a block, so $\lambda_1 = 0$. Labels in different groups (second associates) appear together in one block, so $\lambda_2 = 1$.

A second example is given by the experimental plan in Table 11.5, and it has $g = 4$ groups of $\ell = 3$ labels:

$$(1, 2, 3), \quad (4, 5, 6), \quad (7, 8, 9), \quad (10, 11, 12),$$

and $\lambda_1 = 3$, $\lambda_2 = 1$ (which is not ideal since λ_1 and λ_2 differ by more than 1).

There are four necessary conditions for the existence of a group divisible design with chosen values of $v = g\ell, b, k, r$ namely,

$$g\ell r = bk\,,$$
$$r(k - 1) = \lambda_1(\ell - 1) + \lambda_2 \ell(g - 1)\,,$$
$$r \geq \lambda_1\,,$$
$$rk \geq \lambda_2 v\,,$$

Table 11.5 A group divisible design with $v = 12$, $r = 3$, $b = 6$, $k = 6$, $\lambda_1 = 3$, $\lambda_2 = 1$

Block	Treatments					
I	1	2	3	4	5	6
II	1	2	3	7	8	9
III	1	2	3	10	11	12
IV	4	5	6	7	8	9
V	4	5	6	10	11	12
VI	7	8	9	10	11	12

for integers λ_1 and λ_2.

All group divisible designs with $\lambda_2 = 0$ should be avoided, since not all of the treatment contrasts are estimable. (It can be verified that the disconnected design of Table 11.3 is a group divisible design with groups (1, 3, 5, 7) and (2, 4, 6, 8) and with $\lambda_1 = 2$ and $\lambda_2 = 0$.) Lists of group divisible designs are given by Clatworthy (1973) and in the more recent references listed by Sinha (1991). Good block designs (which may or may not be group divisible) can be obtained by some computer packages (see, for example, `PROC OPTEX` in SAS software, Sect. 11.8.1 and the `ibd` package in R, Sect. 11.9.1).

11.3.3 Cyclic Designs

A *cyclic design* is a design with v treatment labels, each occurring r times, and with bk experimental units grouped into $b = v$ blocks of size $k < v$ in such a way that the units within a block are alike and units in different blocks are substantially different. The experimental plan, using treatment labels $1, 2, \ldots, v$, can be obtained as follows:

 (i) The first block, called the *initial block*, consists of a selection of k distinct treatment labels.
(ii) The second block is obtained from the initial block by *cycling the treatment labels*—that is, by replacing treatment label 1 with 2, 2 with 3, $\ldots$, $v - 1$ with v, and v with 1. The third block is obtained from the second block by cycling the treatment labels once more, and so on until the vth block is reached.

Block design randomization is carried out as in Sect. 11.2.2.

The group divisible design in Table 11.1 is also a cyclic design and has initial block (1, 3, 8). There are three cyclic designs in Table 11.6 all with block size $k = 4$. The first two have initial block (1, 2, 3, 6), but one has $v = 7$ treatment labels and the other has $v = 6$. The third design has initial block (1, 2, 3, 4) and $v = 7$. The first design is also a balanced incomplete block design with $\lambda = 2$. The second design has pairs of treatments occurring together in either $\lambda_1 = 2$ or $\lambda_2 = 3$ blocks, which results in only two possible lengths of confidence intervals for pairwise comparisons, but it is not a group divisible design since the treatment labels cannot be divided into groups of first associates. The third design is less good since pairs of treatments occur in $\lambda_1 = 1$ or $\lambda_2 = 2$ or $\lambda_3 = 3$ blocks, resulting in three different lengths of confidence interval for pairwise comparisons.

A cyclic design can have as many as $v/2$ different values of λ_i, yielding as many as $v/2$ different lengths of confidence intervals for pairwise comparisons of treatment effects. Again, if no balanced incomplete block design exists, the best designs are usually regarded as those with two values of λ_i which differ by one.

Table 11.6 Cyclic designs with $k = 4$ generated by (1, 2, 3, 6) for $v = 7$ and $v = 6$, and generated by (1, 2, 3, 4) for $v = 7$

| Design 1 | | Design 2 | | Design 3 | |
| $v = 7$ | | $v = 6$ | | $v = 7$ | |
Block	Treatments	Block	Treatments	Block	Treatments
1	1 2 3 6	1	1 2 3 6	1	1 2 3 4
2	2 3 4 7	2	2 3 4 1	2	2 3 4 5
3	3 4 5 1	3	3 4 5 2	3	3 4 5 6
4	4 5 6 2	4	4 5 6 3	4	4 5 6 7
5	5 6 7 3	5	5 6 1 4	5	5 6 7 1
6	6 7 1 4	6	6 1 2 5	6	6 7 1 2
7	7 1 2 5			7	7 1 2 3

Some cyclic designs have duplicate blocks, such as that with $v = 8$ and initial block (1, 4, 5, 8). These designs are useful when fewer than v blocks are required, since duplicate blocks can be ignored. Otherwise, designs with distinct blocks are usually better. Lists of cyclic designs are given by John et al. (1972), John (1981), and Lamacraft and Hall (1982).

11.4   Analysis of General Incomplete Block Designs

11.4.1 Contrast Estimators and Multiple Comparisons

The standard block–treatment model for the observation on treatment i in block h in a binary incomplete block design is

$$
\begin{aligned}
Y_{hi} &= \mu + \theta_h + \tau_i + \epsilon_{hi} \,, \\
\epsilon_{hi} &\sim N(0, \sigma^2) \,, \\
\epsilon_{hi}\text{'s are } & \text{mutually independent} \,, \\
h = 1, \ldots, b \,; \quad i = 1, &\ldots, v \,; \quad (h, i) \text{ in the design.}
\end{aligned}
\tag{11.4.2}
$$

The model, which assumes no block–treatment interaction, is almost identical to block–treatment model (10.4.1) for the randomized block design. The only difference is the phrase "(h, i) in the design", which means that the model is applicable only to those combinations of block h and treatment i that are actually observed. The phrase serves as a reminder that not all treatments are observed in each block.

For every experiment, the assumptions on the model should be checked. However, when blocks do not contain every treatment, it is difficult to check the assumption of no block–treatment interaction by plotting the data block by block, as was recommended for complete block designs in Sect. 10.7. Thus, it is preferable that an incomplete block design be used only when there are good reasons for believing that treatment differences do not depend on the level of the blocking factor(s).

The least squares estimators for the treatment parameters in the model for an incomplete block design must include an adjustment for blocks, since some treatments may be observed in "better" blocks than others. This means that the least squares estimator for the pairwise comparison $\tau_p - \tau_i$ is *not* the *unadjusted estimator* $\overline{Y}_{.p} - \overline{Y}_{.i}$ as it would be for a randomized complete block design. For example, if metal alloys 2 and 7 were to be compared via the balanced incomplete block design in the

last column of Table 11.2, we see that alloy 2 is observed on days 1, 2, and 4, and alloy 7 is observed on days 1, 6, and 7. If we were to use $\overline{Y}_{.2} - \overline{Y}_{.7}$ to estimate $\tau_2 - \tau_7$, it would be biased, since

$$
\begin{aligned}
E[\overline{Y}_{.2} - \overline{Y}_{.7}] &= E[\frac{1}{3}(Y_{12} + Y_{22} + Y_{42}) - \frac{1}{3}(Y_{17} + Y_{67} + Y_{77})] \\
&= \frac{1}{3}(3\mu + \theta_1 + \theta_2 + \theta_4 + 3\tau_2) - \frac{1}{3}(3\mu + \theta_1 + \theta_6 + \theta_7 + 3\tau_7) \\
&= (\tau_2 - \tau_7) + \frac{1}{3}(\theta_2 + \theta_4 - \theta_6 - \theta_7) \\
&\neq (\tau_2 - \tau_7).
\end{aligned}
$$

If the experimental conditions were to change over the course of the experiment in such a way that observations on the first few days tended to be higher than observations on the last few days, then θ_2 and θ_4 would be larger than θ_6 and θ_7. If the two alloys do not differ in their tensile strengths, then $\tau_2 = \tau_7$, but the above calculation shows that $\overline{Y}_{.2} - \overline{Y}_{.7}$ would nevertheless be expected to be large. This could cause the experimenter to conclude erroneously that alloy 2 was stronger than alloy 7. Consequently, any estimator for $\tau_2 - \tau_7$ must contain an adjustment for the days on which the alloys were observed.

A general formula for a set of least squares solutions for the parameters τ_i in the block–treatment model (11.4.2) adjusted for block differences can be shown to be

$$
r(k-1)\hat{\tau}_i - \sum_{p \neq i} \lambda_{pi} \hat{\tau}_p = k Q_i, \quad \text{for } i = 1, \ldots, v, \tag{11.4.3}
$$

when $\sum_i \hat{\tau}_i = 0$, $\sum_h \hat{\theta}_h = 0$ are used as the added equations (similar to Sect. 3.4.3). Here, λ_{pi} is the number of blocks containing both treatments p and i, and a formula for calculating Q_i will be given in (11.4.7) in the next section. However, except in special cases, such as that of the balanced incomplete block design, expression (11.4.3) is difficult to solve for the individual τ_i and is usually left for statistical software to calculate. Although the individual τ_i are not uniquely estimable, all contrasts $\sum_i c_i \tau_i$ are estimable if the design is connected, (see Sect. 11.2.3).

The Bonferroni and Scheffé methods of multiple comparisons can be used for simultaneous confidence intervals of estimable contrasts in all incomplete block designs. The method of Tukey is applicable for balanced incomplete block designs, and it is believed to be conservative (true value of α smaller than stated) for other incomplete block designs, but this has not yet been proven. Dunnett's method can be used in balanced incomplete block designs but not in other incomplete block designs without modification to our tables. For each method, the formula for a set of $100(1-\alpha)\%$ simultaneous intervals is

$$
\sum_{i=1}^{v} c_i \tau_i \in \left(\sum_{i=1}^{v} c_i \hat{\tau}_i \pm w \sqrt{\widehat{\text{Var}}\left(\sum_{i=1}^{v} c_i \hat{\tau}_i \right)} \right) \tag{11.4.4}
$$

exactly as in Sect. 4.4. The correct least squares estimate $\sum c_i \hat{\tau}_i$, as well as the estimated variance $\widehat{\text{Var}}(\sum c_i \hat{\tau}_i)$ and the error degrees of freedom can be obtained from computer software for the design being used.

Table 11.7 Analysis of variance table for a binary incomplete block design with b blocks of size k, and v treatment labels appearing r times

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio
Blocks (adj)	$b-1$	$ss\theta_{\text{adj}}$	$ms\theta_{\text{adj}}$	–
Blocks (unadj)	$b-1$	$ss\theta$	–	–
Treatments (adj)	$v-1$	ssT_{adj}	msT_{adj}	$\dfrac{msT_{\text{adj}}}{msE}$
Error	$bk-b-v+1$	ssE	msE	
Total	$bk-1$	$sstot$		

<div align="center">Formulae</div>

$ss\theta = \sum_{h=1}^{b} B_h^2/k - G^2/(bk)$ $\qquad\qquad$ $ssE = sstot - ss\theta - ssT_{\text{adj}}$

$ssT_{\text{adj}} = \sum_{i=1}^{v} Q_i \hat{\tau}_i$ $\qquad\qquad$ $sstot = \sum_{h=1}^{b} \sum_{i=1}^{v} n_{hi} y_{hi}^2 - G^2/(bk)$

$Q_i = T_i - \sum_{h=1}^{b} n_{hi} B_h / k$. $\qquad\qquad$ $ss\theta_{\text{adj}} = sstot - ssE - \left(\sum_{i=1}^{v} T_i^2/r - G^2/(bk) \right)$

11.4.2 Analysis of Variance

For a connected incomplete block design with block–treatment model (11.4.2), the four rows in the center section of Table 11.7 show an outline of the analysis of variance obtained when the block factor is entered into the model before the treatment factor; the formulae are listed in the bottom section of the table. The unadjusted sum of squares for blocks, $ss\theta$, is calculated in a similar way to the block sum of squares in a complete block design; that is

$$ss\theta = \sum_{h=1}^{b} B_h^2/k - G^2/(bk) \,, \qquad (11.4.5)$$

where B_h is the sum of all observations in block h, and G represents the "grand total" of all the observations. The *sum of squares for treatments adjusted for blocks* (i.e. adjusted for the fact that not all treatments are in every block and some blocks are better than others) is

$$ssT_{\text{adj}} = \sum_{i=1}^{v} Q_i \hat{\tau}_i \,, \qquad (11.4.6)$$

where $\hat{\tau}_i$ is the least squares solution for τ_i obtained from (11.4.3), and Q_i is the ith *adjusted treatment total*; that is,

$$Q_i = T_i - \frac{1}{k} \sum_{h=1}^{b} n_{hi} B_h \,, \qquad (11.4.7)$$

where n_{hi} is 1 if treatment i is observed in block h and zero otherwise, B_h is defined above, and T_i is the sum of all observations on treatment i. We could write the three quantities T_i, B_h and G as $y_{.i}$ and $y_{h.}$ and $y_{..}$ in the usual way. The reason for changing notation is as a reminder that some of the y_{hi} are not actually observed and the quantities T_i, B_h and G are more accurately written as

$$T_i = \sum_{h=1}^{b} n_{hi} y_{hi} \,, \quad B_h = \sum_{i=1}^{v} n_{hi} y_{hi} \,, \quad \text{and} \quad G = \sum_{h=1}^{b} \sum_{i=1}^{v} n_{hi} y_{hi} \,.$$

The sum of squares for error, ssE, is

$$ssE = sstot - ss\theta - ssT_{\text{adj}} \,,$$

where $sstot$ is defined as usual as

$$sstot = \sum_{h=1}^{b} \sum_{i=1}^{v} n_{hi} y_{hi}^2 - G^2/(bk) \,.$$

Also as usual, the number of degrees of freedom for treatments is $v - 1$, the number of degrees of freedom for blocks is $b - 1$, and the number of degrees of freedom for error can be obtained by subtraction as

$$df = (n - 1) - (b - 1) - (v - 1) = bk - b - v + 1 \,, \tag{11.4.8}$$

where $n = bk$ is the total number of observations.

A test of H_0^T :{all τ_i are equal} against H_A^T :{at least two of the τ_i's differ} is given by the decision rule

$$\text{reject } H_0^T \quad \text{if} \quad \frac{msT_{\text{adj}}}{msE} > F_{v-1,bk-b-v+1,\alpha}$$

for some chosen significance level α, where $msT_{\text{adj}} = ssT_{\text{adj}}/(v-1)$, and $msE = ssE/(bk-b-v+1)$.

If evaluation of blocking for the purpose of planning future experiments is required, the quantity $ss\theta$ in (11.4.5) *is not the correct value* to use. It has not been adjusted for the fact that every block does not contain an observation on every treatment. In order to evaluate blocks, we would need the *adjusted block sum of squares*, $ss\theta_{\text{adj}}$, whose formula is listed as the last entry in Table 11.7. Some computer packages will give this value under the heading "adjusted" or "Type III" sum of squares. If the program does not automatically generate the adjusted value, it can be obtained from a "sequential sum of squares" by entering treatments in the model before blocks.

11.4.3 Analysis of Balanced Incomplete Block Designs

The set of least squares solutions given in (11.4.3) for the treatment parameters τ_i in the block–treatment model (11.4.2) for the balanced incomplete block design have a simple form:

$$\hat{\tau}_i = \frac{k}{\lambda v} Q_i \,, \quad \text{for} \quad i = 1, \dots, v \,, \tag{11.4.9}$$

where λ is the number of times that every pair of treatments occurs together in a block, and Q_i is the adjusted treatment total, given in (11.4.7). Thus, the sum of squares for treatments adjusted for blocks in (11.4.6) becomes

$$ssT_{\text{adj}} = \sum_{i=1}^{v} \frac{k}{\lambda v} Q_i^2 \,, \tag{11.4.10}$$

and the least squares estimator of contrast $\sum c_i \tau_i$ is

$$\sum_{i=1}^{v} c_i \hat{\tau}_i = \frac{k}{\lambda v} \sum_{i=1}^{v} c_i Q_i \,. \tag{11.4.11}$$

It can be shown that the corresponding variance can be calculated as

$$\text{Var}\left(\sum_{i=1}^{v} c_i \hat{\tau}_i\right) = \sum_{i=1}^{v} c_i^2 \left(\frac{k}{\lambda v}\right) \sigma^2 . \tag{11.4.12}$$

The Bonferroni, Scheffé, Tukey, and Dunnett methods of multiple comparisons can all be used for balanced incomplete block designs with degrees of freedom for error equal to $df = bk - b - v + 1$. The general formula (11.4.4) for simultaneous $100(1 - \alpha)\%$ confidence intervals for a set of contrasts $\Sigma c_i \tau_i$ becomes

$$\sum_{i=1}^{v} c_i \tau_i \in \left(\frac{k}{\lambda v} \sum_{i=1}^{v} c_i Q_i \pm w \sqrt{\sum_{i=1}^{v} c_i^2 \left(\frac{k}{\lambda v}\right) msE}\right), \tag{11.4.13}$$

where the critical coefficients for the four methods are, respectively,

$$w_B = t_{bk-b-v+1, \alpha/2m} \; ; \;\; w_S = \sqrt{(v-1)F_{v-1, bk-b-v+1, \alpha}} \; ;$$
$$w_T = q_{v, bk-b-v+1, \alpha}/\sqrt{2} \; ; \;\; w_{D2} = |t|_{v-1, bk-b-v+1, \alpha}^{(0.5)} \; .$$

For testing m hypotheses of the general form $H_0 : \sum c_i \tau_i = 0$ against the corresponding alternative hypotheses $H_a : \sum c_i \tau_i \neq 0$, at overall significance level α, the decision rule using Bonferroni's method for preplanned contrasts is

$$\text{reject } H_0 \text{ if } \frac{ssc_{\text{adj}}}{msE} > F_{1, bk-b-v+1, \alpha/m} , \tag{11.4.14}$$

and the decision rule using Scheffé's method is

$$\text{reject } H_0 \text{ if } \frac{ssc_{\text{adj}}}{msE} > (v-1)F_{1, bk-b-v+1, \alpha} , \tag{11.4.15}$$

where

$$\frac{ssc_{\text{adj}}}{msE} = \frac{(\sum c_i \hat{\tau}_i)^2}{\left(\frac{k}{\lambda v}\right)(\sum c_i^2) \, msE} = \frac{k(\sum c_i Q_i)^2}{\lambda v (\sum c_i^2) msE} , \tag{11.4.16}$$

(cf. Sections 4.3.3 and 6.7.2).

As for the case of equireplicate completely randomized designs, two contrasts $\Sigma c_i \tau_i$ and $\Sigma d_i \tau_i$ are orthogonal in a balanced incomplete block design if $\Sigma c_i d_i = 0$. The adjusted treatment sum of squares can then be written as a sum of adjusted contrast sums of squares for a complete set of $(v - 1)$ orthogonal contrasts. An example is given in the following section.

11.4.4 A Real Experiment—Detergent Experiment

An experiment to compare dishwashing detergent formulations was described by P. W. M. John in the journal *Technometrics* in 1961. The experiment involved three base detergents and an additive. Detergent I was observed with 3, 2, 1, and 0 parts of the additive, giving four treatments, which we will code 1, 2, 3, and 4. Likewise, Detergent II was observed with 3, 2, 1, and 0 parts of the additive, giving an additional four treatments, which we will code 5, 6, 7, and 8. The standard detergent (Detergent III) with no additive served as a control treatment, which we will code as 9.

   The experiment took place in a location where three sinks were available. Three people took part in the experiment and were instructed to wash plates at a common rate. An observation was the number of plates washed in a sink before the detergent foam disappeared. A block consisted of three observations, one per sink. The refilling of the three sinks with water and detergent constituted the beginning of a new block. The amount of soil on the plates prior to washing was held constant. Differences between blocks were due to differences in the common washing rates, the water temperature, the experimenter fatigue, etc.

   A design was required with blocks of size $k = 3$ and $v = 9$ treatment labels. A balanced incomplete block design was selected with $b = 12$ blocks giving $r = bk/v = 4$ observations per treatment and every pair of treatment labels occurring in $\lambda = r(k-1)/(v-1) = 1$ block. We have shown a possible randomization of this design in Table 11.8 together with the data from the original article. The positions within a block show the allocations of the three basins to treatments. The observations are plotted against treatment in Fig. 11.2, ignoring the block from which the observation was collected.

   Since each pair of treatments occurs together in only one block ($\lambda = 1$), a graphical approach for the evaluation of block–treatment interaction cannot be used. However, it appears from Fig. 11.2 that block differences, block–treatment interaction effects, and random error variability must all be rather small compared with the large detergent differences. A plot of the adjusted data is described below.

Table 11.8   Design and number of plates washed for the detergent experiment

Block	Treatments			Plates washed		
1	3	8	4	13	20	7
2	4	9	2	6	29	17
3	3	6	9	15	23	31
4	9	5	1	31	26	20
5	2	7	6	16	21	23
6	6	5	4	23	26	6
7	9	8	7	28	19	21
8	7	1	4	20	20	7
9	6	8	1	24	19	20
10	5	8	2	26	19	17
11	5	3	7	24	14	21
12	3	2	1	11	17	19

Source John (1961). Copyright © 1961 American Statistical Association. Reprinted with permission

Fig. 11.2   Data plot for the detergent experiment, ignoring block levels
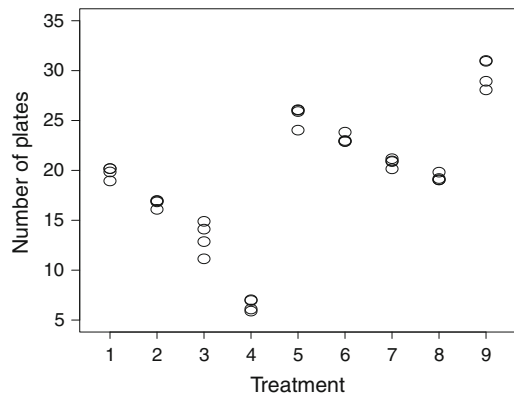
Fig. 11.3 Plot of adjusted observations for the detergent experiment
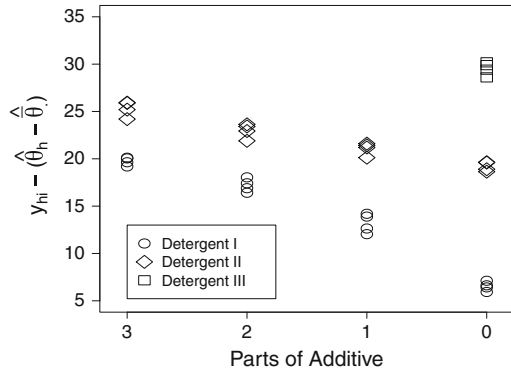
Block–treatment model (11.4.2) for an incomplete block design was fitted to the data. The residual plots might lead us to question some of the error assumptions, but there are only 4 observations per treatment and these are all from different blocks, so it is difficult to make a proper assessment. We will proceed with the standard analysis for a balanced incomplete block design, recognizing that the stated significance levels and confidence levels are only approximate.

Plotting the Data Adjusted for Block Effects

In this detergent experiment, the treatment differences are fairly clear from the plot of the raw data in Fig. 11.2. However, if the block effects had been substantial, such a plot of the raw data could have painted a muddled picture. In such cases, the picture can be substantially improved by adjusting each observation for the block effects before plotting. The observation y_{hi} is adjusted for the block effects as follows,

$$y_{hi}^* = y_{hi} - (\hat{\theta}_h - \overline{\hat{\theta}}_.) \,,$$

where $(\hat{\theta}_h - \overline{\hat{\theta}}_.)$ is the least squares estimate of $(\theta_h - \overline{\theta}_.)$. A SAS program that adjusts the observations for block effects and plots the adjusted observations is given in Sect. 11.8.3 and a corresponding R program in Sect. 11.9.3. It should be noted that since the variability due to block effects has been extracted, a plot of the adjusted observations will appear to exhibit less variability than really exists.

For this particular data set, the block differences are very small, so a plot of the adjusted data would provide information similar to that provided by the plot of the raw data in Fig. 11.2. In Fig. 11.3, the observations adjusted for blocks are plotted against "parts of additive" for each base detergent. It appears that the washing power decreases almost linearly as the amount of additive is decreased and also that the standard detergent is superior to the two test detergents.

Analysis

The analysis of variance table, given in Table 11.9, shows the treatment sum of squares and its decomposition into sums of squares for eight orthogonal contrasts. These contrasts are the linear, quadratic, and cubic trends for each of detergents I and II (as the amount of additive decreases), together with the "I Versus II" contrast that compares the effects of detergents I and II averaged over the levels of the additive, and the "control Versus others" contrast comparing the effect of the control detergent and the average effect of the other eight treatments. For example, the linear trend contrast for detergent I is $-3\tau_1 - \tau_2 + \tau_3 + 3\tau_4$, where the contrast coefficients are obtained from Table A.2. The contrast comparing detergents I and II is the difference of averages contrast

Table 11.9 Analysis of variance table for the detergent experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Blocks (adj)	11	10.06	0.91	–	–
Blocks (unadj)	11	412.75	–	–	–
Treatments(adj)	8	1086.81	135.85	164.85	0.0001
I linear	1	286.02	286.02	347.08	0.0001
I quadratic	1	12.68	12.68	15.38	0.0012
I cubic	1	0.22	0.22	0.27	0.6092
II linear	1	61.34	61.34	74.44	0.0001
II quadratic	1	0.15	0.15	0.18	0.6772
II cubic	1	0.03	0.03	0.04	0.8520
I vs II	1	381.34	381.34	462.75	0.0001
Control vs others	1	345.04	345.04	418.70	0.0001
Error	16	13.19	0.82		
Total	35	1512.75			

$$\frac{1}{4}(\tau_1 + \tau_2 + \tau_3 + \tau_4) - \frac{1}{4}(\tau_5 + \tau_6 + \tau_7 + \tau_8),$$

and the contrast comparing the control detergent with the others is the difference of averages contrast

$$\tau_9 - \frac{1}{8}(\tau_1 + \tau_2 + \tau_3 + \tau_4 + \tau_5 + \tau_6 + \tau_7 + \tau_8).$$

A set of simultaneous 99% confidence intervals for all treatment contrasts using Scheffé's method of multiple comparisons is given by (11.4.13) with $w_S = \sqrt{8F_{8,16,.01}}$, and $F_{8,16,.01} = 3.89, k = 3, \lambda = 1$ and $v = 9$. Using the data shown in Table 11.8, we have treatment totals and grand total

$$T_1 \ T_2 \ T_3 \ T_4 \ T_5 \ T_6 \ T_7 \ T_8 \ T_9 \ \ G$$
$$79 \ 67 \ 53 \ 26 \ 102 \ 93 \ 83 \ 77 \ 119 \ 699$$

and block totals

$$B_1 \ B_2 \ B_3 \ B_4 \ B_5 \ B_6 \ B_7 \ B_8 \ B_9 \ B_{10} \ B_{11} \ B_{12}$$
$$40 \ 52 \ 69 \ 77 \ 60 \ 55 \ 68 \ 47 \ 63 \ \ 62 \ \ 59 \ \ 47$$

Then, from (11.4.7), the first adjusted treatment total is

$$Q_1 = T_1 - \frac{1}{k}[B_4 + B_8 + B_9 + B_{12}] = 79 - \frac{1}{3}[234] = 1.0.$$

The other adjusted treatment totals are calculated similarly, giving

$$Q_1 \ \ \ Q_2 \ \ \ \ \ Q_3 \ \ \ \ \ \ Q_4 \ \ \ \ \ Q_5 \ \ \ \ Q_6 \ \ \ Q_7 \ \ \ \ Q_8 \ \ \ Q_9$$
$$1.00 \ -6.67 \ -18.67 \ -38.67 \ 17.67 \ 10.67 \ 5.00 \ -0.67 \ 30.33$$

and since $k/(\lambda v) = 3/9$, the least squares estimate of the contrast $\Sigma c_i \tau_i$ given by (11.4.11) is $\Sigma c_i \hat{\tau}_i = \Sigma c_i Q_i / 3$ ($i = 1, 2, \ldots, 9$). For example, the least squares estimate for the "control versus others"

contrast is

$$\sum_{i=1}^{9} c_i \hat{\tau}_i = \hat{\tau}_9 - \frac{1}{8} \sum_{i=1}^{8} \hat{\tau}_i = \frac{1}{3} \left(Q_9 - \frac{1}{8} \sum_{i=1}^{8} Q_i \right) = 11.375 \,,$$

with associated estimated variance

$$\widehat{\text{Var}} \left(\sum_{i=1}^{9} c_i \hat{\tau}_i \right) = \sum_{i=1}^{9} c_i^2 \left(\frac{k}{\lambda v} \right) msE = \left(1 + \frac{8}{64} \right) \left(\frac{3}{9} \right) (0.82) = 0.3075 \,,$$

where $msE = 0.82$ is obtained from Table 11.9. Using Scheffé's method of multiple comparisons at overall level 99%, a confidence interval for the control versus others contrast is then

$$11.375 \pm \sqrt{8F_{8,16,0.01}} \sqrt{0.3075} = 11.375 \pm 3.093 = (8.282, 14.468),$$

showing that the control detergent washed between 8.3 and 14.5 more plates than the other detergents on average.

The sum of squares for treatments adjusted for blocks is obtained from (11.4.6), p. 358, as

$$ssT_{\text{adj}} = \frac{k}{\lambda v} \sum Q_i^2 = 1086.81 \,,$$

and since

$$\frac{msT_{\text{adj}}}{msE} = \frac{1086.81/8}{0.82} = 164.85 > F_{8,16,0.01} = 3.89 \,,$$

we reject the hypothesis of no treatment differences.

The eight orthogonal contrasts can be tested simultaneously using the method of Scheffé. For example, the confidence interval for the "control versus others" contrast calculated above as part of a 99% simultaneous set of intervals does not contain zero, so the hypothesis that the control treatment does not differ from the others would be rejected. The overall significance level for all such tests would be $\alpha = 0.01$. Equivalently, the contrasts can be tested by the Scheffé method using the decision rule (11.4.15), p. 360; that is,

$$\text{reject } H_0 : \sum_{i=1}^{v} c_i \tau_i = 0 \text{ if } \frac{ssc_{\text{adj}}}{msE} > 8F_{8,df,.01} = 31.12 \,.$$

The ratios ssc_{adj}/msE are provided in Table 11.9. Comparing their values with 31.12, we see that the linear trends are significantly different from zero for each of the base detergents I and II, as are the comparison of detergents I and II on average and the comparison of the control detergent with the average effects of the other 8 treatments. From significance of the linear trend contrasts, coupled with the direction of the trends, one can conclude that detergents I and II are better with larger amounts of additive.

We cannot use the unadjusted block sum of squares to evaluate the usefulness of blocking. We would need to calculate the adjusted block sum of squares as in Table 11.7, p. 358. Using the values in Table 11.9, this is

$$ss\theta_{\text{adj}} = 1512.75 - 13.19 - \left(\frac{1}{4}(79^2 + 67^2 + \cdots + 119^2) - \frac{1}{36}699^2 \right) = 10.06 \,.$$

So the adjusted block mean square is $ms\theta_{\text{adj}} = 10.06/11 = 0.91$, which is not much larger than the error mean square, $msE = 0.82$, so the blocking did not help with increasing the power of the hypothesis tests. Nevertheless, it was natural to design this experiment as a block design, and the creation of blocks was a wise precaution against changing experimental conditions.

11.4.5 Analysis of Group Divisible Designs

Group divisible designs were described in Sect. 11.3.2, p. 354, and illustrations were shown in Tables 11.1 and 11.5. The least squares solution (obtained with added equations $\sum_i \hat{\tau}_i = 0, \sum_h \hat{\theta}_h = 0$) for the treatment parameters τ_i adjusted for blocks can be shown to be

$$\hat{\tau}_i = \left[\frac{k}{(r(k-1) + \lambda_1)v\lambda_2} \right] \times \left[(v\lambda_2 + (\lambda_1 - \lambda_2))Q_i + (\lambda_1 - \lambda_2) \underset{(1)}{\sum} Q_p \right],$$

$$\tag{11.4.17}$$

where Q_i is the adjusted treatment total as in (11.4.7), p. 358, and where $\sum_{(1)} Q_p$ denotes the sum of the Q_p corresponding to the treatment labels that are the first associates of treatment label i.

The variance of the least squares estimator $\Sigma c_i \hat{\tau}_i$ of an estimable contrast $\Sigma c_i \tau_i$ is

$$\text{Var}\left(\sum_{i=1}^{v} c_i \hat{\tau}_i \right) = \sum_{i=1}^{v} c_i^2 \, \text{Var}(\hat{\tau}_i) + 2 \sum_{i=1}^{v-1} \sum_{p=i+1}^{v} c_i c_p \, \text{Cov}(\hat{\tau}_i, \hat{\tau}_p) \,,$$

where, for a group divisible design,

$$\text{Var}(\hat{\tau}_i) = \frac{k[v\lambda_2 + (\lambda_1 - \lambda_2)]}{v\lambda_2[v\lambda_2 + l(\lambda_1 - \lambda_2)]} \sigma^2$$

and

$$\text{Cov}(\hat{\tau}_i, \hat{\tau}_p) = \begin{cases} \frac{k(\lambda_1 - \lambda_2)\sigma^2}{v\lambda_2[v\lambda_2 + l(\lambda_1 - \lambda_2)]} \,, & \text{if } i \text{ and } p \text{ are first associates,} \\ 0 \,, & \text{if } i \text{ and } p \text{ are second associates.} \end{cases}$$

Using these quantities, the variance of the least squares estimator of the pairwise comparison $\tau_i - \tau_p$ becomes

$$\text{Var}(\hat{\tau}_i - \hat{\tau}_p) = \begin{cases} \frac{2k\sigma^2}{[v\lambda_2 + l(\lambda_1 - \lambda_2)]} \,, & \text{if } i \text{ and } p \text{ are first associates,} \\ \frac{2k[v\lambda_2 + (\lambda_1 - \lambda_2)]\sigma^2}{v\lambda_2[v\lambda_2 + l(\lambda_1 - \lambda_2)]} \,, & \text{if } i \text{ and } p \text{ are second associates.} \end{cases}$$

If λ_1 and λ_2 are as close in value as possible, then the variances for the pairwise comparisons will be as close as possible.

The Bonferroni and Scheffé methods of multiple comparisons can be used for group divisible designs. The Tukey method is believed to be conservative (true value of α smaller than stated). The Dunnett method is not available using our tables, since the critical values can be used only for designs in which $\text{Cov}(\hat{\tau}_i, \hat{\tau}_p)$ are equal for all i and p. However, Dunnett intervals can be obtained from many computer packages (see, for example, Fig 11.7, p. 380, and Table 11.24, p. 387). The analysis of variance table for the group divisible design is that given in Table 11.7, p. 358, with $\hat{\tau}_i$ as in (11.4.17)

above. An example of an experiment designed as a group divisible design is discussed in Sect. 11.5. SAS and R programs are illustrated in Sects. 11.8 and 11.9 that can be used to analyze any group divisible design.

11.4.6 Analysis of Cyclic Designs

Cyclic designs were described in Sect. 11.3.3, p. 355, and illustrated in Tables 11.1 and 11.6. They are incomplete block designs that may or may not possess the properties of balanced incomplete block designs or group divisible designs. When they do not possess these properties, the least squares solutions $\hat{\tau}_i$ have no simple form and are most easily obtained from computer software. The Bonferroni and Scheffé methods of multiple comparisons can be used for all cyclic designs.

In Sect. 11.5 we reproduce the checklist and analysis of an experiment that was designed as a cyclic group divisible design and, in Sects. 11.8 and 11.9, we illustrate SAS and R computer programs that can be used to analyze any cyclic incomplete block design.

11.5   A Real Experiment—Plasma Experiment

The plasma experiment was run by Ernesto Barrios, Jin Feng, and Richard Kibombo in 1992 in the Engineering Research Center at the University of Wisconsin. The following checklist has been extracted from the experimenters' report. The design used was a cyclic group divisible design. Notice that the experimenters moved step (e) of the checklist forward. They had made a list of all potential sources of variation, but they needed a pilot experiment to help determine which sources they could control and which they could not.

Checklist

(a) **Define the objectives of the experiment**.

In physics, plasma is an ionized gas with essentially equal densities of positive and negative charges. It has long been known that plasma can effect desirable changes in the surface properties of materials.
The purpose of this experiment is to study the effects of different plasma treatments of plastic pipet tips on the capillary action of the pipets. Capillary action concerns the movement of a liquid up the pipet—a small tube. Before a plasma treatment, the capillarity conduct of the tips is too narrow to permit water to move up. Changes in capillary action effected by plasma treatment can be measured by suspending the tip of a vertical pipet into a bed of water and measuring the height of the column of water in the tube.

(e) **Run a pilot experiment**.

At this stage we decided to make a test run to become familiar with the process of setting up and running the experiment, to determine the appropriate treatment factor levels, and to help identify the major sources of variation that could be controlled, and to identify other variables that might affect the response but which could not be controlled.

(b) **Identify all sources of variation.**

From the test run, we determined that pressure and voltage could not both be effectively controlled. More generally, it would be difficult to vary all of the variables initially listed (gas flow rate, type of gas, pressure, voltage, presence or absence of a ground shield, and exposure time of the pipet tips to the ionized gas).

The following factors were potential sources of variation.

- Experimenters. Despite the fact that all of the experimenters were to play certain roles during each run of the experiment, it was noted that most of the variation due to the personnel could be attributed to the person who connects the pipet tips to the gas tube in the ionization chamber and takes the readings of the final response using vernier calipers.
- Room conditions. It was thought that variations in both room temperature and atmospheric pressure could have an effect on response.
- Water purity. If the water used to measure the capillarity has a substantial amount of impurities, especially mineral salts, then the response may be greatly affected, either because of variability in cohesion and adhesion forces of different types of substances, or because of a reaction between the impurities (salts) and the pipet tips.
- Materials. Variability in the quality of both the pipet tips and the gases used is likely to introduce some variation in the response. Within an enclosed room such as a laboratory, the composition of air may vary significantly over time.

Taking into account the results of the pilot run, the following decisions were made.

(i) Treatment factors and their levels.

Scale down the variables of interest to three by keeping both the pressure and voltage constant at 100 mm Torres and 5 volts, respectively, and by keeping the ground shield on. Distilled water will be used to control for impurities in the water. Pipet tips from a single package will be used, so the pipets are more likely to be from the same batch and hence more likely to be homogeneous. The only factors that will make up the various treatment combinations are gas flow rate, type of gas, and exposure time. No attempt will be made to control for variation in the composition or purity of the gases used. (This variation will be subsumed into the error variability).

Set the lower and upper levels of each factor far apart in order to make any (linear) effect more noticeable. Also, we decided to include only 6 of the 8 possible treatment combinations, as shown in Table 11.10.

(ii) Experimental units.

The experimental units are the (combinations of) pipets and time order of observations.

(iii) Blocking factors, noise factors, and covariates.

The two blocking factors are "experimenter" and "day."

(No covariates or noise factors were included.)

(c) **Specify a rule by which to assign the experimental units to the treatments**.

The design will be an incomplete block design with blocks of size three, and three blocks of data will be collected on each of two days. We will use the cyclic design for $v = 6 = b$ and $k = 3 = r$ generated by the treatment labels 1, 4, 5. The labels in the design will be randomly assigned to the six treatment combinations, and the treatments within each block will be randomly ordered.

(The selected cyclic design is shown in Table 11.11 in nonrandomized order so that the cyclic nature of the design can be seen more easily. The design also happens to be a group divisible design (see Exercise 14). The smallest balanced incomplete block design with $v = 6$ and $k = 3$ has $r = 5$ and $b = 10$ and would require more observations.)

(d) **Specify the measurements to be made, the experimental procedure, and the anticipated difficulties**.

The height of the water column will be measured for each pipet. In order to make the measurements as uniform as possible, a device has been constructed consisting of a rectangular sheet of plexiglass with a small hole in which to place the pipet tip. Placing this pipet holder on a water vessel suspends

Table 11.10 The six treatment combinations used in the plasma experiment

Treatment	Factors and levels		
	Type of gas	exposure time (sec)	Gas flow rate (cc/sec)
1	Argon	180	10
2	Air	180	10
3	Argon	180	30
4	Argon	60	30
5	Air	60	30
6	Air	60	10

Table 11.11 Design and data for the plasma experiment

Block	Day	Experimenter	Response (Treatment)		
1	1	Feng	0.482 (1)	0.459 (4)	0.458 (5)
2	1	Barrios	0.464 (2)	0.465 (5)	0.467 (6)
3	1	Kibombo	0.473 (3)	0.472 (6)	0.495 (1)
4	2	Feng	0.283 (4)	0.325 (1)	0.296 (2)
5	2	Barrios	0.410 (5)	0.390 (2)	0.248 (3)
6	2	Kibombo	0.384 (6)	0.239 (3)	0.350 (4)

about 2 mm of the tip of the pipet into the water. After 60 seconds, a mark will be made on the pipet indicating the water level reached. The distance of the mark from the tip of the pipet will be measured using a vernier caliper with tenth of a millimeter precision.

The experimental procedure for each observation is as follows: Place a pipet on the tube through which the plasma will flow, screw in a glass tube, turn on the pump and wait 40 seconds, open the Baratron, open the gas, turn a controller to auto, set the flow to a specified level, turn the pressure controller to auto and set the level, set the voltage, time the treatment, turn off flow and shut off the gas, set the pressure to open, wait until the pressure is less than 20, turn off the Baratron, turn off the pump, unscrew the glass tube, then (wearing a glove) take out the pipet, place the pipet in water (using the device for this purpose), and mark the height of the water column, then go on to the next observation.

Anticipated difficulties: Differences in the way people would mark or measure the water column heights would cause variation. Running the experiment consistently.

(f) **Specify the model**.

   (The standard block–treatment model (11.4.2), p. 356, for an incomplete block design was specified.)

(g) **Outline the analysis**.

   An analysis of variance test for equality of the treatment effects will be performed. Then confidence intervals for all pairwise comparisons will be obtained, with a simultaneous 95% confidence level using the method of Scheffé. Model assumptions will be evaluated.

(h) **Calculate the number of observations to be taken**.

   (The number of observations was limited by the time available.)

(i) **Review the above decisions. Revise if necessary**.

   (No revisions were made at this stage.)

Table 11.12   Analysis of variance table for the plasma experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Blocks (adj)	5	0.0805	0.0161	–	–
Blocks	5	0.0992	–	–	–
Treatments (adj)	5	0.0196	0.0039	2.99	0.0932
Error	7	0.0092	0.0013		
Total	17	0.1279			

Table 11.13   Analysis of variance for the plasma experiment—day one only

Source	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Blocks (adj)	2	0.0001213	0.0000607	364.00	–
Block	2	0.0004029	–	–	–
Treatments (adj)	5	0.0007112	0.0001422	853.40	0.026
Error	1	0.0000002	0.0000002		
Total	8	0.0011142			

Results of the Experiment

During the experiment, an unexpected event occurred. A little tube through which the gas passes was broken, allowing for some leaking of gas. We realized this after our first day's runs and tried to fix this problem the next day, using tape, as a new tube was unavailable. As can be seen from the results, given in Table 11.11, the responses from the last nine runs, corresponding to the second day, were consistently smaller than those from the first nine runs. This underscores the advantage of using time as a blocking factor.

Data Analysis

The analysis of variance is given in Table 11.12. It was obtained via a SAS computer program similar to the one in Table 11.18 in Sect. 11.8. The adjusted block mean square, $ms\theta_{adj} = 0.0161$, is twelve times larger than the error mean square, so blocking was certainly worthwhile.

The ratio $msT_{adj}/msE = 2.99$ does not exceed the critical value $F_{5,7,.05} = 3.97$ for testing equality of the treatment effects at the 5% significance level (equivalently, the p-value is greater than 0.05). Based on this result, examination of any individual treatment contrasts may seem unwarranted. However, the broken tube discovered after the first day of runs is an important consideration in this experiment. It is quite possible that the treatments are not the same on day one as on day two, since the broken tube may change the gas flow rate or the type of gas to which the pipet is exposed. So, one must ask the question, "Is there anything to be salvaged from this experiment?"

If the broken tube has in fact changed the treatment effects, and if the breakage occurred after the first day's runs, then it might be most useful to analyze the data for day one or each day separately. The design for each day is no longer a cyclic design or a group divisible design, but it is a connected incomplete block designs and can still be analyzed by computer (see Sects. 11.8 and 11.9).

If a test of the null hypothesis H_0^T of equal treatment effects is conducted separately for each day's data, it can be verified that H_0^T would not be rejected at the 5% significance level for the data collected on day two but would be for the data of day one.

The analysis of variance for day one is shown in Table 11.13. The test ratio is 853.40—which is larger than $F_{5,1,.05} = 230$. The mean square for blocks adjusted for treatments is 0.0000607, which

Fig. 11.4 Plasma data adjusted for block effects—day one only
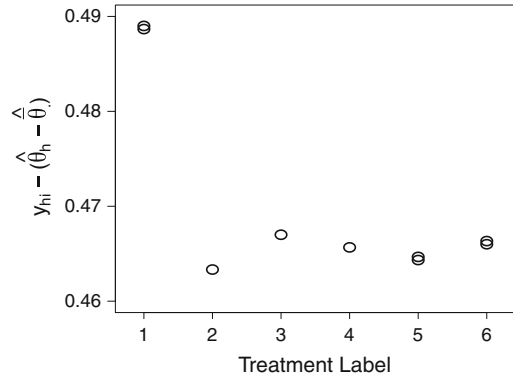
Table 11.14 Pairwise comparisons for the plasma experiment using the Scheffé method and confidence level 95%—day one only

i, p	$\hat{\tau}_i - \hat{\tau}_p$	$\sqrt{\widehat{\text{Var}}(\hat{\tau}_i - \hat{\tau}_p)}$	msd	Significant
1, 2	0.025500	0.000646	0.0219	yes
1, 3	0.021833	0.000553	0.0188	yes
1, 4	0.023167	0.000553	0.0188	yes
1, 5	0.024333	0.000471	0.0160	yes
1, 6	0.022667	0.000471	0.0160	yes
2, 3	−0.003667	0.000745	0.0253	
2, 4	−0.002333	0.000745	0.0253	
2, 5	−0.001167	0.000553	0.0188	
2, 6	−0.002833	0.000553	0.0188	
3, 4	0.001333	0.000745	0.0253	
3, 5	0.002500	0.000646	0.0219	
3, 6	0.000833	0.000553	0.0188	
4, 5	0.001167	0.000553	0.0188	
4, 6	−0.000500	0.000646	0.0219	
5, 6	−0.001667	0.000471	0.0160	

is 364 times larger than msE, so blocking was helpful for the observations collected on day one. With only one degree of freedom for error, use of residuals to check model assumptions is of little value. Figure 11.4 shows the day-one observations adjusted for block effects, $y_{hi} - (\hat{\theta}_h - \bar{\hat{\theta}}_\cdot)$ plotted against treatment. It appears that treatment 1 (Argon at 10 cc per second for 180 seconds) is very different from the other treatments.

Table 11.14 contains information for applying Scheffé's method of multiple comparisons to the day-one data, using a simultaneous 95% confidence level. The least squares estimates were obtained using SAS software (Sect. 11.8, p. 380, and Fig. 11.8). It can be seen that $\hat{\tau}_i - \hat{\tau}_p$ is larger than the minimum significant difference for all pairwise comparisons with $i = 1$, but for none of the others. Consequently, the only confidence intervals that do not contain zero are those involving treatment 1. We conclude that based on the first day's data, treatment 1 is significantly better than each of the other 5 treatments and should be investigated further.

11.6 Sample Sizes

Given the number of treatments v and the block size k, how many blocks b are required to achieve confidence intervals of a specified length or a hypothesis test of specified power? Since for most purposes the balanced incomplete block design is the best incomplete block design when it is available, we start by calculating b and the treatment replication $r = bk/v$ for this design. Then if a balanced incomplete block design cannot be found with b and r close to the calculated values, a group divisible, cyclic, or other incomplete block design can be considered. Since balanced incomplete block designs are the most efficient, other incomplete block designs would generally require b and r to be a little larger.

Example 11.6.1 `Sample size to achieve confidence interval length`

Suppose Tukey's method for all pairwise comparisons will be used to analyze an experiment with $v = 5$ treatments and block size $k = 3$. It is thought unlikely that msE will be larger than 2.0 units^2. Suppose that the experimenters want the length of simultaneous 95% confidence intervals for pairwise comparisons to be at most 3.0 units (that is, a minimum significant difference of at most 1.5). A balanced incomplete block design will ensure that the interval lengths will all be the same.

Using the fact that $bk = vr$ for a block design, the error degrees of freedom (11.4.8) can be written as

$$df = bk - b - v + 1 = vr - vr/k - v + 1 = (v(k-1)r/k) - (v-1), \qquad (11.6.18)$$

so, here, $df = (10r/3) - 4$. For a balanced incomplete block design, the minimum significant difference for a confidence interval for any pairwise treatment comparison, using Tukey's method with an overall 95% confidence level, is given in (11.4.13), p. 360 and, if we set $\lambda = r(k-1)/(v-1)$ from (11.3.1), p. 353, this becomes

$$msd = (q_{v,df,.05}/\sqrt{2})\sqrt{2\left[\frac{k(v-1)}{rv(k-1)}\right] msE} = q_{5,df,.05}\sqrt{\frac{12}{5r}},$$

with $df = (10r/3) - 4$. For the msd to be at most 1.5 units, it is necessary that

$$12q^2/5r \geq 1.5^2; \quad \text{that is,} \quad r \geq 1.0667 q_{5,df,.05}^2.$$

Trial and error shows that around 17–18 observations per treatment would be needed to satisfy the inequality; that is, 85–90 observations in total, which would require 28–30 blocks of size 3. A balanced incomplete block design exists with $v = 5, k = 3, b = 10, r = 6$ (all possible combinations of five treatments taken three at a time as blocks). Repeating this entire design three times would give a balanced incomplete block design with $r = 18$, which will give a minimum significant difference of about

$$q_{5,56,.05}\sqrt{12/(5 \times 18)} \approx 1.46 < 1.5.$$

$\square$

Example 11.6.2 `Sample size to achieve specified power`

Suppose a test of the null hypothesis $H_0 : \{\tau_i \text{ all equal}\}$ is required to detect a difference in the treatment effects of $\Delta = 2$ units with probability 0.95, using significance level $\alpha = 0.05$ for a balanced incomplete block design with $v = 5$ treatments and block size $k = 3$.

The least squares estimator $\hat{\tau}_i - \hat{\tau}_p$ of a pairwise comparison contrast $\tau_i - \tau_p$ for a balanced incomplete block design has variance given by (11.4.12), p. 360, with $\Sigma c_i^2 = 2$. Also, from (11.3.1), p. 353, $\lambda = r(k-1)/(v-1)$ so

$$\text{Var}(\Sigma c_i \hat{\tau}_i) = 2 \frac{k}{\lambda v} \sigma^2 = 2 \left[\frac{k(v-1)}{rv(k-1)} \right] \sigma^2 . \tag{11.6.19}$$

The number r of observations needed per treatment is calculated via a formula similar to (6.6.48), p. 171, with $a = v$ and with $2\sigma^2/b$ replaced by the variance (11.6.19); that is,

$$r = \frac{2v\sigma^2\phi^2}{\Delta^2} \left[\frac{k(v-1)}{v(k-1)} \right] = \frac{2 \times 5 \times \sigma^2\phi^2}{2^2} \left[\frac{3 \times 4}{5 \times 2} \right] .$$

Suppose that σ^2 is believed to be at most 1.0 unit^2; then $r = 3\phi^2$. The power tables in Appendix A.7 can be used to find ϕ^2. The numerator degrees of freedom are $\nu_1 = v-1$ and the denominator degrees of freedom ν_2 are the error degrees of freedom (11.6.18). So for our example, $\nu_1 = 4$ and $\nu_2 = (10r/3)-4$. Trial and error shows that about $r = 9$ observations per treatment are needed to satisfy the equality, requiring about $b = 15$ ($= vr/k$) blocks. A balanced incomplete block design exists with $v = 5$, $k = 3, b = 10, r = 6$ (all possible selections of three treatments taken as blocks). Repeating the entire design twice would give a balanced incomplete block design with $r = 12$, which would give more precision than required. Alternatively, one could use a computer program to seek a different type of incomplete block design with $r = 9$ and $b = 15$ (see Sects. 11.8.1 and 11.9.1, for example).   □

To meet the requirements of each of Examples 11.6.1 and 11.6.2, the resulting designs needed to be large. However, in cases when σ^2 is expected to be small and the block size can be large, the required number of blocks may be smaller than a balanced incomplete design or group divisible design can accommodate. Software such as that illustrated in Sects. 11.8.1 and 11.9.1 can be used to find other incomplete block designs that satisfy the requirements of the experiment.

11.7   Factorial Experiments

11.7.1 Factorial Structure

Any incomplete block design can be used for a factorial experiment by taking the treatment labels to represent treatment combinations. The incomplete block designs that are the most suitable for factorial experiments allow the adjusted treatment sum of squares ssT_{adj} to be written as a sum of the adjusted sums of squares for main effects and interactions. Thus, for an experiment with two factors C and D, for example, we would like to have

$$ssT_{\text{adj}} = ssC_{\text{adj}} + ssD_{\text{adj}} + ssCD_{\text{adj}} .$$

Such block designs are said to have *factorial structure*.

One benefit of this property is that the computations for, and interpretation of, the analysis of variance are simplified. A design with factorial structure requires that main-effect and interaction contrast estimates be adjusted only for block effects. In designs without factorial structure, the contrast estimates have to be adjusted not only for blocks but also for contrasts in all the other main effects

Table 11.15 Design and data for the step experiment

Block	Treatment combination					
	11	12	13	21	22	23
1		75	87	84	93	99
2	93	84	96	90	108	
3	99	93	96		123	129
4	99	108	99	99		120
5	99		111	90	129	141
6	129	135		120	147	153

and interactions. Although computer software can handle this adjustment, uncorrelated estimates are much easier to interpret and are, therefore, preferred.

All balanced incomplete block designs have factorial structure, and the features are illustrated in the following example.

Example 11.7.1 Step experiment

An experiment was run by S. Guerlain, B. Busam, D. Huland, P. Taige, and M. Pavol in 1993 to investigate the effects on heart rate due to the use of a step machine. The experimenters were interested in checking the theoretical model that says that heart rate should be a function of body mass, step height, and step frequency. The experiment involved the two treatment factors "step height" (factor C) and "step frequency" (factor D). Levels of "step height" were 5.75 and 11.5 inches, coded 1 and 2. "Step frequency" had three equally spaced levels, 14, 21, and 28 steps per minute, coded 1, 2, 3. The response variable was pulse rate in beats per minute.

The experiment used $b = 6$ subjects as blocks, and each subject was measured under $k = 5$ of the $v = 6$ combinations of step height and step frequency. The design was a balanced incomplete block design with blocks corresponding to different combinations of subject, run timer, and pulse measurer. All pairs of treatment combinations appeared together in $\lambda = 4$ blocks. The data are shown in Table 11.15.

Writing the treatment combinations as two-digit codes, the block–treatment model (11.4.2), p. 356, becomes

$$Y_{hij} = \mu + \theta_h + \tau_{ij} + \epsilon_{hij} \,,$$

and a set of least squares solutions for the treatment parameters adjusted for subject are given by (11.4.7) and (11.4.9), p. 358 and 359, with two-digit codes; that is,

$$\hat{\tau}_{ij} = \frac{k}{\lambda v} Q_{ij} = \frac{k}{\lambda v} \left[T_{ij} - \frac{1}{k} \sum_{h=1}^{b} n_{hij} B_h \right] \,,$$

where T_{ij} is the total of the $r = 5$ observations on step height i, step frequency j, B_h is the total of the $k = 5$ observations on the hth subject; and n_{hij} is 1 if treatment combination ij is observed for subject h and is zero otherwise. We obtain

$$\begin{array}{cccccc} \hat{\tau}_{11} & \hat{\tau}_{12} & \hat{\tau}_{13} & \hat{\tau}_{21} & \hat{\tau}_{22} & \hat{\tau}_{23} \\ -8.125 & -7.625 & -4.125 & -11.375 & 12.375 & 18.875 \end{array}$$

Table 11.16   Analysis of variance for the step experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Subject (Block) (adj)	5	6685.05	1337.01	–	–
Subject (Block) (unadj)	5	7400.40	–	–	–
Height (C) (adj)	1	1264.05	1264.05	28.63	0.0001
Frequency (D) (adj)	2	1488.90	744.45	16.86	0.0001
Ht×Freq (CD) (adj)	2	990.90	495.45	11.22	0.0006
Error	19	838.95	44.16		
Total	29	11983.20			

For a balanced incomplete block design, the adjusted sums of squares for the main effects of C (step height) and D (step frequency) and their interaction can be obtained by hand by using the values of $\hat{\tau}_{ij}$ in place of $y_{ij.}$, and $k/(\lambda v)$ in place of r in the formulae (6.4.20), (6.4.22), and (6.4.25), p. 157–159, or, equivalently, in Rule 4, p. 209, which leads to

$$ssC_{\text{adj}} = \frac{\lambda v}{k}\left[\frac{1}{d}\sum_{i=1}^{d}\hat{\tau}_{i.}^2 - \frac{1}{cd}\hat{\tau}_{..}^2\right] \; = \; \left(\frac{\lambda v}{k}\right)\frac{1}{d}\sum_{i=1}^{d}\hat{\tau}_{i.}^2$$

$$[3pt] = \left(\frac{24}{5}\right)\frac{1}{3}(-19.875^2 + 19.875^2) \; = \; 1264.05.$$

The adjusted treatment sums of squares, ssT_{adj}, is calculated from (11.4.10), p. 359, and using (11.4.9) becomes

$$ssT_{\text{adj}} = \sum_{i=1}^{v}\left(\frac{k}{\lambda v}\right)Q_i^2 \; = \; \sum_{i=1}^{v}\left(\frac{\lambda v}{k}\right)\hat{\tau}_i^2 \; = \; \left(\frac{4\times 6}{5}\right)\times 779.9688 \; = \; 3743.85\,.$$

The adjusted sums of squares for the main effects of C and D and their interaction are shown in analysis of variance Table 11.16. It can be verified that

$$ssC_{\text{adj}} + ssD_{\text{adj}} + ssCD_{\text{adj}} = ssT_{\text{adj}}\,.$$

Using the p-values in the analysis of variance Table 11.16, the experimenters rejected the hypothesis of negligible interaction. A plot of the data (not shown) suggests that heart rate increases linearly as the step frequency is increased, but that the linear trend is not the same for the two step heights. The experimenters wanted to examine the average behavior of the two factors, so despite this interaction, they decided to examine the main effects. In Exercise 10, the reader is asked to examine the linear trends at each step height separately.

For simplicity of notation, we now drop the subscript "adj." However, all estimates and sums of squares are adjusted for block effects. The experimenters were interested in examining the linear and quadratic trend contrasts for step frequency, that is,

$$D_{\text{L}} = -\overline{\tau}_{.1} + \overline{\tau}_{.3} \; = \; -\frac{1}{2}(\tau_{11} + \tau_{21}) + \frac{1}{2}(\tau_{13} + \tau_{23})\,,$$

$$D_{\text{Q}} = -\overline{\tau}_{.1} + 2\overline{\tau}_{.2} - \overline{\tau}_{.3} \; = \; -\frac{1}{2}(\tau_{11} + \tau_{21}) + \frac{2}{2}(\tau_{12} + \tau_{22}) - \frac{1}{2}(\tau_{13} + \tau_{23})\,.$$

For the balanced incomplete block design, the least squares estimate for a contrast $\sum c_{ij}\tau_{ij}$ and its associated variance are given by (11.4.11) and (11.4.12), p. 359–360; that is,

$$\frac{k}{\lambda v} \sum_i \sum_j c_{ij} Q_{ij} \quad \text{and} \quad \sum_i \sum_j c_{ij}^2 \left(\frac{k}{\lambda v}\right) \sigma^2,$$

respectively. Using these formulae, we find that the least squares estimates of the linear and quadratic trend contrasts for step frequency (adjusted for subjects) are

$$\hat{D}_L = 17.125 \quad \text{and} \quad \hat{D}_Q = -7.125.$$

The linear trend is positive, suggesting that the average pulse rate increases as the step frequency increases, and the quadratic trend is negative, suggesting that the increase in pulse rate is greater from 14 to 21 steps per minute than it is from 21 to 28 steps per minute. The null hypotheses $H_0^L : \{D_L = 0\}$ and $H_0^Q : \{D_Q = 0\}$ should be tested to check whether the perceived trends are significantly different from zero. The variances of the contrast estimators are

$$\left(\frac{4}{4}\right)\left(\frac{5}{24}\right)\sigma^2 = 0.2083\sigma^2 \quad \text{and} \quad \left(\frac{12}{4}\right)\left(\frac{5}{24}\right)\sigma^2 = 0.625\sigma^2,$$

respectively.

The contrast sum of squares for testing the null hypothesis $H_0^L : \{D_L = 0\}$ is obtained from (11.4.16), p. 360, as

$$ss(D_L) = \frac{(\hat{D}_L)^2}{\sum\sum c_{ij}^2 \left(\frac{k}{\lambda v}\right)} = \frac{17.125^2}{0.2083} = 1407.675,$$

and the contrast sum of squares for testing $H_0^Q : \{D_Q = 0\}$ is

$$ss(D_Q) = \frac{(\hat{D}_Q)^2}{\sum\sum c_{ij}^2 \left(\frac{k}{\lambda v}\right)} = \frac{(-7.125)^2}{0.625} = 81.225.$$

The linear and quadratic contrasts are orthogonal in a balanced incomplete block design even after adjusting for blocks, and we can now verify that indeed, $ssD = ss(D_L) + ss(D_Q)$,

To test the null hypotheses H_0^L and H_0^Q against their respective alternative hypotheses that the null hypothesis is false, we compare each of $ss(D_L)/msE = 31.88$ and $ss(D_Q)/msE = 1.84$ with $2F_{2,19,.01} = 7.04$ for Scheffé's method and an overall level of $\alpha = 0.01$. We conclude that the quadratic trend is negligible, but there is a nonnegligible linear trend in the heart rate as the stepping frequency increases (averaged over step height).   □

Table 11.17 SAS program for generation of an efficient incomplete block design

```
* Using proc optex to search for an efficient block design with v = 7,
b = 7, k = 3;
DATA CANDIDATE;
  DO TREATMNT = 1 to 7;
    OUTPUT;
  END;

PROC OPTEX DATA = CANDIDATE SEED = 72145;
  CLASS TREATMNT;
  MODEL TREATMNT;
  * For 7 blocks of size 3;
  BLOCKS STRUCTURE = (7)3;
  EXAMINE DESIGN;
```

11.8   Using SAS Software

11.8.1 Generation of Efficient Block Designs

PROC OPTEX within the SAS software allows one to search for efficient incomplete block designs. Although one cannot specify the type of design to be generated, the software will search for the design that gives the smallest confidence region for all contrasts using the Scheffé method of multiple comparisons. If a balanced incomplete block design exists, it will usually be found by PROC OPTEX.

The first set of lines in the SAS program in Table 11.17 specify that there are 7 treatments and these are stored in a dataset called CANDIDATE. Then, in the second set of lines, PROC OPTEX is run for a block structure "(b)k". In Table 11.17, there are $b = 7$ blocks of size $k = 3$. Since the program is not guaranteed to find the optimal design, it makes 10 independent searches (–this number can be changed by the user). The use of a particular "SEED =" always starts the search at the same point, so that the same set of designs is obtained. This should be removed when starting a new experiment so that random starts of the search are made. The designs found using the seed given in Table 11.17 are listed in the first part of the output in Fig. 11.5. Among the designs found, the one which gives the smallest confidence region for all contrasts is the one with the largest value under the heading "Treatment D-efficiency". This will often coincide with the design with the largest "Treatment A-efficiency" which has the shortest average length of confidence intervals for pairwise comparisons. To search specifically for the best design under A-efficiency, we insert the statement GENERATE CRITERION = A; after the BLOCKS STRUCTURE statement, and the designs will then be rank ordered by the A-efficiency.

The designs found in the 10 searches may or may not be exactly the same, but those listed in Fig. 11.5 are equally good as measured by their efficiencies. The command EXAMINE DESIGN prints out the best design (the one at the top of the list). It can be seen in Fig. 11.5 that, for this design, Block I (listed as the first three "points") consists of treatments 1, 2, 4, while Block II consists of 2, 3, 7, and so on. It can be checked that this is a balanced incomplete block design, although not the same one as that in Table 11.2, p. 351. To examine a design which is not listed as the best, say the third design in the list, replace EXAMINE DESIGN; by EXAMINE NUMBER = 3 DESIGN;.

The quoted value of "A-efficiency" is the ratio of the average variance of the pairwise comparisons in the design being examined relative to the average variance in a randomized block design with the same

Fig. 11.5 SAS output
from PROC OPTEX
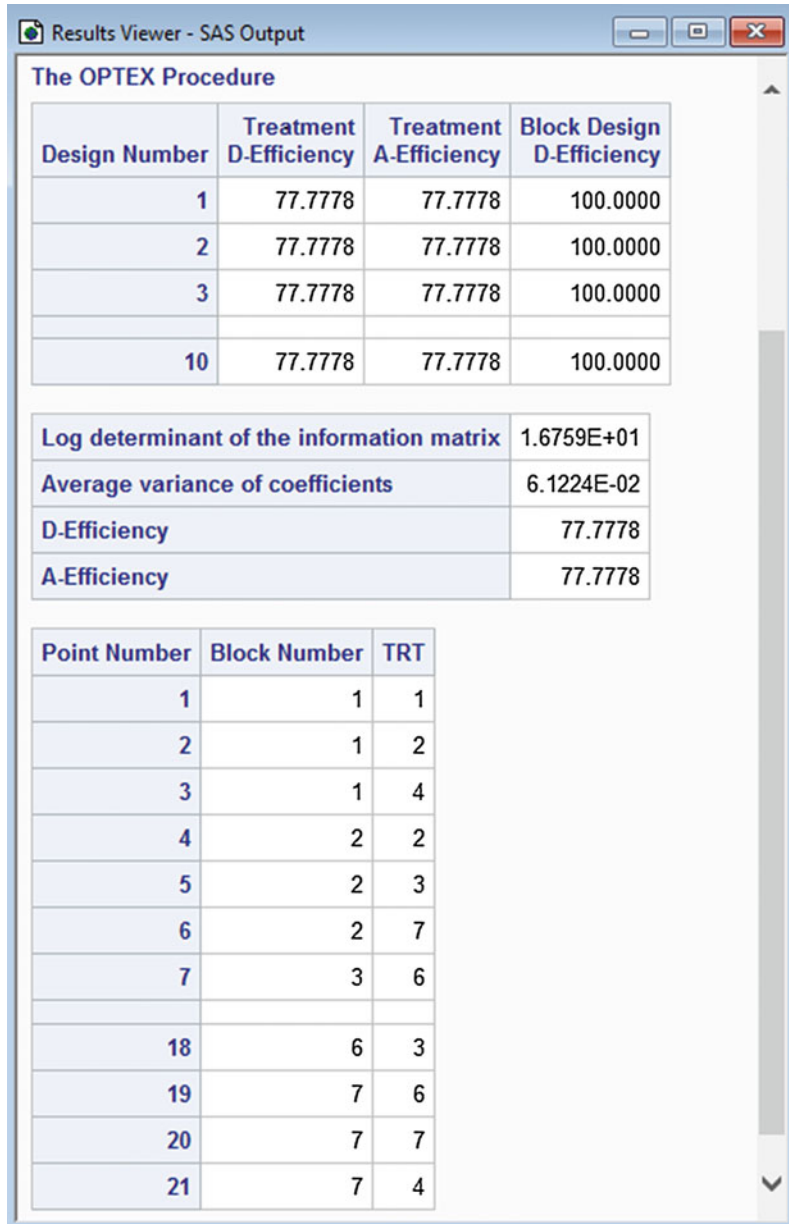
value of r and multiplied by 100%. Since, in this example, the design found is a balanced incomplete block design, the average variance is given by (11.4.12), p. 360, and the ratio is

$$\frac{2\sigma^2/r}{2k\sigma^2/\lambda v}100\% = \frac{\lambda v}{rk}100\% = \frac{v(k-1)}{k(v-1)}100\% = 77.777\% \,.$$

The quoted "average D-efficiency" is related to the volume of the confidence region for all contrasts in the design being examined as compared with that for a randomized block design with the same value of r.

Table 11.18 SAS program for analysis of a balanced incomplete block design—detergent experiment

```
DATA ONE;
  INPUT BLOCK TRTMT Y;
  LINES;
   1 3 13
   : :  :
  12 1 19
;
PROC SGPLOT;
  SCATTER X = TRTMT  Y = Y / GROUP = BLOCK MARKERCHAR = BLOCK;

PROC GLM;
  CLASS BLOCK TRTMT;
  MODEL Y = BLOCK TRTMT;
  OUTPUT OUT = RESIDS PREDICTED = PREDY RESIDUALS = E;
  * contrast sums of squares for 8 orthogonal contrasts;
  CONTRAST 'I linear'         TRTMT -3 -1  1  3  0  0  0  0  0;
  CONTRAST 'I quadratic'      TRTMT  1 -1 -1  1  0  0  0  0  0;
  CONTRAST 'I cubic'          TRTMT -1  3 -3  1  0  0  0  0  0;
  CONTRAST 'II linear'        TRTMT  0  0  0  0 -3 -1  1  3  0;
  CONTRAST 'II quadratic'     TRTMT  0  0  0  0  1 -1 -1  1  0;
  CONTRAST 'II cubic'         TRTMT  0  0  0  0 -1  3 -3  1  0;
  CONTRAST 'I vs II'          TRTMT  1  1  1  1 -1 -1 -1 -1  0;
  CONTRAST 'others vs control' TRTMT 1  1  1  1  1  1  1  1 -8;
  * estimation of  treatment versus control contrasts via LSMEANS;
  LSMEANS TRTMT / PDIFF = CONTROL('9') CL ADJUST = DUNNETT;
  * estimation of  treatment versus control contrast via ESTIMATE;
  ESTIMATE 'Det 9-1' TRTMT -1  0  0  0  0  0  0  0  1;
```

For the requirements of Example 11.6.2, if PROC OPTEX is run with this same seed for $v = 5$ treatments and $b = 15$ blocks of size $k = 3$, the best design found has pairs of treatments appearing together in either $\lambda_1 = 4$ or $\lambda_2 = 5$ blocks. It consists of the 10-block balanced incomplete block design together with an additional 5 blocks comprising a cyclic incomplete block design. (A different seed, or no specified seed, may result in the additional 5 blocks being a non-cyclic incomplete block design but the best design listed will most likely still have $\lambda_2 = \lambda_1 + 1$).

11.8.2 Analysis of Variance, Contrasts, and Multiple Comparisons

In this section, sample programs are given to illustrate the analysis of incomplete block designs using the SAS software. The programs shown are for the detergent experiment of Sect. 11.4.4 and the plasma experiment of Sect. 11.5, but similar programs can be used to analyze the data collected in any incomplete block design.

Table 11.18 contains the first sample program. The data are entered into a data set called ONE, using the variables BLOCK, TRTMT, and Y for the block, treatment, and response value, respectively. PROC SGPLOT is used to plot the observations against treatments, analogous to Fig. 11.2, p. 361, and the legend identifies the block labels by color (the plot is not shown here). The block labels are printed on the plot by inclusion of the command MARKERCHAR = BLOCK. In the next section of Table 11.18, PROC GLM is used to fit the block–treatment model (11.4.2), generate the analysis of

Fig. 11.6 Partial output
from `PROC GLM` for
analysis of an incomplete
block design—detergent
experiment

variance table, and save the predicted values and residuals in the output data set `RESIDS`. Residuals
can be standardized and plotted as in Chap. 6.

Output from `PROC GLM` is reproduced in Fig. 11.6. Since `BLOCK` has been entered before `TRTMT`
in the model statement, the sum of squares for treatments adjusted for blocks is listed under
`Type I` (or sequential) sums of squares as well as under the `Type III` sums of squares. The
adjusted block sum of squares is listed under the `Type III` sums of squares. In order to use the
sequential or `Type I` sums of squares to obtain the adjusted block sum of squares, one would need to
rerun the program with `TRTMT` entered before `BLOCK` in the model statement. In Table 11.18, the sums
of squares corresponding to $v-1 = 8$ orthogonal treatment contrasts are requested via the `CONTRAST`
statements, and it can be verified from Fig. 11.6 that the contrast sums of squares add to the treatment
sum of squares.

Simultaneous confidence intervals for pairwise comparisons can be obtained via the `ESTIMATE`
statements or via `LSMEANS` with options as discussed in Sect. 6.8.2, p. 180. Tukey, Scheffé and Dunnett
methods can all be used for a balanced incomplete block design with code such as:

```
LSMEANS TRTMT / PDIFF=CONTROL('9') CL ADJUST=DUNNETT;
```

Here, the `PDIFF=CONTROL('9')` option for Dunnett's method specifies that level 9 is the control,
as was the case in the detergent experiment. If the designation "`('9')`" had been omitted, then the

Fig. 11.7 Partial output from ESTIMATE and LSMEANS for an incomplete block design—detergent experiment with detergent 9 as the control treatment
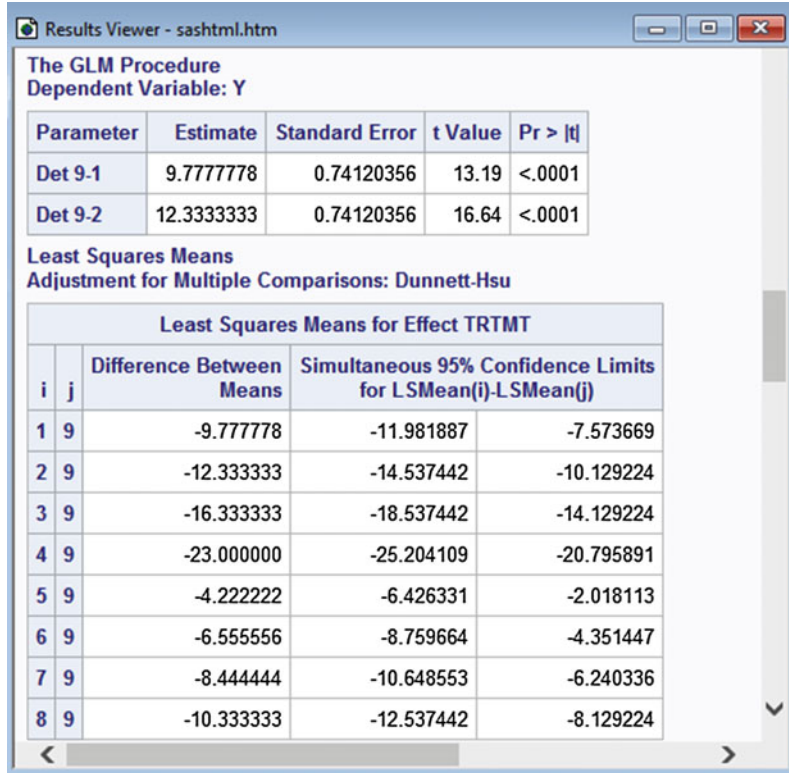
lowest level would have been taken to be the control treatment by default. Partial output is shown in Fig. 11.7. The first section of the table shows the output from the ESTIMATE statements. The second section of the output gives simultaneous 95% confidence intervals for the treatment-versus-control comparisons using Dunnett's method. A word of warning is in order here. If the treatments had been labeled anything other than $1, 2, \ldots, 9$, at this point SAS software would have relabeled them. For example, if the control treatment had been labeled as 0 and the test treatments as $1, \ldots, 8$, SAS software would have relabeled the control as treatment 1 and the test treatments as $2, \ldots, 9$.

Figure 11.8 shows partial output from PROC GLM in the first part of Table 11.19 for day one data from the plasma experiment (Sect. 11.5), which was a nonstandard incomplete block design. The Type I and Type III sums of squares are shown, together with partial output from the LSMEANS statement

```
LSMEANS TRTMT / PDIFF = ALL CL ADJUST=SCHEFFE;
```

Output from the above LSMEANS statement, combined with standard error estimates generated by an ESTIMATE statement for each pairwise treatment contrast like the following one for $\tau_1 - \tau_2$, were used to compile Table 11.14 (p. 370):

```
ESTIMATE 'T1-T2' TRTMT 1 -1 0 0 0 0;
```

Fig. 11.8 Partial output from `PROC GLM` and `LSMEANS` in a SAS program for analyzing an incomplete block design—plasma experiment, day one

Results Viewer - SAS Output

The GLM Procedure
Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	0.00111406	0.00015915	954.90	0.0249
Error	1	0.00000017	0.00000017		
Corrected Total	8	0.00111422			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BLOCK	2	0.00040289	0.00020144	1208.67	0.0203
TRTMT	5	0.00071117	0.00014223	853.40	0.0260

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BLOCK	2	0.00012133	0.00006067	364.00	0.0370
TRTMT	5	0.00071117	0.00014223	853.40	0.0260

Least Squares Means
Adjustment for Multiple Comparisons: Scheffe

		Least Squares Means for Effect TRTMT		
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	0.025500	0.003602	0.047398
1	3	0.021833	0.003081	0.040585
1	4	0.023167	0.004415	0.041919
1	5	0.024333	0.008342	0.040325
1	6	0.022667	0.006675	0.038658
2	3	-0.003667	-0.028952	0.021618
2	4	-0.002333	-0.027618	0.022952

11.8.3 Plots

Table 11.19 contains a sample SAS program illustrating how to plot the data adjusted for blocks against the treatment labels, using the day one data of the plasma experiment, (first three blocks of Table 11.11, p. 368). The program, as written, must be run in three passes. In successive passes, information generated by earlier passes must be added as input in later parts of the program. First, the data are entered into a data set called `ONE`. Since the block effect estimates are needed to adjust the observations, the option `SOLUTION` is included in the `MODEL` statement of `PROC GLM`. This causes a (nonunique) solution to the normal equations for $\hat{\mu}$, $\hat{\tau}_i$, and $\hat{\theta}_h$ to be printed. The solutions will all be labeled "B" for "biased," meaning that the corresponding parameters are not individually estimable (see Fig. 10.8, p. 330, for example).

The least squares solutions $\hat{\theta}_h$ are then entered (as "BHAT") into the data set `TWO` by the user in the second run of the program. `PROC MEANS` is used to compute and print the average value $\bar{\hat{\theta}}_\cdot$. Finally,

Table 11.19 SAS program to plot data adjusted for block effects—plasma experiment, day one only

```
* This program requires 3 runs, adding more information in each run;
DATA ONE;
  INPUT BLOCK TRTMT Y;
  LINES;
   1  4  0.459
   1  5  0.467
   :  :  :
   3  3  0.473
;
* Get block effect estimates;
PROC GLM;
  CLASS BLOCK TRTMT;
  MODEL Y = BLOCK TRTMT / SOLUTION;
  LSMEANS TRTMT / PDIFF = ALL CL ADJUST=SCHEFFE;
PROC SORT;  BY BLOCK;
* Add the following code for the second run;
* values BHAT are solutions for block parameters from first run;
DATA TWO;
  INPUT BLOCK BHAT;
  LINES;
      1    -.0126666667
      2    -.0053333333
      3     0.0000000000
;
PROC MEANS MEAN; * print average of BHAT values;
  VAR BHAT;
* Add the following code for the third run;
* The number -0.006 below is average BHAT calculated in second run;
DATA THREE;
  MERGE ONE TWO;
  BY BLOCK;
  Y_ADJ = Y - (BHAT - (-0.006));
PROC SGPLOT data = THREE;
  SCATTER X = TRTMT Y = Y_ADJ / GROUP = BLOCK;
```

in the third run of the program, the block-effect estimates and their average value are used to adjust the data values. The adjusted values are then plotted against treatment. The SAS plot is not shown here, but it is similar to the plot in Fig. 11.4 (p. 370).

11.9   Using R Software

11.9.1 Generating Efficient Incomplete Block Designs

The R function `bibd` from the package `ibd` can be used to construct balanced incomplete block designs when they exist. The first few lines of the program in Table 11.20 install and load the `ibd` package, then ask the program to search for a balanced incomplete block design with $v = 7$ treatments each appearing $r = 3$ times, $b = 7$ blocks of size $k = 3$, and pairs of treatments appearing together in $\lambda = 1$ block. In general, the `bibd` function checks that the necessary conditions for an existence of a

Table 11.20 R code and output for bibd with $v = 7$ treatments, $b = 7$ blocks of size $k = 3$

```
> # install.packages("ibd")
> library(ibd)
> bibd(v = 7, r=3, b = 7, k = 3, lambda = 1)

$design
     [,1] [,2] [,3]
[1,]   1    3    6
[2,]   1    4    5
[3,]   5    6    7
[4,]   1    2    7
[5,]   3    4    7
[6,]   2    4    6
[7,]   2    3    5

$Aeff
[1] 0.9999999

$Deff
[1] 0.9999999
```

bibd are satisfied. If so, it will either provide the design or respond with "design not found". For the design size in Table 11.20, the design does exist and is shown under the heading $design. The seven blocks of the design are given by the seven rows of three treatment labels. The quoted $Aeff is the ratio of the average variance of the pairwise comparisons in this incomplete block design relative to the average variance (11.4.12) in a balanced incomplete block design with the same values of v and k. The quoted $Deff is related to the volume of the confidence region for all contrasts in the design being examined relative to a balanced incomplete block design with the same values of v and k (which may not actually exist in practice). Since the design found here is, itself, a balanced incomplete block design, A-eff and D-eff are given as 1 (approximately).

If no balanced incomplete block design exists, as for the requirements of Example 11.6.2, with $v = 5$ treatments and $b = 15$ blocks of size $k = 3$, then function ibd, whose inputs are v, b, and k, can be used to construct an incomplete block design. For this example, the best design found is shown in Table 11.21 and has "A.Efficiency = 0.9975" and "D.Efficiency= 0.9987", where these efficiencies are calculated the same way as $Aeff and $Deff in the bibd function. It consists of a balanced incomplete block design with 10 blocks, together with an additional 5 blocks comprising an incomplete block design with $\lambda_2 = \lambda_1 + 1$. The entire 15-block design has pairs of treatments appearing together in either $\lambda_1 = 4$ or $\lambda_2 = 5$ blocks. For a binary design, the values of λ_i are listed as the off-diagonal elements in the array called $conc.mat so, for example, the first row of this array shows that treatment 1 appears in 4 blocks with treatments 2 and 3, and in 5 blocks with treatments 4 and 5. The diagonal elements of the array are the values of r.

The ibd function prints out the design with the highest D.Efficiency that it finds in 5 independent searches. If the design found has low efficiency, ibd may be able to obtain an improved design if the option ntrial=n is added into the ibd statement, where n is some integer larger than the default value of 5; for example,

$$ibd(v = 5, b = 15, k = 3, ntrial = 20).$$

Table 11.21 R code and output for an incomplete block design with $v = 5$ treatments, $b = 15$ blocks of size $k = 3$

```
> library(ibd)
> ibd(v=5, b=15, k=3)

$design
       [,1] [,2] [,3]
 [1,]    1    2    3
 [2,]    1    4    5
 [3,]    1    4    5
 [4,]    3    4    5
 [5,]    1    3    5
 [6,]    1    3    4
 [7,]    2    3    4
 [8,]    1    3    4
 [9,]    2    4    5
[10,]    2    3    4
[11,]    1    2    5
[12,]    1    2    5
[13,]    1    2    4
[14,]    2    3    5
[15,]    2    3    5

$conc.mat
      [,1] [,2] [,3] [,4] [,5]
[1,]     9    4    4    5    5
[2,]     4    9    5    4    5
[3,]     4    5    9    5    4
[4,]     5    4    5    9    4
[5,]     5    5    4    4    9

$A.Efficiency
[1] 0.9975309

$D.Efficiency
[1] 0.9987647
```

11.9.2 Analysis of Variance, Contrasts, and Multiple Comparisons

In this section, sample programs are given to illustrate the analysis of incomplete block designs using the R software. The programs shown are for the detergent experiment of Sect. 11.4.4 and the plasma experiment of Sect. 11.5, but similar programs can be used to analyze the data collected in any incomplete block design.

Table 11.22, p. 385, contains the first sample program. The data are read into the data set detrgnt.data, using the variables Block, Trtmt, and y for the block, treatment, and response value, respectively. Corresponding factor variables fBlock ad fTrtmt are added to the data set. In the second block of code, the plot function is used to plot the observations against treatments, analogous to Fig. 11.2, p. 361, but using block labels as the plotting legend (the plot is not shown here).

In the third block of code, the linear models function lm is used to fit the block–treatment model (11.4.2), then corresponding anova and drop1 functions generate the Type I (or sequential)

Table 11.22 R program for analysis of a balanced incomplete block design—detergent experiment

```
detrgnt.data = read.table("data/detergent.txt", header=T)
head(detrgnt.data)
detrgnt.data = within(detrgnt.data,
  {fBlock = factor(Block); fTrtmt = factor(Trtmt) })

# Plot y vs Trtmt using block level as plotting symbol.
plot(y ~ Trtmt, data=detrgnt.data, xaxt="n", type="n") # Suppress x-axis, pts
  axis(1, at=seq(1,9,1)) # x-axis labels 1:9
  text(y ~ Trtmt, Block, cex=0.75, data=detrgnt.data) # Plot y*Trtmt=Block
  mtext("Block=1,...,12", side=3, adj=1, line=1) # Margin text, TopRt, line 1

# Analysis of variance
model1 = lm(y ~ fBlock + fTrtmt, data=detrgnt.data)
anova(model1)
drop1(model1, ~., test="F")

# Contrast estimates
library(lsmeans)
lsmTrtmt = lsmeans(model1, ~ fTrtmt)
cntrsts = summary(contrast(lsmTrtmt,
                          list(I.linear=c(-3,-1, 1, 3, 0, 0, 0, 0, 0),
                                 I.quad=c( 1,-1,-1, 1, 0, 0, 0, 0, 0),
                                I.cubic=c(-1, 3,-3, 1, 0, 0, 0, 0, 0),
                              II.linear=c( 0, 0, 0, 0,-3,-1, 1, 3, 0),
                                II.quad=c( 0, 0, 0, 0, 1,-1,-1, 1, 0),
                               II.cubic=c( 0, 0, 0, 0,-1, 3,-3, 1, 0),
                                I.vs.II=c( 1, 1, 1, 1,-1,-1,-1,-1, 0),
                           Trt.vs.Ctrl=c( 1, 1, 1, 1, 1, 1, 1, 1,-8))),
                      infer=c(T,T))
# Compute and include contrast sums of squares: ss=t^2*mse
mse = anova(model1)[3,3]; mse
cntrsts = cbind(cntrsts, ss=cntrsts[,"t.ratio"]^2*mse)
options(width=75, digits=3, scipen=2)
cntrsts
options(ooptions)

# Dunnett's method
summary(contrast(lsmTrtmt, method="trt.vs.ctrl", adjust="mvt", ref=9),
        infer=c(T,T))
```

and Type III analysis of variance tables, respectively. Predicted values and residuals are available, and the residuals can be standardized and plotted as in Sect. 6.2.3.

The analysis of variance output from `anova` and `drop1` is reproduced in Table 11.23, p. 386. Since `fBlock` has been entered before `fTrtmt` in the model statement, the sum of squares for treatments adjusted for blocks are obtained as both Type I and Type III sums of squares. The adjusted block sum of squares is listed under the Type III sums of squares whereas, to use the sequential or Type I sums of squares, one would need to rerun the program with `fTrtmt` entered before `fBlock` in the model statement. In the center of Table 11.22, the `summary(contrast(lsmTrtmt...` statement is used to list eight orthogonal treatment contrasts, producing corresponding least squares estimates, standard errors, two-sided *t*-tests, and individual 95% confidence intervals, as shown in the bottom part

Table 11.23 Selected ANOVA and contrast output for analysis of an incomplete block design—detergent experiment

```
> anova(model1)

Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
fBlock    11    413    37.5    45.5 6.0e-10
fTrtmt     8   1087   135.9   164.8 6.8e-14
Residuals 16    13     0.8

> drop1(model1, ~., test="F")

Single term deletions
Model:
y ~ fBlock + fTrtmt
       Df Sum of Sq  RSS   AIC F value  Pr(>F)
<none>                13   3.8
fBlock 11         10  23   2.3    1.11    0.41
fTrtmt  8       1087 1100 147.1  164.85 6.8e-14

> cntrsts

     contrast estimate   SE df lower.CL upper.CL t.ratio  p.value        ss
1    I.linear  -43.667 2.34 16   -48.64   -38.70 -18.630 2.85e-12 286.0167
2      I.quad   -4.111 1.05 16    -6.33    -1.89  -3.922 1.22e-03  12.6759
3     I.cubic   -1.222 2.34 16    -6.19     3.75  -0.521 6.09e-01   0.2241
4   II.linear  -20.222 2.34 16   -25.19   -15.25  -8.628 2.05e-07  61.3407
5     II.quad    0.444 1.05 16    -1.78     2.67   0.424 6.77e-01   0.1481
6    II.cubic   -0.444 2.34 16    -5.41     4.52  -0.190 8.52e-01   0.0296
7     I.vs.II  -31.889 1.48 16   -35.03   -28.75 -21.512 3.10e-13 381.3380
8  Trt.vs.Ctrl -91.000 4.45 16  -100.43   -81.57 -20.462 6.73e-13 345.0417
```

of Table 11.23. The sum of squares for each contrast is subsequently computed from its corresponding t-ratio as $\texttt{ss} = (\texttt{t-ratio})^2 \times mse$, which follows since $(\texttt{t-ratio})^2 = \texttt{ss}/mse$ from (4.3.16) p. 78, where mse is the mean squared error from the `Residuals` line in the analysis of variance table.

Simultaneous confidence intervals for pairwise comparisons can be obtained via the `lsmeans` function as discussed in Sect. 6.9.2, p. 187. For example, Dunnett's method is invoked by the following statements in the bottom of Table 11.22, p. 385.

```
lsmTrtmt = lsmeans(model1, ~ fTrtmt)
summary(contrast(lsmTrtmt, method="trt.vs.ctrl", adjust="mvt", ref=9),
    infer=c(T,T))
```

The `ref=9` option specifies that the 9th level (coincidentally also labeled `"9"`) is the control, as was the case in the detergent experiment. If this designation had been omitted, then the lowest level would have been taken to be the control treatment by default. Table 11.24, p. 387, contains the resulting output, including the estimate of each treatment-versus-control comparison $\tau_i - \tau_9$, as well as the corresponding standard error, simultaneous 95% confidence limits by Dunnett's method, test statistic, and p value for simultaneous tests for whether or not each of the treatment-versus-control comparisons $\tau_i - \tau_9$ is zero, or equivalently whether $\tau_i = \tau_9$, using Dunnett's method.

Table 11.24 Dunnett's method output for an incomplete block design—detergent experiment with detergent 9 as the control treatment

```
> # Dunnett's method
> summary(contrast(lsmT, method="trt.vs.ctrl", adjust="mvt", ref=9),
+          infer=c(T,T))

 contrast estimate      SE df lower.CL upper.CL t.ratio p.value
  1 - 9     -9.7778 0.7412 16 -11.9824  -7.5732 -13.192  <.0001
  2 - 9    -12.3333 0.7412 16 -14.5379 -10.1287 -16.640  <.0001
  3 - 9    -16.3333 0.7412 16 -18.5379 -14.1287 -22.036  <.0001
  4 - 9    -23.0000 0.7412 16 -25.2046 -20.7954 -31.031  <.0001
  5 - 9     -4.2222 0.7412 16  -6.4268  -2.0176  -5.696  0.0002
  6 - 9     -6.5556 0.7412 16  -8.7601  -4.3510  -8.844  <.0001
  7 - 9     -8.4444 0.7412 16 -10.6490  -6.2399 -11.393  <.0001
  8 - 9    -10.3333 0.7412 16 -12.5379  -8.1287 -13.941  <.0001

Results are averaged over the levels of: fB
Confidence level used: 0.95
Confidence-level adjustment: mvt method for 8 estimates
P value adjustment: mvt method for 8 tests
```

Table 11.25, p. 388, shows an R program and partial output for analysis of the day one data from the plasma experiment (Table 11.11, p. 368), which used a nonstandard incomplete block design. After reading the entire set of data, `data/plasma.txt`, and defining the block and treatment factors, the `subset` function is used to create a data set `day1.data` containing only day one data. The Type I and type 3 sums of squares generated by `anova` and `drop1`, respectively, are shown at the top of the table. The bottom of the table contains partial output from the `lsmeans` and `summary(contrast` statements, comparing treatment effects pairwise via Scheffé's method. This is one way to obtain information analogous to that in Table 11.14 (p. 370).

11.9.3 Plots

Table 11.26, p. 389, contains a sample R program illustrating how to plot the data adjusted for blocks against the treatment labels, using the day one data of the plasma experiment, (first three rows of Table 11.11, p. 368). The `subset` function is used to create a data set `day1.data` containing only day one data. Having fit the linear model by the `lm` function and saved the results as `model2`, the least squares estimates of the model parameters are available as `model2$coefficients`. If one were to display these, one would see that the second and third coefficient estimates are the block effect estimates $\hat{\theta}_2 \approx 0.0073333$ and $\hat{\theta}_3 \approx 0.0126667$, but there is no estimate listed for the first block effect, so $\hat{\theta}_1 = 0$. These three block effect estimates are saved in the column `thetahat`, and the mean estimate $\bar{\hat{\theta}}_.$ is saved as `avgest`. These values are used to compute adjusted observations, $\mathtt{yadj} = y - (\hat{\theta}_h - \bar{\hat{\theta}}_.)$, which are then plotted against treatment. The R plot is not shown here, but it is similar to the plot in Fig. 11.4 (p. 370).

Table 11.25 R program and partial output from `anova`, `drop1`, and `lsmeans`, analyzing an incomplete block design—plasma experiment

```
> plasma.data = read.table("data/plasma.txt", header=T)
> plasma.data = within(plasma.data,
+   {fBlock = factor(Block); fTrtmt = factor(Trtmt) })
> head(plasma.data, 3)
> # Analysis of day 1 data
> day1.data = subset(plasma.data, Day==1) # Keep data with "Day==1"
> model2 = lm(y ~ fBlock + fTrtmt, data=day1.data)
> anova(model2)

Analysis of Variance Table
Response: y
          Df    Sum Sq   Mean Sq F value Pr(>F)
fBlock     2 0.000403 0.0002014    1209   0.020
fTrtmt     5 0.000711 0.0001422     853   0.026
Residuals  1 0.000000 0.0000002

> drop1(model2, ~., test="F")

Single term deletions
Model:
y ~ fBlock + fTrtmt
       Df Sum of Sq      RSS     AIC F value Pr(>F)
<none>             0.000000 -144.2
fBlock  2  0.000121 0.000121  -88.9     364  0.037
fTrtmt  5  0.000711 0.000711  -79.0     853  0.026

> # Scheffe's method
> library(lsmeans)
> lsmTrtmt = lsmeans(model2, ~ fTrtmt)
> summary(contrast(lsmTrtmt, method="pairwise", adjust="Scheffe"),
+         infer=c(T,T))

 contrast     estimate          SE df   lower.CL upper.CL t.ratio p.value
 1 - 2      0.02550000 0.00064550  1  0.0036024 0.047398  39.504  0.0429
 1 - 3      0.02183333 0.00055277  1  0.0030814 0.040585  39.498  0.0430
 1 - 4      0.02316667 0.00055277  1  0.0044147 0.041919  41.910  0.0405
   :            :           :     :    :           :        :       :
 5 - 6     -0.00166667 0.00047140  1 -0.0176584 0.014325  -3.536  0.4451

Results are averaged over the levels of: fBlock
Confidence level used: 0.95
Confidence-level adjustment: scheffe method for a family of 6 estimates
P value adjustment: scheffe method for a family of 6 tests
```

Table 11.26 R program to plot data adjusted for block effects—plasma experiment, day one only

```
# Create data set of day 1 data
plasma.data = read.table("data/plasma.txt", header=T)
plasma.data = within(plasma.data,
  {fBlock = factor(Block); fTrtmt = factor(Trtmt) })
day1.data = subset(plasma.data, Day==1) # Keep data with "Day==1"

# Fit model to day 1 data
model2 = lm(y ~ fBlock + fTrtmt, data=day1.data)

# Create columns of block effect estimates, and compute the avg estimate
model2$coefficients # Display all model coefficient estimates
thetahat = c(0, model2$coefficients[2:3]) # Create col of block effect ests
avgest = mean(thetahat) # Compute average block effect estimate

# Compute adjusted y-values: yadj = y - (thetahat.h - thetahat.bar)
day1.data$yadj = day1.data$y - (thetahat[day1.data$Block] - avgest)

# Plotting day 1 data adjusted for block effects
plot(yadj ~ Trtmt, data=day1.data, xlab="Treatment", ylab="y Adjusted")
```

Table 11.27 Three incomplete block designs

Design I		Design II		Design III	
Block	Treatments	Block	Treatments	Block	Treatments
1	1 2	1	1 2 3	1	1 2 6
2	1 3	2	4 5 6	2	3 4 5
3	1 4	3	7 8 9	3	2 6 8
4	2 3	4	1 4 7	4	4 5 7
5	2 4	5	2 5 8	5	1 6 8
6	3 4	6	3 6 9	6	3 5 7
		7	1 5 9	7	1 2 8
		8	2 6 7	8	3 4 7
		9	3 4 8		

Exercises

1. **Connectedness and estimability**

   (a) For each of the three block designs in Table 11.27, draw the connectivity graph for the design, and determine whether the design is connected.
   (b) If the design is connected, determine whether or not it is a balanced incomplete block design.
   (c) For designs II and III, determine graphically whether or not $\tau_1 - \tau_5$ and $\tau_1 - \tau_6$ are estimable.
   (d) For design III, use expected values to show that $\tau_1 - \tau_8$ is estimable.

2. **Connectedness**

   (a) Determine whether or not the cyclic design with initial block $(1, 3, 5)$ is a connected design if $v = 8$ or $v = 9$.
   (b) Determine whether or not the cyclic design with initial block $(1, 4, 7)$ is a connected design if $v = 8$ or $v = 9$.

3. **Randomization**

   Conduct a block design randomization for design II in Table 11.27.

4. **Cyclic designs**

   Determine whether or not the cyclic design obtained from each initial block below is a balanced incomplete block design or a group divisible design or neither.

   (a) Initial block: 1, 3, 4; $v = 7$.
   (b) Initial block: 1, 2, 4, 8; $v = 8$.
   (c) Initial block: 1, 2, 4; $v = 5$.

5. **Balanced incomplete block design**

   In the following questions, consider an experiment to compare $v = 7$ treatments in blocks of size $k = 5$.

   (a) Show that, for this experiment, a necessary condition for a balanced incomplete block design to exist is that r is a multiple of 5 and b is a multiple of 7.
   (b) Show that r must be at least 15.
   (c) Taking all possible combinations of five treatments from seven gives a balanced incomplete block design with $r = 15$. Calculate the number of blocks that must be in this design.

6. **Sample sizes**

   Consider an experiment to compare 7 treatments in blocks of size 5, with an anticipated error variance of at most 30 squared units.

   (a) Assuming that a balanced incomplete block design will be used, how many observations would be needed for the minimum significant difference to be about 50 units for a pairwise comparison using Tukey's method and a 95% simultaneous confidence level?
   (b) Repeat part (a) for a minimum significant difference of 25 units.
   (c) Repeat part (a) using Dunnett's method for treatment versus control comparisons.

7. **Least squares estimator, detergent experiment, continued**

   Consider the balanced incomplete block design in Table 11.8, p. 361, used for the detergent experiment in Sect. 11.4.4.

   (a) Show that the least squares estimator $\hat{\tau}_4 - \hat{\tau}_6$ is an unbiased estimator of $\tau_4 - \tau_6$ under the block–treatment model (11.4.2), p. 356 (that is, show that $E[\hat{\tau}_4 - \hat{\tau}_6] = \tau_4 - \tau_6$).
   (b) Calculate a confidence interval for $\tau_4 - \tau_6$ as part of a 95% set of simultaneous confidence intervals for pairwise comparisons, using Tukey's method.

Table 11.28 Percentage rust observed for the rust experiment

Temperature (°F)	Block									
	I	II	III	IV	V	VI	VII	VIII	IX	X
50	12	19				20	10	21	19	
55	18		33		19		18		18	24
60	24	36		35	39	22				28
65			39	45		43	34	42		31
70		45	52	55	48			50	43	

8. **Rust experiment**

   The rust experiment investigated the effect of temperature on the percentage of surface area of a metal sheet exhibiting rust after a given length of time exposed to certain weathering conditions. Five temperatures were examined in the experiment, but only three could be examined at any one time under identical experimental conditions. A balanced incomplete block design was used, formed from two cyclic designs with initial blocks (1, 2, 3) and (1, 2, 4). The data and design are shown in Table 11.28.

   (a) Write down a model for this experiment and test the hypothesis of no difference in the effects of the temperature on the percentage of rust, against the alternative hypothesis that at least two temperatures differ. Use a significance level of 0.01.
   (b) What does "a significance level of 0.01" in part (a) mean?
   (c) Was blocking worthwhile in this experiment?
   (d) Give a formula for a 95% set of simultaneous confidence intervals for pairwise comparisons among the temperatures in this experiment. Calculate, by hand, the interval comparing temperatures 5 and 4 (that is, $\tau_5 - \tau_4$) for illustration. Compare your answer with that obtained from your computer output.
   (e) Test the hypothesis that there is no linear trend in the percentage of rust as the temperature increases.

9. **Balanced incomplete block design**

   An experiment is to be run to compare the effects of four different formulations of a drug to relieve an allergy. In a pilot experiment, four subjects are to be used, and each is to be given a sequence of three of the four drugs. The measurements are the number of minutes that the subject appears to be free of allergy symptoms. Suppose the design shown in Table 11.29 is selected for the experiment.

   (a) Check that this design is a balanced incomplete block design. (Show what you are checking.)
   (b) Show a randomization of this design, explaining the steps of your randomization.
   (c) The experiment was run as described, and the block–treatment model (11.4.2), p. 356, was used to analyze it. Some information for the analysis is shown in Table 11.29. Using this information and (11.4.7), p. 358, show that Q_1, Q_2, Q_3, Q_4 are $-79.333, 81.667, -158.667, 156.333$, respectively.
   (d) Using the information in part (c), give a confidence interval for $\tau_3 - \tau_2$ assuming that it is part of a set of 95% Tukey confidence intervals.
   (e) Test the hypothesis that there is no difference between the effects of the drugs.

Table 11.29 Balanced incomplete block design of Exercise 9 and partial information for its analysis

Block	Levels of treatment factor			Block totals	Treatment totals
I	1	2	3	$B_1 = 417$	$T_1 = 385$
II	1	2	4	$B_2 = 507$	$T_2 = 582$
III	1	3	4	$B_3 = 469$	$T_3 = 329$
IV	2	3	4	$B_4 = 577$	$T_4 = 674$
				$\bar{y}_{..} = 164.1667$	$msE = 3.683$

(f) The typical model and analysis for a balanced incomplete block design is that of parts (c)–(e). Do you think this is a reasonable model and analysis for the experiment described? Why or why not? (Hint: think about the terms in, and assumptions on, the model.)

10. **Step experiment, continued**

   The step experiment was described in Example 11.7.1 and the data are shown in Table 11.15, p. 373.

   (a) Prepare a plot of the treatment averages and examine the linear trends in the heart rate due to step frequency at each level of step height.
   (b) Fit a block–treatment model to the data with $v = 6$ treatments representing the six treatment combinations.
   (c) Estimate the linear trends in the heart rate due to step frequency at each level of step height separately, and calculate confidence intervals for these.
   (d) Write down a contrast that compares the linear trends in part (c) and test the hypothesis that the linear trends are the same against the alternative hypothesis that they are different.

11. **Beef experiment**

   Cochran and Cox (1957) describe an experiment that was run to compare the effects of cold storage on the tenderness of beef roasts. Six periods of storage (0, 1, 2, 4, 9, and 18 days) were tested and coded 1–6. It was believed that roasts from similar positions on the two sides of the animal would be similar, and therefore the experiment was run in $b = 15$ blocks of size $k = 2$. The response y_{hi} from treatment i in block h is the tenderness score. The maximum score is 40, indicating very tender beef. The design and responses are shown in Table 11.30.

   (a) What is the value of λ for this balanced incomplete block design?
   (b) What benefit do you think the experimenters expected to gain by using a block design instead of a completely randomized design?
   (c) Calculate the least squares estimate of $\tau_6 - \tau_1$ and its corresponding variance.
   (d) Calculate a confidence interval for $\tau_6 - \tau_1$ as though it were part of a set of 95% confidence intervals using Tukey's method of multiple comparisons.
   (e) Calculate a confidence interval for the difference of averages contrast

   $$\frac{1}{3}(\tau_4 + \tau_5 + \tau_6) - \frac{1}{3}(\tau_1 + \tau_2 + \tau_3),$$

Table 11.30 Design and data for the beef experiment

Block	Treatment					
	1	2	3	4	5	6
I	7	17				
II			26	25		
III					33	29
IV	17		27			
V		23			27	
VI				29		30
VII	10			25		
VIII		26				37
IX			24		26	
X	25				40	
XI		25		34		
XII			34			32
XIII	11					27
XIV		24	21			
XV				26	32	

Source Data adapted from *Experimental Designs*, Second Edition, by W. G. Cochran and G. M. Cox, Copyright © 1957, John Wiley & Sons, New York.

assuming that you want the overall level of this interval together with the intervals in part (d) to be at least 94%. What does your interval tell you about storage time and tenderness of beef?

12. **Balanced incomplete block design**

(a) Explain under what circumstances you would choose to use a block design rather than a completely randomized design.

(b) For a balanced incomplete block design, why is it incorrect to estimate the difference in the effects of treatments i and p as $\bar{y}_{.i} - \bar{y}_{.p}$? What is the formula for the correct least squares estimate and its corresponding variance if the design of Table 11.2, p. 351, is used?

(c) Suppose that the design of Table 11.2 is used for an experiment with $v = 7$ treatments and $b = 3$ blocks of size $k = 3$, and suppose that

$$\hat{\tau}_1 = 31.6, \ \hat{\tau}_2 = 22.4, \ \hat{\tau}_3 = 17.9, \ \hat{\tau}_4 = 21.5, \ \hat{\tau}_5 = 30.2, \ \hat{\tau}_6 = 18.6, \ \hat{\tau}_7 = 25.5.$$

Calculate a set of 99% confidence intervals for comparing the effect of treatment 1 with the effects of the other treatments. State your conclusions.

13. **Lithium bioavailability experiment**

An experiment described by W. J. Westlake (*Biometrics*, 1974) used a balanced incomplete block design to compare the amount of four formulations of lithium carbonate that were present in the blood serum of subjects several hours after administration. The four levels of the factor "formulation" consisted of 300 mg capsule (level 1), 250 mg capsule with different inert ingredients (level 2), 450mg delayed release capsule (level 3), 300 mg in solution (level 4). Level 4 is the standard (control) treatment. Table 11.31 shows the amounts of lithium carbonate in the blood serum after 3 hours, measured in mEq (milliequivalent which is a measure of the quantity of ions

Table 11.31   Data for the lithum bioavailability experiment, showing treatment factor level, and response in parenthesis

Subject					
1	2	3	4	5	6
1 (.467)	4 (.450)	3 (.200)	2 (.520)	4 (.600)	4 (.467)
2 (.440)	3 (.156)	1 (.533)	3 (.222)	1 (.467)	2 (.520)

Subject					
7	8	9	10	11	12
2 (.600)	2 (.640)	3 (.156)	1 (.533)	1 (.433)	3 (.156)
1 (.467)	4 (.400)	4 (.467)	4 (.433)	3 (.267)	2 (.520)

Source Data adapted from Westlake, Biometric, 1974, International Biometric Society

in an electrolyte fluid). There were 12 subjects, each of whom was given two formulations (one week apart).

(a) Verify that the design in Table 11.31 is a balanced incomplete block design.
(b) Plot the observations on the treatments adjusted for the block (subject) effects), and explain what the plot shows.
(c) Test, at level .01, the hypothesis that the four formulations have the same effect on the amount of lithium carbonate in the blood serum after 3 hours.
(d) Calculate a set of 99% simultaneous confidence intervals for the pairwise comparisons of the formulations and state your conclusion.
(e) Calculate a set of 99% simultaneous confidence intervals for the comparisons of the three test formulations with the control (level 4). State your conclusion.

14. **Plasma experiment, continued**

In the plasma experiment of Sect. 11.5, the experimenters used the cyclic design for six treatments generated by 1, 4, 5.

(a) Show that the cyclic design is also a group divisible design by determining the groups and the values of λ_1 and λ_2.
(b) The design for day one of the experiment consisted of the following three blocks of size three: (1, 4, 5); (2, 5, 6); (3, 6, 1). Show that this design is connected by drawing the connectivity graph of the design.

15. **Group divisible design**

(a) Show that the design in Table 11.32 with $v = 12$ treatments and $b = 9$ blocks of size $k = 4$ is a group divisible design with $g = 4$ groups of size $\ell = 3$. Find the groups and calculate λ_1 and λ_2.
(b) Using the values of λ_1 and λ_2 in part (a), show that the requirements in Sect. 11.3.2, p. 354, for a group divisible design are met.
(c) Fit model (11.4.2), p. 356, to these data and check the assumptions on the error variables by plotting the residuals.
(d) Treatment 1 is the control treatment. Using the formulae in Sect. 11.4.5, p. 365, calculate the two possible variances of $\hat{\tau}_1 - \hat{\tau}_p$, $p = 2, \ldots, 12$.

Table 11.32 Data for the group divisible design

Block	Treatments (Response)			
1	2 (4.5)	4 (8.7)	10 (19.1)	11 (21.2)
2	5 (11.8)	7 (13.6)	9 (18.2)	11 (22.2)
3	1 (4.8)	8 (17.3)	11 (24.0)	12 (24.8)
4	4 (10.8)	6 (14.5)	7 (16.6)	8 (18.1)
5	2 (7.9)	3 (9.7)	8 (19.1)	9 (20.6)
6	2 (8.7)	5 (14.4)	6 (16.2)	12 (27.7)
7	1 (7.5)	6 (16.9)	9 (20.7)	10 (23.5)
8	3 (11.3)	7 (19.8)	10 (25.5)	12 (29.0)
9	1 (9.1)	3 (9.8)	4 (12.9)	5 (16.9)

(e) Calculate the two possible lengths of 99% confidence intervals for treatment versus control contrasts using Scheffé's method, and verify your calculations using a computer program.

(f) What can you conclude about the control treatment in this experiment?

(g) Using (4.2.4), p. 73, find the linear trend coefficients for 12 equally spaced treatment levels. If the treatment code corresponds to milliliters of added sugar to a product, and the response is a taste score, test the hypothesis that the linear trend in score as the amount of sugar increases is negligible. Use a significance level of $\alpha = 0.01$ and a two-sided alternative hyothesis.

16. **Fabric stain experiment**

The experiment was run by M. Nelson, T. Blake, D. Sullivan, A. Kemme and J. Jia in 2008 to examine how best to remove a stain caused by black waterproof drawing ink. There were three factors under consideration as follows. Factor A was type of cloth (1=cotton, 2=polyester), Factor B was type of pre-treatment (1=gel stain remover, 2=antibacterial dish soap, 3=no pre-treatment), and Factor C was time between staining and washing (0, 1, 2 days, coded as levels 1, 2, 3). The experiment was run in several blocks; in some blocks the washing machine was set to the hot wash setting, and the others to the cold wash setting.

Thirty six pieces of fabric of each type were prepared of approximately the same size. The amount of ink applied to each piece of cloth was held constant, as was the amount of detergent used in each wash. The stained fabric samples were evaluated by a single experimenter using a "Gray Scale" (cf. Chapter 10, Exercise 14) both before and after the wash. The differences between the pairs of evaluations formed the responses and are shown in Table 11.33. The order of the samples was randomized and unknown to the evaluator.

(a) Show that, for this design, pairs of treatments occur together in λ_1 or λ_2 or λ_3 blocks, where these values are different. (This means that there are three different lengths of confidence intervals for treatment contrasts). Thus, argue that the design cannot be a balanced incomplete block design or a group divisible design.

(b) For the standard block–treatment model (11.4.2), p. 356, with $b = 6$ blocks and $v = 18$ treatments, write down the contrast coefficient lists for the following contrasts:

(i) the comparison of treating the stain immediately as compared with the average of the 1 and 2 day delays,

Table 11.33 Responses for the fabric stain experiment

Block	Treatments (Response)					
1	1 (35)	6 (35)	8 (20)	10 (80)	15 (65)	17 (10)
2	2 (30)	4 (25)	9 (10)	11 (75)	13 (60)	18 (15)
3	3 (30)	5 (30)	7 (30)	12 (75)	14 (60)	16 (15)
4	1 (20)	5 (30)	9 (15)	10 (80)	14 (65)	18 (20)
5	2 (20)	6 (15)	7 (10)	11 (80)	15 (65)	16 (15)
6	3 (20)	4 (15)	8 (20)	12 (70)	13 (50)	17 (15)

  (ii) the comparison of no pre-treatment versus the average of the two types of pre-treatment,
 (iii) the comparison of a 1 or 2 day delay for each level of pre-treatment separately.

(c) Prepare an analysis of variance table and test the hypothesis that each contrast in part (b) is negligible. Use an overall significance level of not more than 0.1. State your conclusions.
(d) Using Scheffé's method of multiple comparisons at confidence level 95%, state which pairs of treatments are significantly different and give the confidence intervals for these.
(e) From Scheffé's multiple comparisons calculated in part (d), verify that the confidence intervals for $\tau_1 - \tau_p$, for $p = 2, \ldots, 18$, are of three different lengths, and that the shorter lengths correspond to larger values of λ_i.
(f) Using a factorial model for the treatment effects, prepare an analysis of variance table. State your conclusions and, repeat part (c) for the first two contrasts in (b). Use the mapping of treatment labels to levels of A, B, C as:

$$1 = 111 \quad 2 = 112 \quad 3 = 113 \quad 4 = 121 \quad 5 = 122 \quad 6 = 123$$
$$7 = 131 \quad 8 = 132 \quad 9 = 133 \quad 10 = 211 \quad 11 = 212 \quad 12 = 213$$
$$13 = 221 \quad 14 = 222 \quad 15 = 223 \quad 16 = 231 \quad 17 = 232 \quad 18 = 233$$

(g) Does this design have "factorial structure" as defined in Sect. 11.7.1?

17. **Air rifle experiment**

 This is a dangerous experiment that should not be copied! It requires proper facilities and expert safety supervision.

 An experiment was run in 1995 by C.-Y. Li, D. Ranson, T. Schneider, T. Walsh, and P.-J. Zhang to examine the accuracy of an air rifle shooting at a target. The two treatment factors of interest were the projectile type (factor A at levels 1 and 2) and the number of pumps of the rifle (factor B, 2, 6, and 10 pumps, coded 1, 2, 3). The paper covering the target had to be changed after every four observations, and since there were $v = 6$ treatment combinations, coded 11, 12, 13, 21, 22, 23), an incomplete block design was selected.

 Two copies of the following incomplete block design (called a generalized cyclic design) were used:

Table 11.34 Data for the air rifle experiment

Block	Treatment combination (Response)			
I	22 (2.24)	23 (6.02)	12 (11.40)	11 (26.91)
II	13 (7.07)	22 (8.49)	21 (19.72)	11 (24.21)
III	12 (10.63)	23 (6.32)	21 (9.06)	13 (29.15)
IV	11 (11.05)	22 (6.32)	23 (7.21)	13 (23.02)
V	23 (6.71)	22 (15.65)	12 (11.40)	11 (23.02)
VI	21 (17.89)	12 (8.60)	22 (10.20)	13 (10.05)
VII	11 (18.38)	13 (11.18)	23 (11.31)	22 (11.70)
VIII	22 (1.00)	11 (30.87)	21 (20.10)	13 (17.03)
IX	11 (18.03)	21 (8.25)	23 (6.08)	12 (19.24)
X	21 (15.81)	13 (2.24)	12 (17.09)	22 (7.28)
XI	23 (8.60)	21 (15.13)	12 (14.42)	11 (25.32)
XII	13 (8.49)	12 (14.32)	23 (11.66)	21 (17.72)

Block					
I	II	III	IV	V	VI
11	12	13	21	22	23
21	22	23	11	12	13
13	11	12	23	21	22
22	23	21	12	13	11

The total of 12 blocks were randomly ordered, as were the treatment combinations within each block. The data, shown in Table 11.34, are distances from the center of the target measured in millimeters.

(a) Write down a suitable model for this experiment.
(b) Check that the assumptions on your model are satisfied.
(c) The experimenters expected to see a difference in accuracy of the two projectile types. Using a computer package, analyze the data and determine whether or not this was the case.
(d) For each projectile type separately, examine the linear and quadratic trends in the effects of the number of pumps. State your conclusions.

Designs with Two Blocking Factors

<div style="text-align: right">

12

</div>

12.1 Introduction

In Chaps. 10 and 11, we discussed designs for experiments involving a single system of blocks. However, as we saw in the randomized complete block design of Table 10.1, p. 308, a block label can represent a combination of levels of several factors. The design in Table 10.1 was presented as having six blocks—the six block labels being the six combinations of levels of the factors "run of the oven" and "shelf." When a block design is used in this way, the $b − 1$ degrees of freedom for the block effects include not only those degrees of freedom for the effects of the two factors, but also for their interaction.

In this chapter, we look at designs for experiments that involve two blocking factors that *do not* interact. When the blocking factor interactions can be omitted from the model, fewer observations are needed, and the experiment can be designed with only one observation per combination of levels of the blocking factors. The plan of the design is then often written as an array (that is, a table) with the levels of one blocking factor providing the row headings and those of the other providing the column headings. These designs are called *row–column designs*, and the two sets of blocks are called row blocks and column blocks or, more simply, rows and columns. Such designs are described in Sect. 12.2, where we restrict discussion to the case of exactly one treatment label allocated to each cell of the table. The randomization procedure needed for row–column designs is given in Sect. 12.2.1.

In Sect. 12.2.2, Latin square designs are described. These are the two-dimensional counterparts of randomized complete block designs since the row blocks are complete blocks and the column blocks are also complete blocks. Latin square designs require that the numbers of levels of both blocking factors be the same as (or a multiple of) the number of treatments. Section 12.2.3 concerns Youden designs, in which the column blocks form a randomized block design and the row blocks form a balanced incomplete block design (or vice versa). Cyclic and other row–column designs are discussed in Sect. 12.2.4.

The standard row–column–treatment model for row–column designs is shown in Sect. 12.3, together with an overview of the analysis of variance and confidence intervals for general row–column designs. The analysis becomes greatly simplified for both Latin square designs and Youden designs, and this is shown in Sects. 12.4 and 12.5, respectively. Checks on the model assumptions are described briefly in Sect. 12.6, and an extension of the model to cover factorial experiments in row-column designs is given in Sect. 12.7. All row–column designs can be analyzed using statistical software (see Sects. 12.8 and 12.9 for analysis by SAS and R software).

Table 12.1 A row–column experimental plan with $b = 7$, $c = 4$, $v = 7$, $r = 4$

		Column blocks			
		I	II	III	IV
Row blocks	I	1	3	6	7
	II	2	4	7	1
	III	3	5	1	2
	IV	4	6	2	3
	V	5	7	3	4
	VI	6	1	4	5
	VII	7	2	5	6

Some experiments are designed with more than two blocking factors. Block designs with a single system of blocks can still be used where each block represents some combination of the levels of three or more blocking factors, but designs with more than two blocking factors and a single observation per combination of their levels are not discussed in this book.

12.2   Design Issues

12.2.1   Selection and Randomization of Row–Column Designs

In most row–column designs and, in particular, in all Latin square designs and Youden designs, all treatment contrasts are estimable. However, estimability is not as easy to verify for general row-column designs as it was for incomplete block designs (Sect. 11.2.3) since it cannot be deduced from the row blocks and column blocks separately. One way to check that a miscellaneous row–column design is suitable for a planned experiment is to enter a set of hypothetical data for the design into a computer package and see whether the required contrasts are estimable.

Once the numbers of rows, columns, and treatments have been selected, there are two stages to designing an experiment with two blocking factors and one observation at each combination of their levels. The first design stage is to choose an experimental plan. As an example, the plan in Table 12.1 has two sets of blocks corresponding to two blocking factors—one with 7 levels represented by the $b = 7$ rows, and the other with 4 levels represented by the $c = 4$ columns. There are $v = 7$ treatment labels, each appearing $r = 4$ times in the design, and at most once per row and per column. The second design stage is the random assignment of the labels in the design to the levels of the treatment factors and blocking factors in the experiment, as follows:

 (i) The row block labels in the design are randomly assigned to the levels of the first blocking factor.
 (ii) The column block labels in the design are randomly assigned to the levels of the second blocking factor.
(iii) The treatment labels in the design are randomly assigned to the levels of the treatment factor.

Since there is only one experimental unit in each cell, there is no need for random assignment of experimental units to treatment labels within a cell.

Table 12.2 Latin squares
with $b = c = v = 4$

Square 1	Square 2	Square 3	Square 4
A B C D	A B C D	A B C D	A B C D
B C D A	B A D C	B A D C	B D A C
C D A B	C D A B	C D B A	C A D B
D A B C	D C B A	D C A B	D C B A

12.2.2 Latin Square Designs

A $v \times v$ *Latin square* is an arrangement of v Latin letters into a $v \times v$ array (a table with v rows and v columns) in such a way that each letter occurs once in each row and once in each column. For example, the following 3×3 array is a 3×3 Latin square:

$$
\begin{array}{ccc}
A & B & C \\
B & C & A \\
C & A & B
\end{array}
$$

A *Latin square design* has v treatment labels, and v^2 experimental units arranged in v row blocks and v column blocks, where experimental units within each row block are alike, units within each column block are alike, and units not in the same row block or column block are substantially different. The experimental plan of the design is a $v \times v$ Latin square. Randomization of row block, column block, and treatment labels in the plan is carried out as in Sect. 12.2.1.

If we look only at the row blocks of a Latin square design, ignoring the column blocks, we have a randomized complete block design, and if we look at the column blocks alone, ignoring the row blocks, we also have a randomized complete block design. Each level of the treatment factor is observed $r = v$ times—once in each row block and once in each column block.

The 3×3 Latin square shown above is a "standard, cyclic Latin square." A Latin square is a *standard Latin square* if the letters in the first row and in the first column are in alphabetical order, and it is *cyclic* if the letters in each row can be obtained from those in the previous row by cycling the letters in alphabetical order (cycling back to letter A after the vth letter).

There is only one standard 3×3 Latin square, but there are four standard 4×4 Latin squares, and these are shown in Table 12.2. The first square is the cyclic standard Latin square. A standard cyclic Latin square exists for any number of treatments.

An example of a 6×6 Latin square design was shown in Table 2.5 (p. 19). It was a design that was considered for the cotton-spinning experiment. The row blocks represented the different machines with their attendant operators, and the column blocks represented the different days over which the experiment was to be run. The treatment labels were the combinations of degrees of twist and flyer used on the cotton-spinning machines. After careful consideration, the experimenters decided not to use this design, since it required the same six machines to be available over the same six days, and this could not be guaranteed in the factory setting.

Latin square designs are often used in experiments involving subjects, especially where the subjects are allocated a sequence of treatments over time and where the time effect is thought to have a major effect on the response. For example, in an experiment to compare the effects of v drugs, the rows of the Latin square might correspond to v subjects to whom the drugs are given, and the columns might correspond to v time periods, with each subject receiving one drug during each time period. An experiment of this type, in which each subject receives a sequence of treatments, is called a *crossover experiment*.

In a crossover experiment, it is possible that a treatment will continue to have some effect on the response when another treatment is administered to the same subject in a subsequent time period. Such an effect is called a *carryover effect* or *residual effect* and must be accounted for in the design and analysis of the experiment. This is outside the scope of this book, but for further information, see Ratkowsky et al. (1993) or Jones and Kenward (2003). We will consider experiments in which either there is no carryover effect or in which the gap between the administration of one treatment and the next is sufficient to allow the carryover effect to diminish (and, hopefully, to disappear). The "gap" is called a *washout period*.

To design the experiment, a square is selected from the list of standard squares and the randomization procedure of Sect. 12.2.1 is performed. Thus for an experiment with $b = 4$ levels of the row blocking factor, $c = 4$ levels of the column blocking factor, and $v = 4$ levels of the treatment factor, one of the four standard squares of Table 12.2 would be selected. For larger squares, we have not provided a list of standard squares, since there are so many. However, it is straightforward to write down the standard cyclic Latin square and perform the randomization procedure on this.

Although we have not given a procedure which selects a Latin square at random from the set of all possible Latin squares, the above procedure should be sufficient for the occasional experiment. Fisher and Yates (1973) give more information on this, as well as a complete list of standard squares for $v = 5$ and 6, and selected squares for $v = 7$–12.

Replication of Latin Squares

A design based on a single Latin square has $r = v$ observations on each treatment, which may not be adequate for an analysis of the treatment effects. One way of increasing the number of observations is to piece together a number, s say, of $v \times v$ Latin squares. We will call such a design an *s-replicate Latin square*. Use of an s-replicate Latin square requires the column (or row) blocks to be of size vs. For example, stacking two 3×3 Latin squares, one above the other, we can obtain two possible 2-replicate Latin squares as in Table 12.3. For either plan, the number of observations per treatment is now $r = 6 = 2v$ rather than only $r = 3 = v$. Plan 2 is preferable to Plan 1 because the row blocks consist of each possible ordering of the three treatments (and this will remain true even after randomization).

Suppose we were to use Plan 2 for an experiment to compare the efficacy of $v = 3$ drugs, where there are two blocking factors, say "subjects"—the people to whom the drugs will be administered, and "time period"—the time during which each subject receives a single drug and a response is measured. If rows correspond to subjects, then $b = 6$ subjects would be required over $c = 3$ time periods, but if columns correspond to subjects, then each of $c = 3$ subjects would stay in the study for $b = 6$ periods. In practice, in drug studies, subjects are rarely used for more than 4 time periods, since the subject drop-out rate tends to be high after this length of time.

An s-replicate Latin square may allow some column-treatment interaction contrasts to be measured (after adjusting for row block effects as in Chap. 11). However, these will often not form a full set of

Table 12.3 Two 2-replicate Latin squares for $v = 3, b = 6, c = 3$

	Plan 1			Plan 2	
A	B	C	A	B	C
B	C	A	B	C	A
C	A	B	C	A	B
A	B	C	A	C	B
B	C	A	B	A	C
C	A	B	C	B	A

Table 12.4 Two Youden squares for $v = 4$, $b = 4$, $c = 3$

	Plan 3			Plan 4	
A	B	C	A	C	B
B	C	D	B	D	C
C	D	A	C	A	D
D	A	B	D	B	A

$(v-1)^2$ orthogonal contrasts. For $v = 3$, Plan 2 does allow the adjusted column-treatment interaction to be measured based on the full set of $(v-1)^2 = 4$ degrees of freedom, but Plan 1 does not.

12.2.3 Youden Designs

A $v \times c$ *Youden square* is an arrangement of v Latin letters into a $v \times c$ array (with $c < v$) in such a way that each letter occurs once in each column and at most once in each row. In addition, every pair of treatments occurs in the same number of rows, so the columns form complete blocks and the rows form a balanced incomplete block design.

Notice that a Youden square, written in this way, is not, in fact, square! We have defined a Youden square to have more rows than columns, but the array could be turned so that there are more columns than rows. Plans 3 and 4 in Table 12.4 are examples of 4×3 Youden squares with $v = 4$ treatment labels, with Plan 3 being a standard Youden square.

In general, a *Youden design* has v treatment labels and vc experimental units arranged in $b = v$ row blocks and $c < v$ column blocks, where experimental units within each row block are alike, experimental units within each column block are alike, and experimental units not in the same row block or column block are substantially different. The experimental plan is a $v \times c$ Youden square, and randomization of row, column, and treatment labels is carried out as in Sect. 12.2.1. Each level of the treatment factor is observed $r = c$ times—once in each column and at most once in each row.

The plans of Table 12.4 are both *cyclic Youden designs* prior to randomization. In each case, the row blocks form a cyclic balanced incomplete block design with every pair of treatment labels occurring in $\lambda = 2$ rows. Any cyclic balanced incomplete block design (with the full set of v blocks) can be used as a cyclic Youden design. The design in Table 12.1 is also a cyclic Youden design, having $b = v = 7$ row blocks and $c = 4$ column blocks and every pair of treatments occurring in $\lambda = 2$ row blocks. Youden designs with $c = v - 1$ column blocks can be obtained by deleting any column from a $v \times v$ Latin square design.

Replication of Youden Squares

We can obtain $r = cs$ observations per treatment by stacking s Youden squares one above another (giving $b = vs$ row blocks and $c(<v)$ column blocks). This results in large column-block sizes but may allow for estimation of some column×treatment interaction contrasts after adjusting for row block effects (although usually not the full set of $(c-1)(v-1)$ contrasts). The row blocks still form a balanced incomplete block design, and the column blocks still form a complete block design.

Suppose $v = 7$ drugs are to be compared and that a number of subjects are each to be given a sequence of 4 of the drugs over $c = 4$ time periods, with washout periods in between to avoid carryover effects. The experimental plan in Table 12.1, which uses 7 subjects, would be suitable for the experiment if $r = 4$ observations per drug is deemed adequate. Otherwise, several copies of this design could be stacked. Or column permutations could be made before stacking (cf. stacking the two plans in Table 12.4). If, for example, $s = 3$ copies of the Table 12.1 design are pieced together, the resulting 3–replicate Youden design would require 21 subjects, 4 time periods, and would have $r = cs = 12$

	Plan 5				Plan 6		
	I	II	III		I	II	III
I	1	2	4	I	1	2	7
II	2	3	5	II	2	4	1
III	3	4	6	III	3	7	6
IV	4	5	7	IV	4	8	2
V	5	6	8	V	5	6	8
VI	6	7	1	VI	6	3	5
VII	7	8	2	VII	7	1	3
VIII	8	1	3	VIII	8	5	4

Table 12.5 Incomplete-row complete-column designs with $v = b = 8$ and $r = c = 3$

observations per treatment, with $b = 21$, $c = 4$, $v = 7$, and $\lambda = 6$. The design would be randomized before use as described in Sect. 12.2.1.

12.2.4 Cyclic and Other Row–Column Designs

Any arrangement of v treatment labels into b rows and c columns can be used as a row–column design for an experiment with two blocking factors. As with incomplete block designs, some row–column designs are better than others. The better designs (in terms of average length of confidence intervals for pairwise comparisons) have every pair of treatment labels occurring the same, or nearly the same, number of times in the row blocks and also in the column blocks. This is satisfied by Latin square designs, Youden designs, and also some cyclic designs.

A *cyclic row–column design with complete column blocks* is a row–column design in which the row blocks form a cyclic block design and the column blocks are complete blocks. For example, the experimental plan in Table 12.1 is a cyclic row–column design. The class of cyclic row–column designs is very large, and a design with $b = vs$ rows can always be found when a Youden design does not exist. For example, consider an experiment for comparing $v = 8$ treatments when there are two blocking factors having $b = 8$ and $c = 3$ levels, respectively (so $r = c = 3$). There does not exist a Youden design for $b = v = 8$ and $c = 3$, because there does not exist a balanced incomplete block design for 8 treatments in 8 blocks of size $c = 3$. (Notice that (11.3.1) gives $\lambda = r(c - 1)/(v - 1) = 6/7$ which is not an integer.) A cyclic design that may be suitable for the experiment is given as Plan 5 in Table 12.5. The design is obtained by cycling the labels in the first row block. Treatment pairs (1, 5), (2, 6), (3, 7), and (4, 8) never occur together in a row block, but all other pairs of treatments occur in exactly one row block, so this should be a reasonably good design. The design should be randomized via the procedure in Sect. 12.2.1 before it is used.

Plan 6 of Table 12.5, which is neither a Youden design nor a cyclic design, could also be used, but we might guess that it is not quite as good for pairwise comparisons. Treatment 1, for example, appears in one block with each of treatments 3 and 4, in two blocks with 2 and 7, and not at all with treatments 5, 6, or 8. Thus, the treatments are not quite so evenly distributed in the row blocks of Plan 6 as they are of Plan 5. Nevertheless, Plan 6 was used for the exercise bike experiment described below, and will be analyzed using the SAS and R software in Sects. 12.8 and 12.9, respectively.

Example 12.2.1 Exercise bicycle experiment

Yuedong Wang, Dong Xiang, and Yili Lu conducted an experiment in 1992 at the University of Wisconsin to investigate the effects of exercise on pulse rate. The exercise was performed on a stationary

Table 12.6 Row–column design showing treatments and data for the exercise bicycle experiment

Day	Subject					
	Lu		Wang		Xiang	
1	112	45	212	25	221	18
2	222	27	211	20	212	32
3	212	40	222	23	112	28
4	221	17	112	32	122	24
5	211	30	111	36	222	20
6	122	29	221	13	121	20
7	111	34	121	18	211	25
8	121	21	122	22	111	34

bicycle that included both foot pedals and hand bars. The experiment involved three treatment factors each at two levels. These were "time duration of exercise" with levels 1 and 3 min (coded 1, 2), "exercise speed" with levels 40 and 60 rpm (coded 1, 2), and "pedal type" with levels foot pedal and hand bars (coded 1, 2).

The three experimenters served as the $b = 3$ subjects in the experiment, since they were interested in the effects of the different exercises on themselves rather than on a large population. Had the latter been the case, then the subjects would have been selected at random from the population of interest and regarded as "random effects" (see Chap. 17).

Each subject was assigned a different exercise on each of 8 different days. To minimize any carryover effect, there was a training period prior to the experiment, and there was at least one day of rest between observations. The subject's pulse was taken immediately after completion of the exercise, and the response variable was the time, in seconds, for 50 heart beats.

Plan 6 of Table 12.5 was used. The design, after randomization, and the corresponding data are shown in Table 12.6. The rows were randomly assigned to days in the order V, III, VI, VIII, VII, IV, I, II. The columns were randomly assigned to subjects in the order I = Lu, II = Wang, III = Xiang. The treatment labels were randomly assigned to treatment combinations in the order

$$
\begin{array}{cccccccc}
1 & 5 & 2 & 4 & 7 & 6 & 8 & 3 \\
111 & 112 & 121 & 122 & 211 & 212 & 221 & 222
\end{array}
$$

The data for this experiment are analyzed by computer software in Sects. 12.8 and 12.9.   □

12.3   Analysis of Row–Column Designs

12.3.1   Model for a Row–Column Design

For a row–column design with b rows and c columns, the row–column–treatment model is

$$Y_{hqi} = \mu + \theta_h + \phi_q + \tau_i + \epsilon_{hqi}, \tag{12.3.1}$$
$$\epsilon_{hqi} \sim N(0, \sigma^2),$$
$$\epsilon_{hqi}\text{'s are mutually independent},$$
$$h = 1, \ldots, b; \quad q = 1, \ldots, c; \quad i = 1, \ldots, v; \quad (h, q, i) \text{ in the design},$$

Table 12.7   Analysis of variance table for a connected row–column design with no interactions

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio
Rows (unadj)	$b - 1$	$ss\theta$	—	—
Columns (unadj)	$c - 1$	$ss\phi$	—	—
Treatments(adj)	$v - 1$	ssT_{adj}	msT_{adj}	msT_{adj}/msE
Error	$bc - b - c - v + 2$	ssE	msE	
Total	$bc - 1$	$sstot$		

<div align="center">Formulae</div>

$$Q_i = T_i - \tfrac{1}{c}\Sigma_{h=1}^{b} n_{h.i} B_h - \tfrac{1}{b}\Sigma_{q=1}^{c} n_{.qi} C_q + \tfrac{r}{bc} G \qquad ssT_{\text{adj}} = \Sigma_{i=1}^{v} Q_i \hat{\tau}_i$$

$$sstot = \Sigma_{h=1}^{b}\Sigma_{q=1}^{c}\Sigma_{i=1}^{v} n_{hqi} y_{hqi}^2 - \tfrac{1}{bc} G^2 \qquad ss\theta = \tfrac{1}{c}\Sigma_{h=1}^{b} B_h^2 - \tfrac{1}{bc} G^2$$

$$ssE = sstot - ss\theta - ss\phi - ssT_{\text{adj}} \qquad ss\phi = \tfrac{1}{b}\Sigma_{q=1}^{c} C_q^2 - \tfrac{1}{bc} G^2$$

$$T_i = \Sigma_{h=1}^{b}\Sigma_{q=1}^{c} n_{hqi} y_{hqi} \qquad B_h = \Sigma_{q=1}^{c}\Sigma_{i=1}^{v} n_{hqi} y_{hqi}$$

$$C_q = \Sigma_{h=1}^{b}\Sigma_{i=1}^{v} n_{hqi} y_{hqi} \qquad G = \Sigma_{h=1}^{b}\Sigma_{q=1}^{c}\Sigma_{i=1}^{v} n_{hqi} y_{hqi}$$

where θ_h, ϕ_q, and τ_i denote the effects on the response of row block h, column block q, and treatment i, respectively. The term "(h, q, i) in the design" means that the model is valid for whichever treatment i is observed in the cell defined by the hth row block and qth column block. The model includes the usual assumptions that the error terms ϵ_{hqi} are independent and normally distributed with constant variance. It also includes the assumptions of no interaction between the row and column blocking factors nor between these and the treatment factors.

12.3.2   Analysis of Variance for a Row–Column Design

We first give an overview of the analysis for connected row–column designs with no interactions. This is followed by the specific analyses for Latin square designs and Youden designs in Sects. 12.4 and 12.5, respectively. Analysis of general row–column designs can be done by computer, and this is illustrated via the SAS and R computer packages in Sects. 12.8 and 12.9.

For row–column designs with b row blocks, c column blocks, and r observations on each of v treatments with the row–column–treatment model (12.3.1), the top portion of Table 12.7 shows the form of the analysis of variance obtained when the row and column factors are entered into the model before the treatment factor. The (unadjusted) sum of squares $ss\theta$ for row blocks, and the (unadjusted) sum of squares $ss\phi$ for column blocks are calculated in a similar way to the block sum of squares in a complete block design; that is

$$ss\theta \;=\; \tfrac{1}{c}\Sigma_{h=1}^{b} B_h^2 - \tfrac{1}{bc} G^2 \quad \text{and} \quad ss\phi \;=\; \tfrac{1}{b}\Sigma_{q=1}^{c} C_q^2 - \tfrac{1}{bc} G^2$$

where B_h is the total of the observations in the hth row block, C_q is the total of the observations in the qth column block, and G is the grand total of all the observations.

The sum of squares for treatments adjusted for rows and columns (i.e. adjusted for the fact that not all treatments may be in every row block and every column block, and some blocks are better than others) can be shown to be

$$ssT_{\text{adj}} = \Sigma_{i=1}^{v} Q_i \hat{\tau}_i \ \text{ with } \ Q_i = T_i - \frac{1}{c}\sum_{h=1}^{b} n_{h.i}B_h - \frac{1}{b}\sum_{q=1}^{c} n_{.qi}C_q + \frac{r}{bc}G, \tag{12.3.2}$$

where $n_{h.i}$ is the number of times that treatment i appears in row block h, and and $n_{.qi}$ is the number of times that treatment i appears in column block q. In general, it is not easy to obtain least squares solutions by hand for the treatment parameters τ_i in row–column designs except for the cases of Latin square and Youden designs. Least squares solutions for these special cases are given in Sects. 12.4 and 12.5, respectively. The analysis of a more complicated row–column design is illustrated using the SAS and R computer packages in Sects. 12.8 and 12.9, respectively.

In general, the sum of squares for error for the row–column–treatment model (12.3.1) is obtained by subtraction as usual

$$ssE = sstot - ss\theta - ss\phi - ssT_{\text{adj}}, \tag{12.3.3}$$

where

$$sstot = \sum_h \sum_q \sum_i n_{hqi} y_{hqi}^2 - \frac{1}{bc}G^2,$$

and $n_{hqi} = 1$ if treatment label i is allocated to the combination of row block h and column block q and zero otherwise. As in any connected block design, the numbers of degrees of freedom for rows, columns and treatments are $b - 1$, $c - 1$ and $v - 1$, respectively. The number of degrees of freedom for error is obtained by subtraction:

$$df = (bc - 1) - (b - 1) - (c - 1) - (v - 1) = bc - b - c - v + 2, \tag{12.3.4}$$

where bc is the total number of observations.

A test of the null hypothesis $H_0^T : \{\tau_1 = \tau_2 = \cdots \tau_v\}$ against the general alternative hypothesis $H_A^T : \{$at least two of the τ_i's differ$\}$ is given by the decision rule

$$\text{reject } H_0^T \text{ if } msT_{\text{adj}}/msE > F_{v-1,bc-b-c-v+2,\alpha} \tag{12.3.5}$$

for some chosen significance level α. The test (12.3.5) is most conveniently set out in an analysis of variance table, as in the top portion of Table 12.7. The bottom portion of the table shows the specific formulae.

12.3.3   Confidence Intervals and Multiple Comparisons

The multiple-comparison methods of Bonferroni and Scheffé can be applied for any row–column design. Simultaneous $100(1 - \alpha)\%$ confidence intervals for treatment contrasts $\Sigma_i d_i \hat{\tau}_i$ are of the form

$$\left(\sum_i d_i \hat{\tau}_i \pm w \sqrt{\widehat{\text{Var}}(\sum_i d_i \hat{\tau}_i)}\right), \tag{12.3.6}$$

where the critical coefficient w is

$$w_B = t_{bc-b-c-v+2,\alpha/2m} \text{ or } w_S = \sqrt{(v-1)F_{v-1,bc-b-c-v+2,\alpha}},$$

for the Bonferroni or Scheffé methods, respectively. Tukey and Dunnett methods of pairwise comparisons can be used in the special cases of Latin square designs and Youden designs (Sects. 12.4.2 and 12.5.2).

12.4   Analysis of Latin Square Designs

12.4.1   Analysis of Variance for Latin Square Designs

Table 12.7, p. 406, shows the analysis of variance table for the row–column–treatment model (12.3.1) for any row–column design. In the table, T_i, B_h, C_q, and G are, respectively, the sum of the observations on the ith level of the treatment factor, the hth level of the row blocking factor, the qth level of the column blocking factor, and the grand total of all the observations. The constant n_{hqi} is equal to 1 if treatment i is observed in the combination of row block h and column block q, and n_{hqi} is equal to zero otherwise. For an s-replicate Latin square design, every treatment appears once in every row block and s times in every column block, so there are $b = vs$ rows, $c = v$ columns, and we have $n_{h.i} = 1$, $n_{.qi} = s$. It can be shown that, for Latin square designs, least squares solutions for the treatment parameters τ_i in model (12.3.1) are given by

$$\hat{\tau}_i = \frac{1}{vs} Q_i, \quad \text{with} \quad Q_i = T_i - \frac{1}{c} \sum_h n_{h.i} B_h = T_i - \frac{1}{v} G. \tag{12.4.7}$$

Thus, for an s-replicate Latin square design, the treatment sum of squares (12.3.2) become

$$ssT_{\text{adj}} = \sum_{i=1}^v Q_i \hat{\tau}_i = \frac{1}{vs} \sum_{i=1}^v \left(T_i - \frac{1}{v} G \right)^2,$$

and we see that the "adjusted" treatment sum of squares for a Latin square design is actually *not* adjusted for row-block effects nor for column-block effects. This is because every treatment is observed the same number of times in every row block and in every column block (even though only bc of the bcv row–column–treatment combinations are actually observed). The computational formula for ssT_{adj} is obtained by expanding the terms in parentheses to give

$$ssT_{\text{adj}} = \frac{1}{vs} \sum_i T_i^2 - \frac{1}{v^2 s} G^2.$$

The error sum of squares and degrees of freedom are obtained by subtraction, as in (12.3.3). For an s-replicate Latin square design with $b = vs$ rows and $c = v$ columns, we have error degrees of freedom equal to

$$df = bc - b - c - v + 2 = v^2 s - vs - 2v + 2 = (vs - 2)(v - 1). \tag{12.4.8}$$

The test for equality of treatment effects compares the ratio of the treatment and error mean squares with the corresponding value from the F-distribution, in the usual way (see Example 12.4.1). We will not utilize a test for the hypothesis of negligible row-block or of negligible column-block effects. However, we conclude that the current experiment has benefited from the use of the row (or column) blocking factor if the row (or column) mean square exceeds the mean square for error.

Table 12.8 Unrandomized 2-replicate Latin Square (Plan 2, Table 12.3), and possible randomization for the dairy cow experiment (with data in parentheses)

Row	Column			Cow	Period		
	1	2	3		1	2	3
1	A	B	C	6	2 (63)	3 (101)	1 (1)
2	B	C	A	4	3 (76)	1 (86)	2 (46)
3	C	A	B	2	3 (86)	2 (109)	1 (39)
4	A	C	B	5	1 (35)	2 (75)	3 (34)
5	B	A	C	1	2 (25)	1 (38)	3 (15)
6	C	B	A	3	1 (72)	3 (124)	2 (27)

Source Data adapted from *Experimental Designs*, Second Edition, by Cochran and Cox (1957), Copyright © 1957, John Wiley & Sons, New York.

Example 12.4.1 `Dairy cow experiment`

Cochran and Cox (1957, p. 135) described an experiment that studied the effects of three diets on the milk production of dairy cows. The $v = 3$ diets (levels of a treatment factor) consisted of roughage (level 1), limited grain (level 2), and full grain (level 3). A crossover experiment was used, with each of the $b = 6$ cows being fed each diet in turn, each for a six-week period. The response variable was the milk yield for each cow during each of the $c = 3$ periods. Data were collected using a randomization of the 2-replicate Plan 2 Latin square in Table 12.3, p. 403, with $v = 3$ treatment labels, $b = vs = 6$ rows, $c = 3$ columns, and with $v = 3$ treatments each observed $r = vs = 6$ times. A possible randomization and the data collected are shown in Table 12.8.

Provided that each measurement on milk yield has been taken after a cow has been on a new diet for a sufficiently long period of time to allow the effect of the previous diet to "wash out" of the system, we need not be concerned with carryover effects. For this experiment, a model without carryover effects describes the data fairly well.

Table 12.9 shows the analysis of variance table. To test the null hypothesis H_0^T that all three diets have the same effect, the decision rule is

$$\text{reject } H_0^T \text{ if } msT/msE > F_{2,8,.01} = 8.65,$$

with a probability of $\alpha = 0.01$ of making a Type I error. Since $msT/msE = 1138.39/103.06 = 11.05 > 8.65$, we reject the null hypothesis at the $\alpha = 0.01$ significance level and conclude that the diets do not all have the same effect.

To evaluate whether or not blocking was worthwhile, observe that the mean squares for periods and cows are both considerably larger than the error mean square. Thus, inclusion of both blocking factors in the experiment was beneficial in reducing the experimental error. (Since this is a Latin square design, no adjustment is needed to the row and column sums of squares to draw this conclusion.) □

12.4.2 Confidence Intervals for Latin Square Designs

All treatment contrasts are estimable in Latin square designs. Using the row–column–treatment model (12.3.1), the least squares estimate of a treatment contrast $\Sigma d_i \tau_i$, with $\Sigma d_i = 0$, is obtained from (12.4.7) as

$$\sum d_i \hat{\tau}_i = \frac{1}{vs} \sum d_i T_i, \tag{12.4.9}$$

Table 12.9 Analysis of variance table for the dairy cow experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Cow (row)	5	5781.11	1156.22	–	–
Period (column)	2	11480.11	5740.06	–	–
Diets (treatment)	2	2276.78	1138.39	11.05	0.0050
Error	8	824.44	103.06		
Total	17	20362.44			

and, since T_i is the sum of vs observations, the corresponding variance is

$$\text{Var}\left(\sum d_i \hat{\tau}_i\right) = \sum d_i^2 \left(\frac{\sigma^2}{vs}\right), \tag{12.4.10}$$

The multiple-comparison methods of Bonferroni, Scheffé, Tukey, and Dunnett can be used for s-replicate Latin square designs. Formulae for confidence intervals for treatment contrasts $\Sigma d_i \tau_i$ are of the form

$$\left(\frac{1}{vs}\Sigma_i d_i T_i \ \pm \ w \sqrt{msE\left(\frac{\Sigma d_i^2}{vs}\right)}\right), \tag{12.4.11}$$

where the appropriate critical coefficients w for the four methods are,

$$w_B = t_{(vs-2)(v-1), \alpha/2m} \ ; \ \ w_S = \sqrt{(v-1)F_{v-1,(vs-2)(v-1),\alpha}} \ ;$$
$$w_T = q_{v,(vs-2)(v-1),\alpha}/\sqrt{2} \ ; \ \ w_{D2} = |t|^{(0.5)}_{v-1,(vs-2)(v-1),\alpha} \ .$$

Example 12.4.2 `Dairy cow experiment, continued`

The dairy cow experiment was described in Example 12.4.1, p. 409. Tukey's method of all pairwise comparisons can be applied to determine which diets, if any, are significantly different from any of the others. The treatment sample means, which can be computed from the data in Table 12.8, are

$$\frac{1}{vs}T_1 = 45.167, \quad \frac{1}{vs}T_2 = 57.500, \quad \frac{1}{vs}T_3 = 72.667,$$

so that the least squares estimates of the pairwise differences are

$$\hat{\tau}_2 - \hat{\tau}_1 = 57.500 - 45.167 = 12.333,$$
$$\hat{\tau}_3 - \hat{\tau}_1 = 72.667 - 45.167 = 27.500,$$
$$\hat{\tau}_3 - \hat{\tau}_2 = 72.667 - 57.500 = 15.167.$$

The value of the error mean square is $msE = 103.06$ from the analysis of variance table, Table 12.9. The error degrees of freedom are given by (12.4.8) as

$$df = (vs - 2)(v - 1) = (2 \times 3 - 1)(3 - 1) = 8.$$

Using a simultaneous confidence level of 95%, we obtain the critical coefficient for Tukey's method as $w_T = q_{3,8,.05}/\sqrt{2} = 4.04/\sqrt{2}$. Hence, using (12.4.11), the minimum significant difference for

pairwise differences is

$$msd = (4.04/\sqrt{2})\sqrt{msE(2/6)} = 16.743 \,.$$

The simultaneous 95% confidence intervals are therefore

$$\tau_3 - \tau_1 \in (27.500 \ \pm \ 16.743) \ = \ (\ 10.58, 44.24) \,,$$
$$\tau_2 - \tau_1 \in (12.333 \ \pm \ 16.743) \ = \ (-4.41, 29.08) \,,$$
$$\tau_3 - \tau_2 \in (15.167 \ \pm \ 16.743) \ = \ (-1.58, 31.91) \,.$$

From these intervals, we can deduce that at overall significance level $\alpha = 0.05$, the full grain diet (level 3) results in a mean yield of 10.58–44.24 units higher than the roughage diet (level 1), but the limited grain diet (level 2) is not significantly different from either of the other two.   □

12.4.3   How Many Observations?

The formula for determining the sample size needed to achieve a power $\pi(\Delta)$ of detecting a difference Δ in the treatment effects for given v, α, and σ^2, using a Latin square design, is the same as (10.6.13) for a randomized complete block design, but with b replaced by v (since $r = vs$ rather than $r = bs$). Thus, after simplification of the formula, we need to find s to satisfy

$$s \ \geq \ \frac{2\sigma^2\phi^2}{\Delta^2} \,. \tag{12.4.12}$$

Alternatively, the confidence interval formula (12.4.11) can be used to calculate the sample sizes needed for achieving confidence intervals of a desired width (see Example 12.4.3).

Example 12.4.3 `Sample-size calculation for an` s`-replicate Latin square`
`design`

Consider an experiment to compare $v = 3$ computer keyboard designs with respect to the time taken to type an article of given length. Typists and time periods are the two blocking factors, with each of $b = 3s$ typists using each of the keyboards in a sequence of $c = 3$ time periods. An s-replicate Latin square design will be used, with $r = 3s$ observations per keyboard layout (treatment), and with sufficient time between observations to prevent a carryover effect.

   With 3 treatments, there are 3 pairwise comparisons. Using Tukey's method of multiple comparisons and a simultaneous confidence level of 95%, suppose that the experimenters want confidence intervals with half-width (minimum significant difference) of 10 min or less. The error standard deviation is expected to be at most 15 min (a variance of at most 225 min^2). The error degrees of freedom, as given in (12.4.8), are

$$df = (vs - 2)(v - 1) = 2(3s - 2) = 6s - 4 \,.$$

   Then, using the confidence interval formula (12.4.11) for Tukey's method of pairwise comparisons, the minimum significant difference is

$$msd \approx \frac{q_{3,6s-4,.05}}{\sqrt{2}}\sqrt{\frac{225 \times 2}{vs}} = q_{3,6s-4,.05}\sqrt{\frac{225}{3s}} \,.$$

To obtain $msd \le 10$, we require $0.75q^2_{3,6s-4,.05} \le s$. Sample size is then computed by trial and error as follows:

s	$6s - 4$	$q_{3,6s-4,.05}$	$0.75q^2_{3,6s-4,.05}$	Required
∞		3.31	8.22	$s \ge 9$
9	50	3.42	8.77	$s = 9$

So, $s = 9$, and a 9-replicate Latin square would be needed, giving $msd \approx 3.42\sqrt{\frac{225}{27}} = 9.87\,\text{min}$ and requiring $b = vs = 27$ typists and $r = vs = 27$ observations per keyboard layout. $\square$

12.5   Analysis of Youden Designs

12.5.1   Analysis of Variance for Youden Designs

The s-replicate Youden design with v treatments has complete column blocks of size $b = vs$ and row blocks forming a balanced incomplete block design with blocks of size c. Each treatment is observed $r = cs$ times. For the row–column–treatment model (12.3.1) with no interactions, a solution to the normal equations can be shown to be similar to the solution (11.4.9), p. 359, for a balanced incomplete block design; that is,

$$\hat{\tau}_i = \frac{c}{\lambda v}Q_i, \quad \text{where } Q_i = T_i - \frac{1}{c}\sum_h n_{h.i}B_h,\tag{12.5.13}$$

for $i = 1, \ldots, v$, where $\lambda = r(c - 1)/(v - 1)$.

As in the analysis of a balanced incomplete block design, the estimators of the treatment effects in a Youden design are not independent of the estimators of the row-block effects. As a result, the treatment-effect estimators must be adjusted for row blocks. However, since the column blocks form a randomized complete block design, no adjustment is needed for these. Table 12.7, p. 406, shows the analysis of variance table for a row–column design with no interactions. Using (12.5.13), the adjusted treatment sum of squares given in the table becomes

$$ssT_{\text{adj}} = \sum_{i=1}^{v} Q_i\hat{\tau}_i = \frac{c}{\lambda v}\sum_{i=1}^{v} Q_i^2,$$

and, because of the complete column blocks, this is exactly the same formula as for a balanced incomplete block design with blocks of size $k = c$ (see Table 11.7, p. 358).

To test the null hypothesis H_0^T of no treatment differences, at significance level α, the decision rule is

$$\text{reject } H_0^T \quad \text{if } \frac{msT_{\text{adj}}}{msE} > F_{v-1,df,\alpha},$$

where, as given in Table 12.7, the number of error degrees of freedom for an s-replicate Youden square is

$$\begin{aligned}
df &= bc - b - c - v + 2\\
&= vsc - vs - c - v + 2\\
&= (vs - 1)(c - 1) - (v - 1).
\end{aligned}\tag{12.5.14}$$

12.5.2   Confidence Intervals for Youden Designs

All treatment contrasts $\Sigma d_i \tau_i$ are estimable in all Youden designs. The least squares estimate of the contrast $\Sigma d_i \tau_i$ is

$$\sum_i d_i \hat{\tau}_i \ = \ \frac{c}{\lambda v} \sum_i d_i Q_i \ = \ \left(\frac{v-1}{vs(c-1)}\right) \sum_i d_i Q_i \,,$$

since $\lambda = r(c-1)/(v-1)$ as for a balanced incomplete block design with $k = c$. The variance of the corresponding estimator is

$$\mathrm{Var}(\sum_i d_i \hat{\tau}_i) \ = \ \left(\frac{c}{\lambda v}\right) \sum_i d_i^2 \sigma^2 \ = \ \left(\frac{v-1}{vs(c-1)}\right) \sum_i d_i^2 \sigma^2 \,. \tag{12.5.15}$$

Confidence interval formulae for treatment contrasts $\Sigma d_i \tau_i$ are the same as those for a balanced incomplete block design with block size $k = c$; that is,

$$\left(\frac{v-1}{vs(c-1)}\right) \sum_i d_i Q_i \ \pm \ w \sqrt{msE \left(\frac{v-1}{vs(c-1)}\right) \sum d_i^2} \,, \tag{12.5.16}$$

where w, as usual, is the critical coefficient for the Bonferroni, Scheffé, Tukey, or Dunnett method, given by

$$w_B = t_{(vs-1)(c-1)-(v-1),\alpha/2m} \ ; \ \ w_S = \sqrt{(v-1)F_{v-1,(vs-1)(c-1)-(v-1),\alpha}} \ ;$$
$$w_T = q_{v,(vs-1)(c-1)-(v-1),\alpha}/\sqrt{2} \ ; \ \ w_{D2} = |t|_{v-1,(vs-1)(c-1)-(v-1),\alpha}^{(0.5)} \ .$$

12.5.3   How Many Observations?

The methods of sample-size calculation for an *s*-replicate Youden square are analogous to those for computing sample sizes for a balanced incomplete block design. To calculate the number of observations $r = cs$ per treatment that are required to achieve a power $\pi(\Delta)$ of detecting a difference Δ in the treatment effects for given v, α, and σ^2, the power tables in Appendix A.7 can be used in the same way as for a balanced incomplete block design (Sect. 11.6). Thus, we need to find s satisfying

$$cs \geq \frac{2v\sigma^2\phi^2}{\Delta^2}\left[\frac{c(v-1)}{v(c-1)}\right] \,,$$

which reduces to

$$s \geq \frac{2\sigma^2\phi^2(v-1)}{\Delta^2(c-1)} \,.$$

Alternatively, the confidence interval formula (12.5.16) can be used to calculate the sample sizes needed for achieving confidence intervals of a desired width (see Example 12.5.1).

Example 12.5.1

Suppose an experiment is run to compare six paint additive formulations (levels 2–7 of the treatment factor) with a standard "control" formulation (level 1) with respect to the drying time (in minutes). The

paint is sprayed through $c = 4$ different nozzles so that $c = 4$ paints can be sprayed simultaneously. A total of $b = 7s$ panels will each be painted with strips of 4 of the 7 paints (using a standard formulation between each test panel to clean the nozzles and create a washout period). The error variability is expected to be at most $25\,\text{min}^2$ (standard deviation at most 5 min), and simultaneous 90% confidence intervals with half-width of at most 3.5 min are required for the six treatment-versus-control contrasts.

A Youden square with 7 rows, 4 columns, and 7 treatments is shown in Table 12.1. Suppose that s copies (or column randomizations) of this basic square are to be stacked, giving an s-replicate Youden design with the same number of observations on the experimental and control treatments. The number of error degrees of freedom, given in (12.5.14), is then

$$df = (vs - 1)(c - 1) - (v - 1) = (7s - 1)(4 - 1) - (7 - 1) = 21s - 9 \,.$$

Using Dunnett's method of multiple comparisons for treatment versus control, the minimum significant difference is

$$msd = w_{D2}\sqrt{msE \times \left(\frac{(v - 1)}{vs(c - 1)}\right) \times 2} \,,$$

so we require

$$msd \approx w_{D2}\sqrt{\frac{(25)(6)(2)}{(7)s(3)}} \le 3.5 \,,$$

that is,

$$w_{D2}^2 \le 0.8575s \,,$$

where $w_{D2} = |t|_{6,21s-9,.1}^{(.5)}$. Using Table A.10 for the values of $w_{D2} = |t|_{6,21s-9,.1}^{(.5)}$, we have

s	$21s - 9$	$w_{D2} = |t|_{6,21s-9,.1}^{(.5)}$	w_{D2}^2	$0.8575s$	Action
20	411	2.30	5.29	17.15	Decrease s
6	117	2.32	5.38	5.15	Increase s
7	138	2.32	5.38	6.00	Decrease s

So we see that $s = 7$ stacked Youden squares would be adequate, requiring $b = vs = 49$ panels and giving $r = sc = 28$ observations on each of the $v = 7$ paint formulations, both experimental and control. If 49 panels are not available for the experiment, then the experimenter would have to be satisfied with wider confidence intervals. □

12.6   Checking the Assumptions on the Model

The error assumptions on the row–column–treatment model (12.3.1) can be checked by plotting the standardized residuals against the run order, the predicted values $\hat{y}_{ijk}$, the levels of the row blocking factor, the levels of the column blocking factor, the levels of the treatment factor, and the normal scores.

The data collected from a row–column design can be examined by plotting the adjusted observations against the treatment labels. The observations are adjusted by subtracting the least squares estimates for row blocks and column blocks (which are adjusted for treatments):

$$y_{hqi}^* = y_{hqi} - (\hat{\theta}_h - \hat{\bar{\theta}}_.) - (\hat{\phi}_q - \hat{\bar{\phi}}_.) \,. \tag{12.6.17}$$

Fig. 12.1 Adjusted data
for the dairy cow
experiment
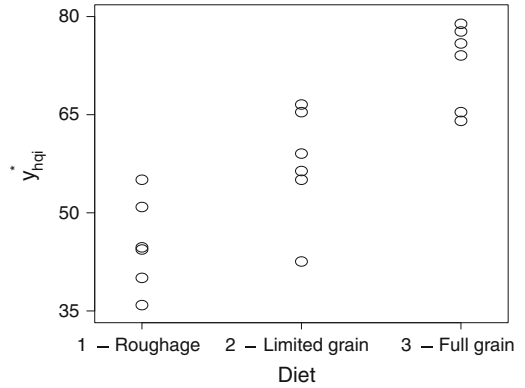
For the s-replicate Latin square design with $b = vs$ row blocks and $c = v$ column blocks, and the row–column–treatment model (12.3.1), the row-block and column-block effect estimators are independent of treatment effects, and (12.6.17) becomes

$$y_{hqi}^* = y_{hqi} - \left(\frac{1}{vs} B_h - \frac{1}{v^2 s} G \right) - \left(\frac{1}{v} C_q - \frac{1}{v^2 s} G \right) .$$

For other designs, the adjusted observations can be calculated by computer, as shown in Sects. 12.8 and 12.9. Since the variability due to the blocking factors has been extracted from the adjusted observations, the data plots will exhibit less variability than really exists.

Example 12.6.1 `Dairy cow experiment, continued`

For the dairy cow experiment, which was run as a Latin square design, the adjusted observations are plotted against treatment labels in Fig. 12.1. The plot shows how milk yield tended to increase as the quality of the diet improves from roughage (level 1) to limited grain (level 2) to full grain (level 3). The plot is consistent with the results of Example 12.4.1, where simultaneous confidence intervals (with an overall 95% confidence level) showed a significant difference in diets 1 and 3, but were unable to distinguish between diets 1 and 2 and between diets 2 and 3.   □

The row–column–treatment model (12.3.1) assumes that the two blocking factors do not interact with each other nor with the treatment factor. If interactions are present, the error variance estimate will be inflated, decreasing the powers of hypothesis tests and widening confidence intervals for treatment contrasts. When the column blocks are complete blocks, the column–treatment interaction can be checked by plotting the row-adjusted observations (as in Chap. 11) against treatments, with plot labels being the column-block labels. Interaction is indicated by nonparallel lines. The row–column interaction can be investigated in the same way using the treatment-adjusted observations and plotting against row labels. The row–treatment interactions can only be investigated if the row blocks are also complete blocks as in Latin square designs.

12.7   Factorial Experiments in Row–Column Designs

The exercise bike experiment of Example 12.2.1, p. 404, was a factorial experiment with three treatment factors having two levels each. These were "time duration of exercise" (1 and 3 min, coded as 1 and 2), "exercise speed" (40 and 60 rpm, coded as 1 and 2), and "pedal type" (foot pedal and hand bars, coded as 1 and 2). The data were shown in Table 12.6.

If we use the row–column–treatment model with three-digit labels for the treatment combinations, then we can write the model as

$$Y_{hqijk} = \mu + \theta_h + \phi_q + \tau_{ijk} + \epsilon_{hqijk} \,,$$

where θ_h is effect of day h, ϕ_q is the effect of subject q, and τ_{ijk} is the effect of treatment combination ijk. If we now rewrite τ_{ijk} in terms of its constituent main effects and interactions, we obtain the following form of the row–column–treatment model:

$$\begin{aligned}
Y_{hqijk} \quad &= \mu + \theta_h + \phi_q + \alpha_i + \beta_j + \gamma_k \\
&\quad + (\alpha\beta)_{ij} + (\alpha\gamma)_{jk} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{hqijk} \,, \\
h = 1, &\ldots, 8 \,; \; q = 1, 2, 3 \,; \; i = 1, 2 \,; \; j = 1, 2 \,; \; k = 1, 2 \,;
\end{aligned}$$

where α_i is the effect of the ith duration, β_j is the effect of the jth speed, and γ_k is the effect of the kth pedal type, and the other terms represent interactions between the treatment factors. The analysis of this experiment by the programs SAS and R software is shown in Sects. 12.8 and 12.9, respectively, for both forms of the model. If some of the treatment interactions are thought to be negligible, they can be dropped from the latter model.

12.8   Using SAS Software

In this section, a sample program is given to illustrate the analysis of row–column designs using the SAS software. The program uses the data of the exercise bicycle experiment, which was described in Example 12.2.1, p. 404. The design is a cyclic row–column design with $v = 8$ treatment labels representing the eight treatment combinations shown in Table 12.6, and with $b = 8$ row blocks representing days and $c = 3$ column blocks representing subject. The treatment combinations were combinations of the levels of the three treatment factors "time duration of exercise," "exercise speed," and "pedal type."

A SAS program for analyzing this experiment is shown in Table 12.10. The data are entered into a data set called BIKE, using DAY and SUBJECT as the two blocking factors, and using DURAT, SPEED, and PEDAL as the three treatment factors, and PULSE for the response variable "pulse rate." The combinations of the levels of the three treatment factors are recoded as levels of a factor TRTMT.

The MODEL statement in the first GLM procedure causes generation of the analysis of variance table shown in Fig. 12.2. The blocking factors are entered into the model before the treatment factor, so the treatment sum of squares is adjusted for block effects whether one looks at the Type I or Type III sums of squares. The row and column effects are independent of each other, since there is one observation at each combination of levels of the row and column blocking factors. Consequently, the Type I sums of squares reproduce the analysis of variance table, Table 12.7, with unadjusted block effects and adjusted treatment effects. In this particular experiment, the column blocks (subjects) are complete blocks, and

Table 12.10   SAS program for analysis of a row–column design—Exercise bicycle experiment

```
DATA BIKE;
  INPUT  DAY SUBJECT  PULSE  DURAT$ SPEED$ PEDAL$;
  TRTMT = trim(DURAT)||trim(SPEED)||trim(PEDAL);
  LINES;
    1 1  45  1 1 2
    1 2  25  2 1 2
    1 3  18  2 2 1
    2 1  27  2 2 2
    : :   :  : : :
    8 3  34  1 1 1
;
PROC PRINT;
*    row-column-treatment model;
PROC GLM;
  CLASS DAY SUBJECT TRTMT;
  MODEL PULSE = DAY SUBJECT TRTMT / SOLUTION;
  OUTPUT OUT=RESIDS PREDICTED=PRED RESIDUAL=Z;
  ESTIMATE 'DURATION DIFF'  TRTMT -1 -1 -1 -1  1  1  1  1 / DIVISOR=4;
  ESTIMATE 'SPEED DIFF'     TRTMT -1 -1  1  1 -1 -1  1  1 / DIVISOR=4;
  ESTIMATE 'PEDAL DIFF'     TRTMT -1  1 -1  1 -1  1 -1  1 / DIVISOR=4;
* Standardize residuals and compute normal scores;
PROC STANDARD STD=1.0; VAR Z;
PROC RANK NORMAL=BLOM; VAR Z; RANKS NSCORE;
* Residual plots can now be generated using PROC SGPLOT;
```

so the treatment sum of squares is actually only adjusted for row-block (day) effects. The Type III sums of squares show the row-block (day) effects adjusted for the treatment effects.

The ESTIMATE statements request the SAS software to calculate the information needed to calculate three confidence intervals. The three selected contrasts are the main-effect contrasts comparing the effect on pulse rate of the two levels of each treatment factor (averaging over any interaction that might be present). The output, shown in Fig. 12.3, gives the least squares estimates of the contrasts and their associated standard errors. From this information, confidence intervals can be calculated in the usual way. The contrast estimates are adjusted for the incomplete row blocks.

12.8.1   Factorial Model

To write the row–column–treatment model in terms of factorial treatment combinations, the model statement in Table 12.10 is replaced by

```
MODEL PULSE = DAY SUBJECT DURAT SPEED PEDAL DURAT*SPEED
              DURAT*PEDAL SPEED*PEDAL DURAT*SPEED*PEDAL;
```

The Type III sums of squares are shown in Fig. 12.4, and we can see that the sums of squares for the main effects and interactions of the three treatment factors do not add to the Type III treatment sum of squares in Fig. 12.2 due to the individual adjustments for block effects and for the other treatment factor effects.

Fig. 12.2 Analysis of variance for a row–column design—Exercise bicycle experiment

Fig. 12.3 Output from the ESTIMATE statement—Exercise bicycle experiment
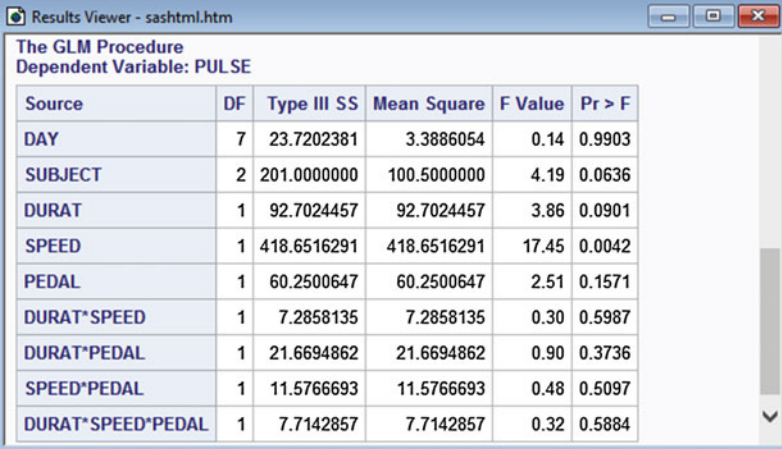
Parameter	Estimate	Standard Error	t Value	Pr > |t|
DURATION DIFF	-5.3437500	2.71854736	-1.97	0.0901
SPEED DIFF	-10.3214286	2.47086937	-4.18	0.0042
PEDAL DIFF	4.3080357	2.71854736	1.58	0.1571

Fig. 12.4 Analysis of variance for a row–column design—Exercise bicycle experiment

The `ESTIMATE` statements for the factorial model become

```
ESTIMATE 'DURATION DIFF'  DURAT -1 1;
ESTIMATE 'SPEED DIFF'     SPEED -1 1;
ESTIMATE 'FOOT/HAND DIFF' PEDAL -1 1;
```

and give output identical to that of Fig. 12.3.

12.8.2 Plots

The statements needed for calculating the standardized residuals and normal scores are also shown in Table 12.10. Residual plots (not shown here) can then be generated as in Sects. 5.8.1 and 6.8.3.

The `SOLUTION` option in the `MODEL` statement causes a set of least squares solutions to the normal equations to be printed. These are subsequently used to calculate the adjusted data values needed for examining the data.

For obtaining the adjusted observations, the statements in Table 12.11 would be added to the program in Table 12.10 for a second and third run of the program. The data are adjusted using (12.6.17); that is,

$$y_{hqi}^* = y_{hqi} - (\hat{\theta}_h - \hat{\bar{\theta}}.) - (\hat{\phi}_q - \hat{\bar{\phi}}.)\,,$$

where $\hat{\theta}_h$ and $\hat{\phi}_q$ are obtained from the output of the `SOLUTION` option shown in Fig. 12.2. The second run of the program takes as input the values of $\hat{\theta}_h$ and $\hat{\phi}_q$ from the first run of the program and calculates $\hat{\bar{\theta}}. = \Sigma\hat{\theta}_h/3 = 1.25$ and $\hat{\bar{\phi}}. = \Sigma\hat{\phi}_q/8 = 1.66$, which are then copied by hand into the statements for the third run of the program. The symbols `@@` allow the input to be entered with more than one record per line.

In the third run of the program, the data set `BIKE5` is created as a copy of the data set `BIKE`. This is done to create the new variable `YADJ`, which contains the values of `PULSE`, adjusted for the row and column effects. The `SGPLOT` procedure is used to plot the adjusted observations by treatment. The SAS plot is analogous to that in Fig. 12.1 (p. 415) and is not shown here.

Table 12.11 SAS code for calculating and plotting the adjusted observations—Exercise bicycle experiment

```
* Add the following code for the second run of the program;
DATA BIKE3; * input subject effect estimates from first run;
  INPUT SUBJECT SBJHAT; @@;
  LINES;
   1  5.25     2 -1.50     3  0.00
PROC MEANS MEAN; * print average of the subject effect estimates;
  VAR SBJHAT;
DATA BIKE4; * input day effect estimates from first run;
  INPUT DAY DHAT @@;
  LINES;
   1 3.9911     2 1.5625     3 2.5714     4 0.8036
   5 2.5179     6 1.6334     7 0.2054     8 0.0000
PROC MEANS MEAN; * print average of the day effect estimates;
  VAR DHAT;
* Add the following code for the third run;
* Adjust data for subject and day effects, then plot adjusted data;
DATA BIKE5; SET BIKE;
  IF SUBJECT = 1 THEN YADJ = PULSE-(5.25-1.25);
    ELSE IF SUBJECT = 2 THEN YADJ = PULSE-(-1.50-1.25);
      ELSE IF SUBJECT = 3 THEN YADJ = PULSE-(0.00-1.25);
  IF DAY = 1 THEN YADJ = YADJ-(3.9911-1.660);
    ELSE IF DAY = 2 THEN YADJ = YADJ-(1.5625-1.6607);
      ELSE IF DAY = 3 THEN YADJ = YADJ-(2.5714-1.6607);
        ELSE IF DAY = 4 THEN YADJ = YADJ-(0.8036-1.6607);
          ELSE IF DAY = 5 THEN YADJ = YADJ-(2.5179-1.6607);
            ELSE IF DAY = 6 THEN YADJ = YADJ-(1.6334-1.6607);
              ELSE IF DAY = 7 THEN YADJ = YADJ-(0.2054-1.6607);
                ELSE IF DAY = 8 THEN YADJ = YADJ-(0.0000-1.6607);
PROC SGPLOT;
  SCATTER X = TRTMT Y = YADJ;
  X AXIS LABEL = 'Treatment'; REFLINE 0/ AXIS = X;
  Y AXIS LABEL = 'Adjusted Observations'; REFLINE 0/ AXIS = Y;
```

12.9   Using R Software

In this section, a sample program is given to illustrate the analysis of row–column designs using the R software. The program uses the data of the exercise bicycle experiment, which was described in Example 12.2.1, p. 404. The design is a cyclic row–column design with $v = 8$ treatment labels representing the eight treatment combinations shown in Table 12.6, and with $b = 8$ row blocks representing days and $c = 3$ column blocks representing subject. The treatment combinations were combinations of the levels of the three treatment factors "time duration of exercise," "exercise speed," and "pedal type."

An R program for analyzing this experiment is shown in Table 12.12. The data are entered into a data set called `bike.data`, using `fDay` and `fSubject` as the two blocking factors, and using `fDurat`, `fSpeed`, and `fPedal` as the three treatment factors, and `Pulse` for the response variable "pulse rate." The combinations of the levels of the three treatment factors are recoded as levels of a factor `fTrtmt`.

In the second block of code, the linear models function `lm` is used to fit the row–column–treatment model (12.3.1), then corresponding `anova` and `drop1` functions generate the type 1 and type 3

Table 12.12 R program for analysis of a row–column design—Exercise bicycle experiment

```
bike.data = read.table("data/exercise.bicycle.txt", header=T)
head(bike.data, 3)

      Day Subject Durat Speed Pedal Pulse Trtmt
   1   1       1     1     1     2    45   112
   2   2       1     2     2     2    27   222
   3   3       1     2     1     2    40   212

# Create factor variables
bike.data = within(bike.data,
  {fDay=factor(Day); fSubject=factor(Subject); fDurat=factor(Durat);
   fSpeed=factor(Speed); fPedal=factor(Pedal); fTrtmt=factor(Trtmt) })

# ANOVA: treatment combinations
modelTC = lm(Pulse ~ fDay + fSubject + fTrtmt, data=bike.data)
anova(modelTC)
drop1(modelTC, ~., test = "F")
# Treatment contrasts
library(lsmeans)
lsmTrtmt = lsmeans(modelTC, ~ fTrtmt)
summary(contrast(lsmTrtmt,
        list(DurationDiff=c(-1,-1,-1,-1, 1, 1, 1, 1)/4,
                SpeedDiff=c(-1,-1, 1, 1,-1,-1, 1, 1)/4,
                PedalDiff=c(-1, 1,-1, 1,-1, 1,-1, 1)/4)),
        infer=c(T,T))

# Compute variables for residual plots
bike.data = within(bike.data, {
   # Compute predicted, residual, and standardized residual values
   ypred = fitted(modelTC); e = resid(modelTC); z = e/sd(e);
   # Compute Blom's normal scores
   n = length(e); q = rank(e); nscore = qnorm((q-0.375)/(n+0.25))
   # Compute vbl TLevel with levels 1:8 for equispaced plotting vs TC's
   TLevel = 4*(Durat-1) + 2*(Speed-1) + (Pedal-1) })
# Residual plots
plot(z ~ TLevel, data=bike.data, xaxt="n", xlab="Treatment")
   axis(1, at=bike.data$TLevel, labels=bike.data$fTrtmt)
plot(z ~ ypred + Day + Subject + nscore, data=bike.data)
```

analysis of variance tables, respectively, shown at the top of Table 12.13. The blocking factors are entered into the model before the treatment factor, so the treatment sum of squares is adjusted for block effects whether one looks at the Type I or Type III sums of squares. The row and column effects are independent of each other, since there is one observation at each combination of levels of the row and column blocking factors. Consequently, the Type I sums of squares reproduce the analysis of variance table, Table 12.7, with unadjusted block effects and adjusted treatment effects. In this particular experiment, the column blocks (subjects) are complete blocks, and so the treatment sum of squares is actually only adjusted for row-block (day) effects. The Type I sums of squares show the row-block (day) effects adjusted for the treatment effects.

The `summary` and `contrast` functions of the `lsmeans` package are coupled to calculate estimates, standard errors, tests and confidence for the three main-effect contrasts comparing the effect

Table 12.13 Analysis of variance and contrast estimation for a row–column design—Exercise bicycle experiment

```
> # ANOVA: treatment combinations
> modelTC = lm(Pulse ~ fDay + fSubject + fTrtmt, data=bike.data)
> anova(modelTC)

Analysis of Variance Table

Response: Pulse
          Df Sum Sq Mean Sq F value Pr(>F)
fDay       7    202    28.9    1.20  0.406
fSubject   2    201   100.5    4.19  0.064
fTrtmt     7    854   122.1    5.09  0.024
Residuals  7    168    24.0

> drop1(modelTC, ~., test = "F")

Single term deletions

Model:
Pulse ~ fDay + fSubject + fTrtmt
         Df Sum of Sq  RSS    AIC F value Pr(>F)
<none>                 168   80.7
fDay      7        24  192   69.9    0.14  0.990
fSubject  2       201  369   95.6    4.19  0.064
fTrtmt    7       854 1022  110.0    5.09  0.024

> # Treatment contrasts
> library(lsmeans)
> lsmTrtmt = lsmeans(modelTC, ~ fTrtmt)
> summary(contrast(lsmTrtmt,
+                   list(DurationDiff=c(-1,-1,-1,-1, 1, 1, 1, 1)/4,
+                        SpeedDiff=c(-1,-1, 1, 1,-1,-1, 1, 1)/4,
+                        PedalDiff=c(-1, 1,-1, 1,-1, 1,-1, 1)/4)),
+         infer=c(T,T))

 contrast      estimate     SE df lower.CL upper.CL t.ratio p.value
 DurationDiff   -5.3438 2.7185  7 -11.7721   1.0846  -1.966  0.0901
 SpeedDiff     -10.3214 2.4709  7 -16.1641  -4.4788  -4.177  0.0042
 PedalDiff       4.3080 2.7185  7  -2.1203  10.7364   1.585  0.1571

Results are averaged over the levels of: fDay, fSubject
Confidence level used: 0.95
```

on pulse rate of the two levels of each treatment factor (averaging over any interaction that might be present). The corresponding output is shown in the bottom of Table 12.13.

Predicted values and residuals are available, and the residuals can be standardized and plotted as in Sects. 5.9.1 and 6.9.3. Sample code is provided at the bottom of Table 12.12.

Table 12.14 Analysis of variance for a row–column design—Exercise bicycle experiment

```
> # ANOVA: factorial effects
> modelFE = lm(Pulse ~ fDay + fSubject + fDurat*fSpeed*fPedal,
+               data=bike.data)
> anova(modelFE)
> drop1(modelFE, ~., test = "F")

Single term deletions

Model:
Pulse ~ fDay + fSubject + fDurat * fSpeed * fPedal
                    Df Sum of Sq RSS   AIC F value Pr(>F)
<none>                           168  80.7
fDay                 7      23.7 192  69.9    0.14  0.990
fSubject             2     201.0 369  95.6    4.19  0.064
fDurat               1     129.2 297  92.4    5.38  0.053
fSpeed               1     269.1 437 101.6   11.22  0.012
fPedal               1       1.1 169  78.9    0.05  0.833
fDurat:fSpeed        1      15.0 183  80.7    0.63  0.455
fDurat:fPedal        1      36.5 204  83.4    1.52  0.257
fSpeed:fPedal        1      24.3 192  81.9    1.01  0.348
fDurat:fSpeed:fPedal 1       7.7 176  79.8    0.32  0.588
```

12.9.1 Factorial Model

To write the row–column–treatment model in terms of factorial effects, the linear models statement in Table 12.12 is replaced by

```
modelFE = lm(Pulse ~ fDay + fSubject + fDurat*fSpeed*fPedal,
             data=bike.data)
```

The Type III sums of squares are shown in Table 12.14, and we can see that the sums of squares for the main effects and interactions of the three treatment factors do not add to the Type III treatment sum of squares in Table 12.13 due to the individual adjustments for block effects and for the other treatment factor effects.

The statements to estimate contrasts for the factorial model become

```
lsmDurat = lsmeans(modelFE, ~ fDurat)
summary(contrast(lsmDurat,  list(DurationDiff=c(-1, 1))), infer=c(T,T))
lsmSpeed = lsmeans(modelFE, ~ fSpeed)
summary(contrast(lsmSpeed, list(SpeedDiff=c(-1, 1))), infer=c(T,T))
lsmPedal = lsmeans(modelFE, ~ fPedal)
summary(contrast(lsmPedal, list(PedalDiff=c(-1, 1))), infer=c(T,T))
```

and give output identical to that of Table 12.13.

Table 12.15 R code for calculating and plotting the adjusted observations—Exercise bicycle experiment

```
bike.data = read.table("data/exercise.bicycle.txt", header=T)
# Create factor variables, plus numeric variable TLevel for plotting
bike.data = within(bike.data,
     {fDay=factor(Day); fSubject=factor(Subject); fDurat=factor(Durat);
      fSpeed=factor(Speed); fPedal=factor(Pedal); fTrtmt=factor(Trtmt);
      # Vbl TLevel with treatment levels 1:8 for equispaced plotting;
      TLevel = 4*(Durat-1) + 2*(Speed-1) + (Pedal-1) })

# ANOVA: treatment combinations
modelTC = lm(Pulse ~ fDay + fSubject + fTrtmt, data=bike.data)

# Plotting data adjusted for row and column effects
modelTC$coefficients # Display all model coefficient estimates

    (Intercept)         fDay2         fDay3         fDay4         fDay5         fDay6
        41.7500       -2.4286       -1.4196       -3.1875       -1.4732       -2.3571

          fDay7         fDay8     fSubject2     fSubject3     fTrtmt112     fTrtmt121
        -3.7857       -3.9911       -6.7500       -5.2500       -1.2143      -14.7054

      fTrtmt122     fTrtmt211     fTrtmt212     fTrtmt221     fTrtmt222
        -9.5714      -10.1875       -4.1339      -19.9018      -12.6429

thetahat = c(0, modelTC$coefficients[2:8]) # Row effect estimates
thetamean = mean(thetahat) # Mean row effect estimate
phihat = c(0, modelTC$coefficients[9:10]) # Column effect estimates
phimean = mean(phihat) # Mean column effect estimate
# Compute adjusted y-values
bike.data$yadj = ( bike.data$Pulse
                     - (thetahat[bike.data$Day] - thetamean)
                     - (phihat[bike.data$Subject] - phimean) )
# Plot adjusted response versus treatment
plot(yadj ~ TLevel, data=bike.data, xaxt="n", xlab="Treatment",
     ylab="y Adjusted")
axis(1, at=bike.data$TLevel, labels=bike.data$fTrtmt)
```

12.9.2   Plots

The statements needed for calculating the standardized residuals and normal scores and for generating the residual plots are also shown in Table 12.12. These are as discussed in Sects. 5.9.1 and 6.9.3. The `plot` statement is used to generate the usual residual plots (not shown here).

The R program in Table 12.15 illustrates how to compute the adjusted observations—namely, the observations adjusted for row and column effect estimates—and how to plot the adjusted observations versus treatment. The data are adjusted using (12.6.17); that is,

$$y_{hqi}^* = y_{hqi} - (\hat{\theta}_h - \bar{\hat{\theta}}_.) - (\hat{\phi}_q - \bar{\hat{\phi}}_.) \,.$$

Toward this end, having saved the fitted row-column design model as `modelTC`, the statement `modelTC$coefficients` displays a set of least squares solutions to the normal equations. These are shown for example in Table 12.15. The row and column effect estimates $\hat{\theta}_h$ and $\hat{\phi}_q$ are obtained from the column `modelTC$coefficient` of displayed least squares estimates by specifying the appropriate entries, and the mean of each collection of estimates is computed. Given this information, the adjusted observations are computed.

The `plot` function is used to plot the adjusted observations versus treatment. The adjusted observations are actually plotted against treatment levels 1–8, so the treatment combinations are equally spaced on the x-axis of the plot, but these equispaced levels are replaced by the treatment combination labels 111–222. The R plot is analogous to that in Fig. 12.1 (p. 415) and is not shown here.

Exercises

1. **Randomization**

   (a) Randomize the plan in Table 12.1 (p. 400) so that it can be used for an experiment with seven subjects, each being assigned a sequence of four out of a possible seven antihistamines over four time periods.
   (b) Discuss whether or not one would need to include a carryover effect in the model, or whether this could be avoided through the design of the experiment.

2. **Latin squares**

   (a) Show that there is only one standard 3×3 Latin square. (Hint: Given the letters in the first row and the first column, show that there is only one way to complete the Latin square.)
   (b) Show that there are exactly four standard 4×4 Latin squares.

3. **Sample sizes**

   Consider an experiment to compare 4 degrees of twist in a cotton-spinning experiment with respect to the number of breaks per 100 pounds. A replicated Latin square design is to be used, with time periods and machines being the row and column blocking factors.

   (a) Determine the number s of Latin squares and the number r of observations per degree of twist to include in the experiment if each interval in a simultaneous set of 95% confidence intervals for all pairwise comparisons is to have a minimum significant difference (half-width) of 5 breaks per 100 pounds. The error standard deviation is thought to be at most 6 breaks per 100 pounds. Investigate both the Tukey and the Bonferroni methods.
   (b) Discuss how the resulting design would be randomized.

4. **Youden design randomization**

   (a) Find a Youden square (plan of treatment labels in rows and columns) for 5 treatments in 5 rows and 4 columns.
   (b) Randomize the design found in part (a), assigning the rows to 5 different drying temperatures, the columns to 4 different paint nozzles, and the treatment labels to 5 different paint formulations.

Table 12.16 Latin square design showing treatments and data for the video game experiment

						Day						
		1		2		3		4		5		
Time order	1	1	94	3	100	4	98	2	101	5	112	
	2	3	103	2	111	1	51	5	110	4	90	
	3	4	114	1	75	5	94	3	85	2	107	
	4	5	100	4	74	2	70	1	93	3	106	
	5	2	106	5	95	3	81	4	90	1	73	

5. **Row–column design randomization**

Consider an experiment to compare 5 protocols with respect to a resting metabolism rate measurement. A row–column design is to be used, blocking on subjects and time periods. Since subjects prefer to stay in a study for a short length of time, only 3 time periods will be used, with each subject assigned a different protocol in each of the 3 time periods. For 10 subjects, the following experimental plan with 10 rows and 3 columns could be used:

	Column				Column		
Row	I	II	III	Row	I	II	III
I	1	2	3	VI	1	2	4
II	2	3	4	VII	2	3	5
III	3	4	5	VIII	3	4	1
IV	4	5	1	IX	4	5	2
V	5	1	2	X	5	1	3

(a) Does this experimental plan have treatments evenly distributed across rows and columns? Explain what you mean by "evenly distributed."
(b) Determine the number of replicates s of this experimental plan, and the corresponding number of observations r per protocol to include in the experiment if each interval in a set of simultaneous 95% confidence intervals for all pairwise comparisons is to have a minimum significant difference (half-width) of 150 units. The error standard deviation is thought to be at most 250 units. Investigate both the Tukey and the Bonferroni methods.
(c) Discuss how the resulting design would be randomized.

6. **Video game experiment**

Professor Robert Wardrop, of the University of Wisconsin, conducted an experiment in 1991 to evaluate in which of five sound modes he best played a certain video game. The first three sound modes corresponded to three different types of background music, as well as game sounds expected to enhance play. The fourth mode had game sounds but no background music. The fifth mode had no music or game sounds. Denote these sound modes by the treatment factor levels 1–5, respectively.

The experimenter observed that the game required no warm up, that boredom and fatigue would be a factor after 4–6 games, and that his performance varied considerably on a day-to-day basis. Hence, he used a Latin square design, with the two blocking factors being "day" and "time order of the game." The response measured was the game score, with higher scores being better. The design and resulting data are given in Table 12.16.

(a) Write down a possible model for these data and check the model assumptions. If the assumptions appear to be approximately satisfied, then answer parts (b)–(f).
(b) Plot the adjusted data and discuss the plot.
(c) Complete an analysis of variance table.
(d) Evaluate whether blocking was effective.
(e) Construct simultaneous 95% confidence intervals for all pairwise comparisons, as well as the "music versus no music" contrast

$$\frac{1}{3}(\tau_1 + \tau_2 + \tau_3) - \frac{1}{2}(\tau_4 + \tau_5)$$

and the "game sound versus no game sound" contrast

$$\frac{1}{4}(\tau_1 + \tau_2 + \tau_3 + \tau_4) - \tau_5 .$$

(f) What are your conclusions from this experiment? Which sound mode(s) should Professor Wardrop use?

7. **Video game experiment, continued**

Suppose that in the video game experiment of Exercise 6, Professor Wardop had run out of time and that only the first four days of data had been collected. The design would then have been a Youden design. Repeat parts (c), (e), and (f) of Exercise 6. Do your conclusions remain the same? Is this what you expected? Why or why not?

8. **Air freshener experiment**

A. Cunningham and N. O' Connor (1968, *European Journal of Marketing* 2, 147–151) conducted a two-replicate Latin square design to compare the effects of four price-and-display treatments on the sales of a brand of air fresheners. Treatments 1–3 corresponded to high, middle, and low prices, respectively, and each had an extra display. Treatment 4 corresponded to the middle price and no extra display. The response variable was the unit sales for a one-week period. The experiment involved two blocking factors defined by stores ($c = 8$ levels) and one-week periods ($b = 4$ levels). The design and data are given in Table 12.17.

(a) Factors such as product location and shelf stocking could affect sales. Discuss how these factors might be controlled in such an experiment.
(b) Check the model assumptions.
(c) Plot the adjusted data and comment on the results.
(d) Complete an analysis of variance table.

Table 12.17 Latin square design and data for the air freshener experiment

Week	Store															
	1		2		3		4		5		6		7		8	
1	2	31	1	23	3	12	4	3	1	10	3	30	2	23	4	14
2	1	19	4	16	2	14	3	4	2	21	4	25	3	17	1	14
3	4	15	3	30	1	12	2	6	3	12	1	47	4	5	2	3
4	3	16	2	27	4	5	1	11	4	12	2	38	1	13	3	6

Source Data adapted from Cunningham and O'Connor, European Journal of Marketing, Copyright © Emerald Group Publishing Limited

(e) Evaluate whether blocking was effective.

(f) Test for equality of treatment effects using a 5% significance level.

(g) Construct simultaneous 95% confidence intervals for all pairwise comparisons of the treatments. What would you recommend for the sales of air fresheners if the results of this experiment are still valid today?

9. **Air freshener experiment, continued**

Suppose that the air freshener experiment of Exercise 8 had to be stopped after only 3 weeks. The resulting design would then be a replicated Youden design. Repeat parts (d)–(g) of Exercise 8. Do your conclusions remain the same? Is this what you expected? Why or why not?

10. **Quantity perception experiment**

An experiment was run in 1996 by M. Gbenado, A. Veress, L. Heimenz, J. Monroe, and S. Yu to investigate the effect of color on the perception of quantity. Subjects were recruited at random from the student population. A number of small candies of a specific color were tipped onto a flat tray. A subject was allowed to view the tray for 3 sec and then asked to make a guess as to the number of candies on the tray. The response was the difference between the actual number of candies on the tray and the number guessed by the subject.

The treatment factors of interest were "actual number of candies on the tray" and "color." The selected levels were 17, 29, and 41 for the treatment factor "number", and yellow, orange, brown for the factor "color." The experimenters decided that each subject should view all nine treatment combinations, and they based their design on 9×9 Latin squares.

We consider only part of the original study, constituting a 2-replicate Latin square, as shown in

Table 12.18 2-replicate Latin square design and data in parentheses for the quantity perception experiment—data are 'true number' minus 'guessed number'

Subj	Time Order								
	1	2	3	4	5	6	7	8	9
1	23 (4)	22 (−3)	11 (0)	12 (−3)	32 (−1)	13 (−3)	31 (−6)	33 (−9)	21 (−1)
2	12 (2)	31 (16)	32 (21)	21 (9)	22 (4)	33 (16)	23 (9)	11 (2)	13 (2)
3	21 (4)	23 (−1)	12 (7)	13 (−5)	33 (−1)	11 (−13)	32 (−9)	31 (−19)	22 (−16)
4	32 (21)	12 (4)	22 (3)	31 (11)	13 (0)	21 (4)	11 (0)	23 (4)	33 (11)
5	31 (7)	11 (−2)	21 (2)	33 (3)	12 (−3)	23 (3)	13 (−4)	22 (−5)	32 (−7)
6	11 (3)	33 (7)	31 (14)	23 (11)	21 (12)	32 (17)	22 (10)	13 (5)	12 (0)
7	22 (11)	21 (14)	13 (0)	11 (1)	31 (16)	12 (1)	33 (13)	32 (14)	23 (7)
8	13 (7)	32 (16)	33 (16)	22 (4)	23 (4)	31 (16)	21 (14)	12 (7)	11 (2)
9	33 (21)	13 (2)	23 (10)	32 (24)	11 (6)	22 (13)	12 (2)	21 (8)	31 (20)
10	33 (16)	31 (20)	22 (6)	21 (6)	11 (7)	23 (6)	12 (2)	13 (3)	32 (14)
11	12 (2)	22 (4)	32 (11)	13 (2)	21 (9)	33 (1)	23 (4)	31 (−4)	11 (7)
12	13 (−4)	23 (−11)	33 (−4)	11 (−3)	22 (4)	31 (11)	21 (−1)	32 (1)	12 (−3)
13	21 (4)	12 (−1)	11 (2)	32 (11)	31 (11)	13 (−3)	33 (1)	22 (−1)	23 (11)
14	22 (2)	13 (−7)	12 (−9)	33 (8)	32 (−2)	11 (−6)	31 (−9)	23 (4)	21 (2)
15	31 (21)	32 (21)	23 (14)	22 (14)	12 (4)	21 (16)	13 (0)	11 (5)	33 (11)
16	11 (2)	21 (9)	31 (21)	12 (6)	23 (9)	32 (18)	22 (9)	33 (16)	13 (2)
17	32 (6)	33 (6)	21 (−1)	23 (−1)	13 (2)	22 (4)	11 (−3)	12 (−1)	31 (6)
18	23 (4)	11 (2)	13 (2)	31 (11)	33 (6)	12 (2)	32 (6)	21 (−1)	22 (4)

Table 12.18. Subjects represent the row blocks, and time order represents the column blocks. The treatment combinations have been coded as follows:

$$1 = (17, \text{yellow}) \quad 2 = (17, \text{orange}) \quad 3 = (17, \text{brown})$$
$$4 = (29, \text{yellow}) \quad 5 = (29, \text{orange}) \quad 6 = (29, \text{brown})$$
$$7 = (41, \text{yellow}) \quad 8 = (41, \text{orange}) \quad 9 = (41, \text{brown})$$

(a) The experiment was conducted in a busy hallway at The Ohio State University. Subjects were recruited from the population of noncolorblind students walking past the table. Recruited subjects were not allowed to view the experiment in progress with previous subjects, but they were paid for their participation in candies. Discuss whether or not the subjects in this study are likely to be representative of some larger population of subjects. Are the conclusions of the study likely to be relevant to noncolorblind people in general?

(b) Fit a model that includes the effects of the two blocking factors, "subject" and "time-order", the treatment effect, and the treatment×time-order interaction. By looking at a computer calculated analysis of variance table, verify that the interaction (adjusted for subjects) can be measured with the full set of $(v - 1)^2 = 64$ degrees of freedom.

(c) For the model that includes treatment×time-order interaction, check whether the residuals are approximately normally distributed and whether they have approximately the same variance for each treatment. Do you prefer to use the original response variable "guessed number" or the transformed response "square root of guessed number" or "(true number − guessed number)/(true number)" or some other transformation?

(d) Present an analysis of variance table and test any hypotheses that you think are of interest. State your conclusions.

(e) Rewrite your model in terms of main effects and interactions of the two treatment factors. Redo your analysis of variance table. What can you conclude from the experiment?

(f) If you were to plan a followup experiment, what would you wish to study? Write up a checklist for such an experiment.

11. **Caffeine experiment**

An experiment was run by Lisa Carpinello in 2001 to study the effects of caffeine on a subject's blood pressure readings. Two treatment conditions were studied; subject abstaining from caffeine for two hours in advance of the blood pressure readings (level 0), and subject consuming 12 ounces of coffee with one tablespoon of non-dairy creamer one hour before the readings (level 1).

The investigator measured the subject's blood pressure at 9am and 2pm each day for eight days, with the subject sitting quietly for five minutes before the readings were taken. The ordered pairs of treatments were randomly assigned to the days in such a way that on four of the days, the treatment order was level 0 at 9am, level 1 at 2pm, and, on the other four days, the treatment order was reversed (i.e. 1 at 9am, 0 at 2pm). The design can be represented as a randomized 4-replicate

Table 12.19 Latin square design and systolic blood pressure readings (mm Hg) for the caffeine experiment

	Time: 9am	Time: 2pm		Time: 9am	Time: 2pm
Day 1	1 121	0 118	Day 5	0 126	1 130
Day 2	1 123	0 119	Day 6	1 129	0 123
Day 3	0 111	1 117	Day 7	1 127	0 118
Day 4	0 120	1 129	Day 8	0 121	1 130

Latin square design with rows representing days and columns time of day. The design and data are shown in Table 12.19.

(a) Plot the data and comment on the results.
(b) Write down a model for this experiment and check the assumptions on your model.
(c) Complete an analysis of variance table and evaluate whether blocking was effective.
(d) Test for equality of treatment effects using a 5% significance level.
(e) Construct a 95% confidence interval for comparing the treatments, and interpret the results.

12. **Golf driver experiment**

An experiment was conducted by Dale Meyer in 2001 to compare four different golf clubs (drivers) in terms of the distance golf balls travel when hit. Two expensive drivers (coded 1 and 2) and two cheaper ones (coded 3 and 4) were used. Four golfers each hit four shots in each of four rounds, each golfer using each of the four drivers once per round, and balancing for the order in which each driver was used. A different Latin square was used as the design in each round. The final design is shown in Table 12.20 with driver as the treatment factor, and rows and columns defined by golfer and order. Notice that, not only do the combinations of golfer/order/driver constitute a Latin square in each round, but so do combinations of round/order/driver for each golfer, and so do combinations of round/golfer/driver for each order. So the design is extremely well balanced and this allows the golfer×driver interaction to measured without adjusting for order or round. The experimenter was interested in comparing the driver effects and the golfer×driver interaction. No other interaction effects were anticipated. If the golfer×driver interaction were to be significant, then the experimenter also wanted to compare the driver effects separately for each golfer. For each observation, a golf ball was hit in the golf simulator of a golf store, and the simulated distances of the shots (in yards) were recorded and are shown in Table 12.20.

Table 12.20   Driver-distance (yards) pairs for the golf driver experiment

Round	Golfer	Order							
		1		2		3		4	
1	1	1	175	2	90	3	176	4	102
1	2	2	157	1	149	4	166	3	140
1	3	3	196	4	182	1	199	2	158
1	4	4	147	3	216	2	197	1	134
2	1	2	147	1	191	4	137	3	165
2	2	1	124	2	135	3	171	4	79
2	3	4	180	3	195	2	206	1	184
2	4	3	167	4	161	1	143	2	121
3	1	3	187	4	166	1	176	2	181
3	2	4	135	3	123	2	168	1	131
3	3	1	197	2	165	3	200	4	171
3	4	2	216	1	158	4	181	3	180
4	1	4	176	3	192	2	193	1	154
4	2	3	176	4	125	1	95	2	187
4	3	2	150	1	188	4	201	3	192
4	4	1	162	2	170	3	214	4	123

(a) Suggest a model for this experiment. By looking at computer calculated interaction degrees of freedom and the adjusted and unadjusted sums of squares, verify that the golfer×driver interaction can be measured based on the full set of $(v - 1)^2 = 9$ degrees of freedom, and without adjusting for order or round. Test whether the golfer×driver interaction is significantly different from zero using $\alpha = 0.01$.

(b) If there is no significant golfer×driver interaction, compare the driver effects (averaged over golfer, order and round) using simultaneous 99% confidence intervals. Which method of multiple comparisons did you use and why? What can you conclude?

(c) Compare the pairwise effects of the drivers for each golfer, using individual 99% confidence intervals. Which method of multiple comparisons did you use and why? State your overall confidence level and interpret your results.

Confounded Two-Level Factorial
Experiments

13

13.1 Introduction

In Chaps. 6 and 7 we discussed factorial experiments arranged as completely randomized designs, and in Chaps. 10 and 11 we looked at factorial experiments arranged as block designs. Factorial experiments that involve several treatment factors tend to be large. Even a modest experiment with four factors having 2, 2, 3, and 3 levels has a total of 36 treatment combinations. Since experimenters generally are working to a restricted budget and since observations cost time and money, many factorial experiments are *single-replicate* experiments (one observation per treatment combination). In this chapter we consider single-replicate experiments arranged in blocks where every treatment factor has two levels. This will be extended in Chap. 14 to cover treatment factors with more than two levels. Then, in Chap. 15, we will look at experiments in which only a fraction of the treatment combinations can be observed.

In Sect. 13.2, we discuss alternative codings of treatment combinations, and the general problem of confounding together with its implications for analysis. Methods of designing single-replicate experiments so that information is lost on as few lower-order treatment contrasts as possible are the main focus of Sects. 13.3 and 13.4, followed by an example in Sect. 13.5. Section 13.6 describes plans for the design of confounded experiments.

In Sects. 13.7–13.10 we return to the subject of multi-replicate factorial experiments in blocks and compare the traditional incomplete block designs with the multiple use of single-replicate confounded designs. Analysis of confounded factorial experiments by the computer packages SAS and R is considered briefly in Sects. 13.11 and 13.12.

13.2 Single Replicate Factorial Experiments

13.2.1 Coding and Notation

A factorial experiment that involves two treatment factors each having two levels is known as a 2×2, or 2^2, experiment. Similarly, an experiment with two factors each having 3 levels is known as a 3×3, or 3^2, experiment. A $2^4 \times 3^2$ experiment has six treatment factors, the first four having two levels each, and the last two having three levels each. Other factorial experiments are described in a similar manner. A factorial experiment is called *symmetric* if all factors have the same number of levels. Otherwise, it is called *asymmetric*. In this chapter we will deal only with symmetric 2^p experiments. Other situations will be discussed in Chap. 14.

The levels of a two-level treatment factor are often referred to as the "low" and "high" levels, and in Chaps. 6 and 7 we coded these as 1 and 2. The codings 0 and 1, or -1 and $+1$, are also commonly used. A 2^2 experiment then has four treatment combinations coded as (11, 12, 21, 22) or as (00, 01, 10, 11) or as $(-1-1, -1+1, +1-1, +1+1)$. A fourth standard coding for the treatment combinations is $((1), b, a, ab)$, where the letter a or b appears if the corresponding factor A or B is at its high level, and is absent if the corresponding factor is at its low level. The symbol (1) means that both factors are at their low level.

Coding is a matter of personal choice. Although we have coded the levels as 1 and 2 until now, the other three codings are more usual when talking about single-replicate factorial experiments. We will code the levels as 0 and 1 throughout this and the next two chapters.

13.2.2   Confounding

A factorial experiment with v treatment combinations uses $v - 1$ degrees of freedom to measure all of the main effects and interactions. In a single-replicate experiment, there are only v observations and $v - 1$ total degrees of freedom. Thus, the experiment is not large enough to allow measurement of all of the factorial effects and also estimation of the error variance. Three ways around this problem for statistical inference in completely randomized designs were discussed in Sects. 6.7 and 7.5.

The problem is worse when the experiment is to be run as a block design. If there are b blocks in the design, $b - 1$ of the total degrees of freedom are used to measure the block differences, leaving only $(v - 1) - (b - 1) = v - b$ degrees of freedom available for measuring the treatment contrasts and the error variance. The result of this is that $b - 1$ of the treatment contrasts can no longer be measured. They cannot be distinguished from block contrasts and are said to be *confounded* with blocks. Such a design is useful only when at most $v - b$ treatment contrasts are to be measured.

Care is required in designing this type of experiment. If v treatment combinations are arbitrarily divided into b blocks of size v/b, the important treatment contrasts will not necessarily be estimable. The estimable contrasts are those that are orthogonal to the confounded contrasts. This means that the experiment should be designed in such a way that the confounded contrasts belong only to interactions that are expected to be negligible. Fortunately, in some cases this is not difficult to achieve, and we will examine these cases in this chapter. For a 2^p experiment, we will restrict attention to designs with $b = 2^s$ blocks of size $k = 2^{p-s}$.

13.2.3   Analysis

The standard block–treatment model for a single-replicate factorial experiment arranged as an incomplete block design has the same form as model (11.4.2), p. 356, used for incomplete block designs in Chap. 11; that is,

$$Y_{hi} = \mu + \theta_h + \tau_i + \epsilon_{hi} \,,$$
$$\epsilon_{hi} \sim N(0, \sigma^2) \,,$$
$$\epsilon_{hi}\text{'s are mutually independent} \,,$$
$$h = 1, \ldots, b; \quad i = 1, \ldots, v; \quad (h, i) \text{ is in the design.} \tag{13.2.1}$$

As usual, Y_{hi} is the random variable representing the observation on treatment combination i in block h (if it appears in the design), ϵ_{hi} is the corresponding error random variable, μ is a constant, τ_i is the effect of the ith treatment combination, and θ_h is the effect of the hth block. The block×treatment interaction is assumed to be negligible.

Table 13.1 Outline analysis of variance table for single-replicate factorial experiments constructed by the methods of this chapter

Source of variation	Degrees of freedom	Sum of squares
Blocks	$b - 1$	$ss\theta = \frac{1}{k} \Sigma B_h^2 - \frac{1}{v} G^2$
⋮	⋮	⋮
$\Sigma c_i \tau_i$	1	$ssc = \frac{(\Sigma \Sigma c_i y_{hi})^2}{\Sigma c_i^2}$
(at most $v - b$ of these)		
⋮	⋮	⋮
Error	df (by subtraction)	ssE (by subtraction)
Total	$v - 1$	$sstot = \Sigma \Sigma y_{hi}^2 - \frac{1}{v} G^2$

Analysis of all single-replicate designs described in this chapter is straightforward. Because of the way in which the designs will be constructed, contrasts in the important main effects and interactions will be completely orthogonal to block contrasts. As a consequence, these contrasts will have no adjustment for blocks, and their estimates and sums of squares can be calculated in exactly the same way as for completely randomized designs (see Chaps. 6 and 7). An outline of an analysis of variance table is shown in Table 13.1. The maximum number of degrees of freedom available for estimating main effects and interactions is $v - b$ if no estimate of the error variance is required; otherwise, it is $v - b - 1$. The unadjusted sum of squares for blocks $ss\theta$ can be calculated either as the total of all the confounded contrast sums of squares or by the usual formula, which was given in Table 11.7, p. 358, as

$$ss\theta = \frac{1}{k} \Sigma_h B_h^2 - \frac{1}{v} G^2 \,, \tag{13.2.2}$$

where B_h is the total of the observations in the hth block and G is the grand total of all the observations.

13.3   Confounding Using Contrasts

13.3.1   Contrasts

Treatment contrasts for factorial experiments were discussed in Sects. 6.3, 7.2.4, and 7.3. When there are two factors, A and B, each having two levels, there are $v = 4$ treatment combinations in total, and it is possible to find a set of three orthogonal contrasts, one for the main effects of each of A and B and one for their interaction. The coefficient lists $[c_{00}, c_{01}, c_{10}, c_{11}]$ for these contrasts are

$$\begin{aligned}
\text{For } A: \quad & [-1, -1, \ \ 1, \ \ 1 \,], \\
\text{For } B: \quad & [-1, \ \ 1, -1, \ \ 1 \,], \\
\text{For } AB: \quad & [\ \ 1, -1, -1, \ \ 1 \,].
\end{aligned}$$

Each coefficient for the AB interaction is the product of the corresponding coefficients for the main effects of A and B. Similarly, for three factors, the coefficient lists for the seven contrasts are shown in Table 13.2 written as columns. The interaction coefficients are, again, the product of corresponding main-effect coefficients. The row labels in Table 13.2 are the treatment combinations (in lexicographical order) whose observations are to be multiplied by the contrast coefficients when estimating the contrast. The seven contrasts are orthogonal. This can be verified by multiplying together corresponding digits

Table 13.2 Contrasts for a 2^3 experiment

	A	B	C	AB	AC	BC	ABC
000	−1	−1	−1	1	1	1	−1
001	−1	−1	1	1	−1	−1	1
010	−1	1	−1	−1	1	−1	1
011	−1	1	1	−1	−1	1	−1
100	1	−1	−1	−1	−1	1	1
101	1	−1	1	−1	1	−1	−1
110	1	1	−1	1	−1	−1	−1
111	1	1	1	1	1	1	1

in any two columns and showing that the sum of the products is zero. A table of orthogonal contrasts, such as Table 13.2, is sometimes called an *orthogonal array*.

Such contrasts in a 2^p experiment all have the same variance, since $\Sigma c_i^2 = v = 2^p$ for all contrasts and $\text{Var}(\Sigma c_i \hat{\tau}_i) = \Sigma c_i^2 \sigma^2 = v\sigma^2$. The main effect of A is often measured by the A contrast divided by $v/2$, so that it compares the average of all treatment combinations at the high level of A with the average of the treatment combinations at the low level. If the interaction contrast coefficients are also divided by $v/2$, the contrast estimators all have variance $\sum c_i^2 \sigma^2 = 4\sigma^2/v$, and the AB interaction, for example, then compares the average response when factors A and B are at the same level with the average response when they are at different levels. We shall use either $v/2$ or 1 for the divisor for all contrasts in 2^p experiments both here and in Chap. 15.

13.3.2   Experiments in Two Blocks

We start with an example. Suppose that a single-replicate 2^3 experiment is to be run in two blocks of size four. Suppose also that the experimenter knows that one of the factors, say factor A, does not interact with either of the other two factors. This means that the interactions AB, AC, and ABC may be assumed to be negligible and that the contrasts labeled A, B, C, and BC in Table 13.2 are the only contrasts to be measured.

Since there will be $b = 2$ blocks, it follows that $b - 1 = 1$ degree of freedom will be used to measure block differences and one treatment contrast will be confounded with blocks. Without too much difficulty, we can ensure that the confounded contrast is one of the negligible contrasts. For example, we can confound the negligible ABC contrast by placing in one block those treatment combinations corresponding to -1 in the ABC contrast, and placing in the second block those treatment combinations corresponding to $+1$ in the same contrast. Referring to Table 13.2, we can see that the design in Table 13.3 results. The ABC contrast is now identical to a block contrast that compares Block I with Block II, and consequently, the ABC contrast is confounded with blocks. The other two negligible contrasts, AB and AC, provide two degrees of freedom to estimate σ^2. Since all the nonnegligible factorial contrasts are orthogonal to AB, AC, and ABC, they can be measured as though there were no blocks present. Block design randomization (see Sect. 11.2.2) needs to be carried out before the design in Table 13.3 can be used in practice.

Table 13.3 2^3 experiment in 2 blocks of 4, confounding ABC

Block I	000	011	101	110
Block II	001	010	100	111

Table 13.4   Data for the field experiment (*ABCD* is confounded)

Block I		Block II	
TC	Yield	TC	Yield
0000	58	0001	55
0011	51	0010	45
0101	44	0100	42
0110	50	0111	36
1001	43	1000	53
1010	50	1011	55
1100	41	1101	41
1111	44	1110	48

Source Data adapted from *Experimental Designs*, Second Edition, by W.G. Cochran and G.M. Cox, 1957, John Wiley & Sons, New York

A similar method of confounding can be used for any 2^p experiment in $b = 2$ blocks of size $k = 2^{p-1}$. All factorial contrasts except for the one confounded contrast can be estimated.

Example 13.3.1 `Field experiment`

The data shown in Table 13.4 form part of the results of a field experiment on the yield of beans using various types of fertilization. The experiment was conducted at Rothamsted Experimental Station in 1936 and was reported by W.G. Cochran and G.M. Cox in their book *Experimental Designs*. There were four treatment factors each at two levels. Factor *A* was the amount of dung (0 or 10 tons) spread per acre, factors *B*, *C*, and *D* were the amounts of nitrochalk (0 and 45 lb), superphosphate (0 and 67 lb), and muriate of potash (0 and 112 lb), respectively, per acre. The experimental area was divided into two possibly dissimilar blocks of land, each of which was subdivided into eight plots (experimental units). Since this was a single-replicate experiment with $2^4 = 16$ treatment combinations (TC) divided into $b = 2$ blocks of size $k = 8$, one treatment contrast had to be confounded. The experimenters chose to confound the *ABCD* contrast, since the four-factor interaction was of least interest.

The *ABCD* contrast is shown below, and it can be verified that the treatment combinations corresponding to contrast coefficient $+1$ appear in Block I of Table 13.4, while those corresponding to coefficient -1 appear in Block II. All the other factorial contrasts are orthogonal to the *ABCD* contrast, so they can all be estimated without adjusting for the block effects. We take as examples the *B* and *BC* contrasts shown below.

TC	0000	0001	0010	0011	0100	0101	0110	0111
y_{hijkl}	58	55	45	51	42	44	50	36
B	-1	-1	-1	-1	1	1	1	1
BC	1	1	-1	-1	-1	-1	1	1
ABCD	1	-1	-1	1	-1	1	1	-1

TC	1000	1001	1010	1011	1100	1101	1110	1111
y_{hijkl}	53	43	50	55	41	41	48	44
B	-1	-1	-1	-1	1	1	1	1
BC	1	1	-1	-1	-1	-1	1	1
ABCD	-1	1	1	-1	1	-1	-1	1

Using rule 10 of Sect. 7.3, the least squares estimate of the *B* contrast is $\overline{y}_{.1..} - \overline{y}_{.0..} = \frac{1}{8}\Sigma c_{ijkl} y_{hijkl}$, where the sum is to be taken over the four subscripts, and the contrast coefficients c_{ijkl} are given in

standard order above. Multiplying the contrast coefficients by the data values, we obtain

$$\text{For } B : \tfrac{1}{8} \sum c_{ijkl} y_{hijkl} \;=\; -8.00.$$

Similarly, if we divide the BC contrast shown above by the same divisor $v/2$, we obtain the contrast estimate

$$\text{For } BC : \tfrac{1}{8} \sum c_{ijkl} y_{hijkl} \;=\; \tfrac{1}{8}(y_{..00.} - y_{..01.} - y_{..10.} + y_{..11.}) \;=\; 2.25.$$

Using (6.7.53), the sum of squares for testing the hypothesis that the main effect of B is negligible is

$$ssB = \frac{\left(\tfrac{1}{8} \sum c_{ijkl} y_{hijkl}\right)^2}{\sum \left(\tfrac{1}{8} c_{ijkl}\right)^2} = \frac{(-8.00)^2}{\frac{16}{64}} = 256.0.$$

Similarly, the sum of squares for testing the hypothesis that the interaction between B and C is negligible is

$$ss(BC) = \frac{(2.25)^2}{\frac{16}{64}} = 20.25.$$

An alternative way to calculate the sums of squares is to use the method of Sect. 7.3. Following the rules in that section, we obtain

$$ssB = acd \sum_j \bar{y}^2_{.j..} - abcd\,\bar{y}^2_{.....}$$
$$= 8\left[(51.25)^2 + (43.25)^2\right] - 16(47.25)^2 \;=\; 256.0\,,$$

and

Table 13.5 Analysis of variance for the field experiment

Source of variation	Degrees of freedom	Sum of squares	Contrast estimate (divisor $v/2$)
Block ($ABCD$)	1	2.25	
A	1	2.25	-0.75
B	1	256.00	-8.00
C	1	0.25	0.25
D	1	20.25	-2.25
AB	1	6.25	1.25
AC	1	81.00	4.50
AD	1	0.00	0.00
BC	1	20.25	2.25
BD	1	12.25	-1.75
CD	1	1.00	0.50
ABC	1	16.00	-2.00
ABD	1	16.00	2.00
ACD	1	20.25	2.25
BCD	1	121.00	-5.50
Total	15	575.00	

$$ss(BC) = ad \sum_j \sum_k \bar{y}^2_{..jk.} - acd \sum_j \bar{y}^2_{..j..} - abd \sum_k \bar{y}^2_{...k.} + abcd\bar{y}^2_{.....}$$

$$= 4\left[(52.25)^2 + (50.25)^2 + (42.00)^2 + (44.50)^2\right]$$

$$- 8\left[(51.25)^2 + (43.25)^2\right] - 8\left[(47.125)^2 + (47.375)^2\right]$$

$$+ 16(47.25)^2 = 20.25.$$

The complete analysis of variance table is shown in Table 13.5. The important contrasts can be identified using one of the methods of Sect. 7.5. A half-normal probability plot of the 14 contrast estimates is shown in Fig. 13.1. Note that we have not included the confounded $ABCD$ contrast in the half-normal probability plot. Although it is not the case here, the block effect is usually expected to be large and may draw attention away from the important treatment contrasts.

The contrasts B, BCD, and AC may or may not be important, since their plotted points roughly line up with those of the seven smallest absolute estimates, but not with those of the 11 smallest absolute estimates. Suppose they are considered important. We notice that the contrast estimate for B is negative, suggesting that the addition of nitrochalk decreased the yield of beans when averaged over the levels of A, C, and D. The interaction plot for BCD is shown in Fig. 13.2 and that for AC in Fig. 13.3. We see from Fig. 13.2 that the BCD interaction can be characterized by the fact that the CD interaction changes as B changes from its low level to its high level. If the objective of the experiment is to increase yield, then comparison of the two plots in Fig. 13.2 suggests that B should be set at its low level, unless

Fig. 13.1 Half-normal probability plot for the contrast absolute estimates of the field experiment
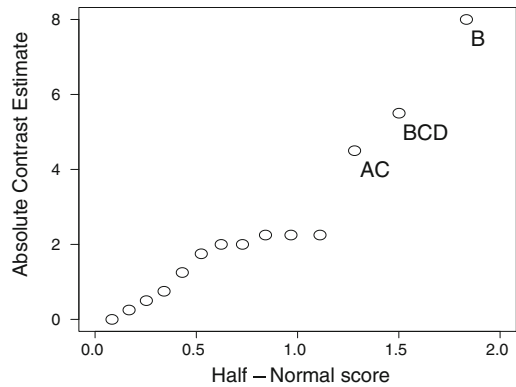
Fig. 13.2 BCD interaction plot for the field experiment

(a) B level 0

(b) B level 1
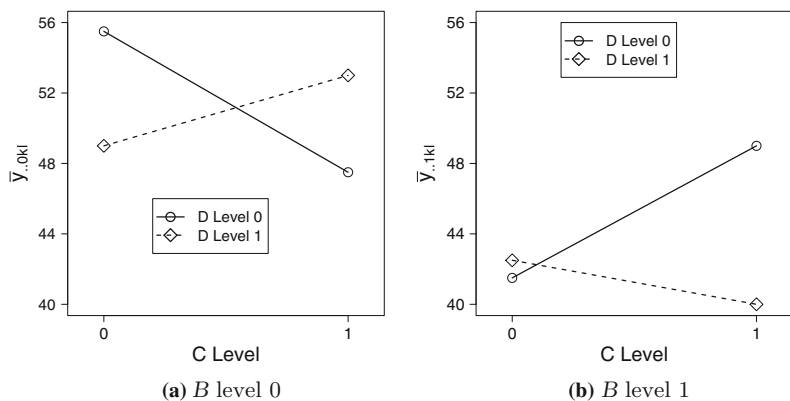
Fig. 13.3 *AC* interaction
plot for the field
experiment
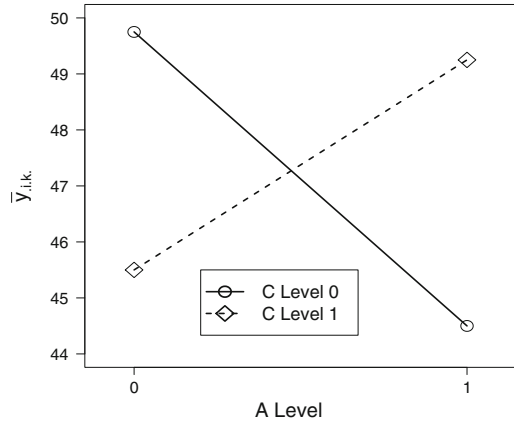
the high level of *C* and the low level of *D* are used. This tends to agree with the earlier observation
that the contrast estimate for *B* is negative, suggesting that the low level is better. If *B* is set at its low
level, the left-hand graph of Fig. 13.2 suggests that both *C* and *D* should be at their low levels. The
AC interaction plot in Fig. 13.3 shows that either both *C* and *A* should be at their low levels or both *C*
and *A* should be at their high levels. The contrast estimators for *A* and *D* are both negative, suggesting
that on average the low level is better, although the difference in yield is minor. More importantly, the
low levels in this experiment are cheaper. Therefore, all the evidence points towards not adding any
fertilizer ingredients in the quantities studied in the experiment. A followup experiment could be run
with the same four factors but with an increased "high" level of *C* and lower "high" levels of *A*, *B*,
and *D*. Since it is possible that the response is quadratic for each of the factors, a 3^4 experiment could
be run.

Suppose the experimenters had known ahead of time that factors *A* and *D* do not interact, so that
interactions *AD*, *ABD* and *ACD* could have been assumed negligible; then the corresponding terms
would have been omitted from the model. There would then have been 3 degrees of freedom for
estimating the error variance. The error sum of squares would have been the total of the sums of
squares for *AD*, *ABD*, and *ACD* listed in Table 13.5, so that

$$\hat{\sigma}^2 = msE = \frac{1}{3}(ss(AD) + ss(ABD) + ss(ACD)) = \frac{1}{3}(36.25) = 12.0833.$$

The analysis of variance table would then have been as shown in Table 13.6, and we see that at an
overall significance level of at most $\alpha = 11(0.005) = 0.055$, none of the contrasts would have been
judged as significantly different from zero, since $F_{1,3,0.005} = 55.55$.

Confidence intervals could be calculated for each contrast at an overall confidence level of at least
94.5%, using the Bonferroni method (formula (4.4.21)) with each interval at individual level 99.5%,
with error degrees of freedom $df = 3$ and with $r_i = 8$ being the number of observations averaged over
to obtain the estimate. For example, a confidence interval for the difference in the high and low levels
of *B* is

$$\beta_1^* - \beta_0^* \in \left(\frac{1}{8}\Sigma c_{ijkl}y_{hijkl} \pm t_{3,0.0025}\sqrt{msE\,(16/64)}\right)$$
$$= (-8 \pm 7.4532 \times 1.7381) = (-20.954, 4.954).$$

Table 13.6   Analysis of variance for the field experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Block	1	2.25	2.25	–	–
A	1	2.25	2.25	0.186	0.6952
B	1	256.00	256.00	21.186	0.0193
C	1	0.25	0.25	0.021	0.8947
D	1	20.25	20.25	1.676	0.2861
AB	1	6.25	6.25	0.517	0.5240
AC	1	81.00	81.00	6.703	0.0811
BC	1	20.25	20.25	1.676	0.2861
BD	1	12.25	12.25	1.014	0.3882
CD	1	1.00	1.00	0.083	0.7923
ABC	1	16.00	16.00	1.324	0.3332
BCD	1	121.00	121.00	10.014	0.0507
Error	3	36.25	12.0833		
Total	15	575.00			

We remind the reader that if both of the above analyses are done, that is, if the interactions AD, ABD, and ACD are dropped from the model after examining the half-normal probability plot, then it is no longer meaningful to talk about the significance levels of the tests or the confidence interval levels (see Sect. 6.5.6, p. 170).   □

13.3.3   Experiments in Four Blocks

We can extend the method of confounding that we used for two blocks to obtain $b = 4 = 2^2$ blocks. We then need to use two contrasts to divide up the treatment combinations. For example, suppose that in a 2^4 experiment, all interactions except for the two-factor interactions are thought to be negligible. We can select one of the negligible interactions to produce two blocks of size 8 and then select a second interaction to subdivide each of these two blocks into two smaller blocks, giving a total of 4 blocks of size 4. Since $b - 1 = 3$ degrees of freedom are needed to measure blocks, a third treatment contrast must also be confounded. We require this third contrast to be among the negligible contrasts, so care must be taken as to which pair of contrasts is initially selected for confounding. The choice of $ABCD$ and ABC, for example, is a very poor choice even if these high-order interactions may be thought to be negligible. We can see this by examining the design and the third confounded contrast as follows.

The two contrasts and the corresponding treatment combinations (TC) are shown in Table 13.7. The treatment combinations are divided into 2 blocks of size 8 according to the coefficients in the $ABCD$ contrast. Each of these blocks is then subdivided into 2 blocks of size 4 according to the coefficients in the ABC contrast. The $b = 4$ blocks of size $k = 4$ are therefore determined by the pairs of coefficients $(ABCD, ABC) = (-1, -1)$ or $(-1, 1)$ or $(1, -1)$ or $(1, 1)$ in the two contrasts. The resulting design is shown in Table 13.7 (prior to randomization).

Examination of the blocks in the design shows that all of the treatment combinations in Block I and Block IV have the fourth digit equal to 1, and all of those in Blocks II and III have the fourth digit equal to 0. This means that the high and low levels of factor D cannot be compared within the same block, and therefore the contrast for the main effect of D must be the third contrast confounded with blocks.

Table 13.7 2^4 experiment in 4 blocks of 4, confounding $ABCD$, ABC, D

TC	0000	0001	0010	0011	0100	0101	0110	0111
ABCD	1	−1	−1	1	−1	1	1	−1
ABC	−1	−1	1	1	1	1	−1	−1
TC	1000	1001	1010	1011	1100	1101	1110	1111
ABCD	−1	1	1	−1	1	−1	−1	1
ABC	1	1	−1	−1	−1	−1	1	1

Block	Contrast coefficients ($ABCD$, ABC)	Treatment combinations			
I	(−1, −1)	0001	0111	1011	1101
II	(−1, 1)	0010	0100	1000	1110
III	(1, −1)	0000	0110	1010	1100
IV	(1, 1)	0011	0101	1001	1111

We could have predicted this outcome, since if corresponding coefficients of the two contrasts $ABCD$ and ABC shown in Table 13.7 are multiplied together, the coefficients of the D contrast results. Notice that in symbols, we can write

$$(ABCD)(ABC) = A^2 B^2 C^2 D = D,$$

where any letter with exponent 2 is ignored. The squared coefficients of any 2^p factorial contrast are all $+1$, so multiplying the D contrast by C^2, say, is the same as multiplying the contrast coefficients by $+1$ and $C^2 D = D$. Multiplication of the contrast names in this way gives a quick, easy method of checking which third contrast is confounded without writing out the contrasts and without writing out the design.

The above design is not suitable for the stated experiment. Suppose that the contrasts ABD and BCD were selected for confounding instead. The third confounded contrast would then be $(ABD)(BCD) = AB^2CD^2 = AC$. This, too, is not suitable since all two-factor interactions were to have been measured. Unfortunately, there is no choice that will meet the specifications of this particular experiment. The number of blocks and the block sizes are too small to measure everything that is required. At least one two-factor interaction would have to be sacrificed, or a larger experiment must be run.

Example 13.3.2 2^5 experiment in 4 blocks of 8
In Sect. 13.5 we will describe a 2^5 experiment that was run in $b = 4$ blocks of size $k = 8$. In designing such an experiment, one needs to select two contrasts for confounding and to check that the third confounded contrast is acceptable. As in the discussion above, selecting the 5-factor interaction for confounding will generally be a poor choice, since no matter which other interaction is selected for the second confounded interaction, a 2-factor interaction or main effect will be among the confounded contrasts. For example,

$$(ABCD)(ABCDE) = E \quad \text{and} \quad (ABC)(ABCDE) = DE.$$

If an experimenter knew ahead of time that factors D and E do not interact, then the second choice might be acceptable. In general, though, most experimenters would prefer not to confound low-order interactions. So, a selection of a 3-factor interaction and a 4-factor interaction with as few letters in common as possible will generally be the best choice. For example,

$$(ABCD)(CDE) = ABE.$$

There are many selections of this type, and the experimenter would wish to avoid confounding any 3-factor interaction that might be of some interest. The selection made in Sect. 13.5 is ABD, BCE, and their product $ACDE$. If the treatment combinations are written out in standard order together with the contrast coefficients for ABD and BCE, it can be verified that the pairs of contrast coefficients give the four blocks shown in Table 13.11 (p. 448).   □

13.3.4   Experiments in Eight Blocks

If an experiment is required in $b = 2^3$ blocks, then three contrasts must be selected for confounding. A single-replicate design in eight blocks confounds $b - 1 = 7$ treatment contrasts in total, including the three contrasts initially selected and all products of these. For example, suppose that a 2^6 experiment is required in $b = 2^3 = 8$ blocks of 8, and that the two-factor interactions are of interest together with the four three-factor interactions ACE, ACD, ADE, and CDE (these are the four 3-factor interactions that do not contain B or F). A suitable choice might be to confound the interactions BCD, ABE, and ADF. The other four confounded contrasts would be

$$(BCD)(ABE) = ACDE,$$
$$(BCD)(ADF) = ABCF,$$
$$(ABE)(ADF) = BDEF,$$
$$(BCD)(ABE)(ADF) = CEF.$$

The list of seven confounded contrasts is called the *confounding scheme* for the design. The reader is invited to write out the three selected contrasts and verify that the design in Table 13.10 (p. 446) results. The fact that $ACDE$, $ABCF$, $BDEF$, and ABE are also confounded can be verified by showing that the coefficients of each of these four contrasts are constant for all treatment combinations within each block.

Note that the same design will be obtained for any initial selection of three of the above seven confounded contrasts, *provided that no selected contrast is a product of the other two*. For example, suppose that $ABCF$, ABE, and CEF are initially selected. The selected contrasts $ABCF$ and ABE divide the treatment combinations into four blocks, but CEF does not subdivide these blocks further, since it is the third contrast automatically confounded in the four blocks, i.e., $(ABCF)(ABE) = CEF$. A different third contrast needs to be chosen, and any of the remaining four contrasts will do. Three selected contrasts satisfying the requirement that no selected contrast be a product of the other two is called a set of three *independent* contrasts.

13.3.5   Experiments in More Than Eight Blocks

The same ideas can be used for 2^p experiments in 2^s blocks of size $k = 2^{p-s}$ by selecting s independent contrasts to subdivide the treatment combinations into blocks; s contrasts are independent if none can be obtained as the product of two or more of the other $s - 1$ contrasts. Although the multiplication of contrast names is a convenient method to determine the list of confounded contrasts, it becomes harder to use the contrasts themselves for constructing the design as the number of factors increases.

Table 13.8 2^3 experiment
in 2 blocks of 4,
confounding ABC

Block I	000	011	101	110
Block II	001	010	100	111

In the next section, we present a method of constructing block designs for single-replicate factorial experiments that avoids writing out the contrasts.

13.4 Confounding Using Equations

13.4.1 Experiments in Two Blocks

The design in Table 13.8 (given previously in Table 13.3) was constructed by allocating the treatment combinations to blocks in such a way that the contrast that compares Block I with Block II is identical to the ABC contrast. The ABC contrast is confounded with blocks and cannot be estimated, but all of the other factorial contrasts are estimable because they are orthogonal to the confounded contrast.

   If the treatment combinations in the two blocks of the design are examined closely, an interesting property becomes apparent. All the treatment combinations in the first block have an even number of 1's, and all those in the second block have an odd number of 1's. We could, in fact, have predicted this property. The ABC contrast is the product of the A, B, and C contrasts (see Sect. 13.3.1), and so the only way to achieve a coefficient -1 in the ABC contrast is for there to be an odd number of -1's among the corresponding coefficients in the A, B, and C contrasts. This means that there must be an odd number of 0's and an even number of 1's in the corresponding treatment combination. Thus, if we want to confound ABC without writing out the contrasts, we can simply allocate a treatment combination $a_1a_2a_3$ to Block I if it has an even number of 1's among its digits, and to Block II if it has an odd number of 1's. Equivalently, we allocate $a_1a_2a_3$ to Block I if $a_1 + a_2 + a_3$ is an even number and to Block II if $a_1 + a_2 + a_3$ is an odd number. Instead of writing "$a_1 + a_2 + a_3$ is an even number," we write "$a_1 + a_2 + a_3 = 0$ (mod 2)" and instead of writing "$a_1 + a_2 + a_3$ is an odd number," we write "$a_1 + a_2 + a_3 = 1$ (mod 2)". Working mod 2, or modulo 2, means that we subtract 2 repeatedly from the number until we reach either 0 or 1, or equivalently, we divide by 2 and take the remainder which is either 0 or 1. For example, $5 = 1$ (mod 2), but $8 = 0$ (mod 2). We call the pair of equations "$a_1 + a_2 + a_3 = 0$ (mod 2)" and "$a_1 + a_2 + a_3 = 1$ (mod 2)" the *confounding equations*. The design of Tables 13.3 and 13.8 is constructed using the following rule:

$$\text{Block I: Treatment combinations with } a_1 + a_2 + a_3 = 0 \text{ (mod 2)},$$
$$\text{Block II: Treatment combinations with } a_1 + a_2 + a_3 = 1 \text{ (mod 2)}.$$

   If the AC contrast were to be confounded in the 2^3 experiment instead of the ABC contrast, only the first and third digits of each treatment combination would be used to allocate it to a block. This is because only the first and third digits of the treatment combination govern the coefficients in the AC contrast. We would use the pair of confounding equations $a_1 + a_3 = 0$ (mod 2) and $a_1 + a_3 = 1$ (mod 2), and the design would be constructed using the rule

$$\text{Block I: Treatment combinations with } a_1 + a_3 = 0 \text{ (mod 2)},$$
$$\text{Block II: Treatment combinations with } a_1 + a_3 = 1 \text{ (mod 2)},$$

giving the design of Table 13.9. It can be verified that AC is indeed confounded with blocks in this design, since the coefficients of the AC contrast (shown in Table 13.2) are all equal to $+1$ for the

Table 13.9 2^3 experiments in 2 blocks of 4, confounding *AC*					
	Block I	000	010	101	111
	Block II	001	011	100	110

treatment combinations in Block I and -1 for those in Block II. All other contrasts are estimable because they are orthogonal to the confounded *AC* contrast.

We may now generalize to 2^p experiments in 2 blocks of size 2^{p-1}. If the interaction $A^{z_1}B^{z_2}C^{z_3}\cdots$ P^{z_p} is to be confounded with blocks, where $z_i = 1$ if the factor is present in the interaction and $z_i = 0$ if it is not, the blocks of the design are

$$\text{Block I: Treatment combinations with}$$
$$z_1 a_1 + z_2 a_2 + z_3 a_3 + \cdots + z_p a_p = 0 \ (\text{mod } 2),$$
$$\text{Block II: Treatment combinations with}$$
$$z_1 a_1 + z_2 a_2 + z_3 a_3 + \cdots + z_p a_p = 1 \ (\text{mod } 2).$$

13.4.2   Experiments in More Than Two Blocks

We obtain designs with 4, 8, 16, ... blocks by using more than one pair of confounding equations. We label the pairs of confounding equations as L_1, L_2, etc. For example, the (unsatisfactory) design

Block	Treatment Combinations
I	0000 0110 1010 1100
II	0011 0101 1001 1111
III	0001 0111 1011 1101
IV	0010 0100 1000 1110

from Table 13.7 for a 2^4 experiment in 4 blocks of size 4 was produced by confounding the *ABCD* and *ABC* contrasts. Using the *ABCD* contrast to produce two blocks is equivalent to using the pair of confounding equations $L_1 = a_1 + a_2 + a_3 + a_4 = 0 \ (\text{mod } 2)$ and $L_1 = 1 \ (\text{mod } 2)$, and using the *ABC* contrast is equivalent to using the pair of confounding equations $L_2 = a_1 + a_2 + a_3 = 0 \ (\text{mod } 2)$ and $L_2 = 1 \ (\text{mod } 2)$. Thus there are four possible values for the pair (L_1, L_2), and it can be verified that the blocks of the design satisfy

$$\text{Block I:} \quad L_1 = a_1 + a_2 + a_3 + a_4 = 0; \ L_2 = a_1 + a_2 + a_3 = 0 \ (\text{mod } 2);$$
$$\text{Block II:} \quad L_1 = a_1 + a_2 + a_3 + a_4 = 0; \ L_2 = a_1 + a_2 + a_3 = 1 \ (\text{mod } 2);$$
$$\text{Block III:} \ L_1 = a_1 + a_2 + a_3 + a_4 = 1; \ L_2 = a_1 + a_2 + a_3 = 0 \ (\text{mod } 2);$$
$$\text{Block IV:} \ L_1 = a_1 + a_2 + a_3 + a_4 = 1; \ L_2 = a_1 + a_2 + a_3 = 1 \ (\text{mod } 2).$$

We already know from Sect. 13.3.3 that a third contrast, namely contrast *D*, is confounded in this design. If we add the two confounding equations used to create each block, we have, for Block I,

$$
\begin{aligned}
L_1 = {}& a_1 + a_2 + a_3 + a_4 = 0 \ (\text{mod } 2)\\
+\, L_2 = {}& a_1 + a_2 + a_3 \qquad\ \ = 0 \ (\text{mod } 2)\\
\hline
L_1 + L_2 = {}& 2a_1 + 2a_2 + 2a_3 + a_4 = 0 \ (\text{mod } 2)
\end{aligned}
$$

If all coefficients are reduced modulo 2, the sum gives $a_4 = 0 \ (\text{mod } 2)$, which indicates that the *D* contrast is also confounded.

As has been demonstrated, there is a correspondence between the contrasts, the contrast names, and the confounding equations. The contrast names are the most convenient for checking the total list of confounded contrasts, and the equations are the most convenient for constructing the design.

Design construction can be done in several ways. One way is to examine each of the 2^p treatment combinations and allocate them to blocks according to their values obtained in the left sides of the confounding equations. Another way is to identify the treatment combinations that make the confounding equations equal to zero. These will form Block I of the design. The other blocks of the design are obtained by adding (modulo 2) to the treatment combinations in Block I any treatment combination that has not yet appeared in a block. This is illustrated in the following example.

Example 13.4.1 2^6 experiment in 8 blocks of 8
A confounding scheme was found in Sect. 13.3.4 for a 2^6 experiment in 8 blocks of 8 by selecting contrasts *BCD*, *ABE*, and *ADF* for confounding. It was shown that contrasts *ACDE*, *ABCF*, *BDEF*, and *CEF* were also confounded. Writing out the equations for the three selected contrasts, we have

$$
\begin{aligned}
L_1 &= \phantom{a_1 + {}} a_2 + a_3 + a_4 = 0 \text{ or } 1 \ (\text{mod } 2)\,, \\
L_2 &= a_1 + a_2 + a_5 = 0 \text{ or } 1 \ (\text{mod } 2)\,, \\
L_3 &= a_1 + a_4 + a_6 = 0 \text{ or } 1 \ (\text{mod } 2)\,.
\end{aligned}
$$

The equations corresponding to the other four confounded contrasts are obtained by setting $L_1 + L_2$, $L_1 + L_3$, $L_2 + L_3$, and $L_1 + L_2 + L_3$ equal to 0 or 1 (mod 2).

The design is shown in Table 13.10. It can be constructed systematically as follows. The first treatment combination 000000 gives 0 for each of L_1, L_2, L_3 and is allocated to Block I. The second treatment combination 000001 gives values 0, 0, 1 for the three L_i and is allocated to Block II, and so on.

Alternatively, one can look for the eight treatment combinations that give zero for each of L_1, L_2, and L_3 and construct Block I first. Solving $L_1 = 0$, $L_2 = 0$, and $L_3 = 0$ each for the last a_i gives $a_4 = a_2 + a_3 \ (\text{mod } 2)$, $a_5 = a_1 + a_2 \ (\text{mod } 2)$ and $a_6 = a_1 + a_4 \ (\text{mod } 2)$, respectively. For each one

Table 13.10 2^6 experiment in 8 blocks of 8, confounding *ABE*, *ADF*, *BCD*, *CEF*, *ABCF*, *ACDE*, *BDEF*

Block	L_1, L_2, L_3	Treatment Combinations			
I	0, 0, 0	000000	001101	010111	011010
		100011	101110	110100	111001
II	0, 0, 1	000001	001100	010110	011011
		100010	101111	110101	111000
III	0, 1, 0	000010	001111	010101	011000
		100001	101100	110110	111011
IV	0, 1, 1	000011	001110	010100	011001
		100000	101101	110111	111010
V	1, 0, 0	000101	001000	010010	011111
		100110	101011	110001	111100
VI	1, 0, 1	000100	001001	010011	011110
		100111	101010	110000	111101
VII	1, 1, 0	000111	001010	010000	011101
		100100	101001	110011	111110
VIII	1, 1, 1	000110	001011	010001	011100
		100101	101000	110010	111111

of of the eight combinations a_1, a_2, a_3 of factors A, B, and C, the corresponding values of a_4, a_5, and a_6 can thus be computed to obtain one of the eight treatment combinations of Block I. For example, if $a_1 = a_2 = a_3 = 1$, then $a_4 = a_2 + a_3 = 1 + 1 = 0 \pmod 2$, $a_5 = a_1 + a_2 = 1 + 1 = 0 \pmod 2$, and $a_6 = a_1 + a_4 = 1 + 0 = 1 \pmod 2$, so the treatment combination 111001 is in Block I.

Each of the other blocks is obtained by adding a new treatment combination to those in Block I—"new" meaning not yet in a block. For example, the treatment combination 000001 is not in Block I and if 000001 is added modulo 2 to the eight treatment combinations in Block I, then Block II results. For example, 000001 is added to 111001 of Block I by adding corresponding digits and reducing modulo 2; that is,

$$000001 + 111001 = 111002 = 111000 \pmod 2,$$

so 111000 is also in Block II.

A third block can be obtained by taking any treatment combination not in the first two blocks, and adding it to each treatment combination in Block I. Proceeding in this fashion, blocks are constructed until each treatment combination has been allocated to some block. □

13.5 A Real Experiment—Mangold Experiment

O. Kempthorne in his book *Design and Analysis of Experiments* describes an experiment run at Rothamsted Agricultural Station to investigate the effects of five different fertilizers on the growth of mangold roots. The five factors were Sulphate of ammonia (factor A at levels 0 or 0.6 cwt per acre), Superphosphate (factor B at levels 0 or 0.5 cwt per acre), Muriate of potash (factor C at levels 0 or 1.0 cwt per acre), Agricultural salt (factor D at levels 0 or 5 cwt per acre), and Dung (factor E at levels 0 or 10 tons per acre). The experimental area was divided into $b = 4$ blocks of size $k = 8$. All 3-, 4-, and 5-factor interactions were expected to be negligible. The two three-factor interactions *ABD*, *BCE*, and their product *ACDE* were selected for confounding.

The division of the 32 treatment combinations into the four blocks was then determined from the confounding equations corresponding to *ABD* and *BCE*; that is,

$$L_1 = a_1 + a_2 \quad + a_4 \quad = 0 \text{ or } 1 \pmod 2,$$
$$L_2 = \quad a_2 + a_3 \quad + a_5 = 0 \text{ or } 1 \pmod 2.$$

The blocks can be formed by systematically working through all 32 treatment combinations and assigning them to the blocks according to the values of L_1 and L_2 and the rule

$$\begin{aligned} \text{Block I:} \quad & L_1 = 0 \pmod 2 \quad \text{and} \quad L_2 = 0 \pmod 2, \\ \text{Block II:} \quad & L_1 = 0 \pmod 2 \quad \text{and} \quad L_2 = 1 \pmod 2, \\ \text{Block III:} \quad & L_1 = 1 \pmod 2 \quad \text{and} \quad L_2 = 0 \pmod 2, \\ \text{Block IV:} \quad & L_1 = 1 \pmod 2 \quad \text{and} \quad L_2 = 1 \pmod 2. \end{aligned}$$

Alternatively, we can notice that $L_1 = 0$ gives $a_4 = a_1 + a_2 \pmod 2$ and $L_2 = 0$ gives $a_5 = a_2 + a_3 \pmod 2$. Computing these for each combination of levels $a_1 a_2 a_3$ of the first three factors gives Block I as

Block I: 00000 10010 00101 10111
11100 01110 11001 01011

Any treatment combination that is not in Block I can be added to the treatment combinations in Block I to obtain a second block. So, if we add 00001, for example, we obtain

Block II: 00001 10011 00100 10110
11101 01111 11000 01010

Any treatment combination not in Blocks I and II can now be added to the treatment combinations in Block I to obtain a third block. Since 00010, for example, has not appeared in Blocks I or II, we can add it to the treatment combinations in Block I to obtain Block III. Block IV can then be obtained by adding a treatment combination, say 00011, that has not appeared in the previous three blocks.

Block III: 00010 10000 00111 10101
11110 01100 11011 01001
Block IV: 00011 10001 00110 10100
11111 01101 11010 01000

The order of the treatment combinations was randomized within each block, and the order of the blocks was randomized. The final design, together with the resulting yields (in pounds) of mangold roots, is shown in Table 13.11.

The contrasts for the main effects and interactions are obtained as usual. The contrast for comparing the high and low levels of superphosphate (B), for example, is $\sum_i c_i \tau_i$, where τ_i is the effect of the ith treatment combination, and the coefficient c_i is the ith element of

$$\frac{1}{16}[-1, \ -1, \ -1, \ -1, \ -1, \ -1, \ -1, \ -1, \ \ 1, \ \ 1, \ \ 1, \ \ 1, \ \ 1, \ \ 1, \ \ 1, \ \ 1,$$
$$-1, \ -1, \ -1, \ -1, \ -1, \ -1, \ -1, \ -1, \ \ 1, \ \ 1, \ \ 1, \ \ 1, \ \ 1, \ \ 1, \ \ 1, \ \ 1 \ ].$$

Table 13.11   Yields (in pounds) of mangold roots for the mangold experiment

| Block | Treatment Combinations | | | |
	(Yield)			
I	01101	00011	10100	11111
	(844)	(1104)	(1156)	(1508)
	11010	00110	10001	01000
	(1312)	(1000)	(1176)	(888)
II	00001	01111	00100	10011
	(1248)	(1100)	(784)	(1376)
	11101	10110	11000	01010
	(1356)	(1376)	(1008)	(964)
III	00101	11001	01011	01110
	(896)	(1284)	(996)	(860)
	10010	11100	00000	10111
	(1184)	(984)	(740)	(1468)
IV	10101	11110	00111	11011
	(1328)	(1292)	(1008)	(1324)
	01001	01100	00010	10000
	(1008)	(692)	(780)	(1108)

Source Design and Analysis of Experiments, by O. Kempthorne, 1976. Reprinted by permission of Krieger Publishing company Inc.

Equivalently, $c_i = -1/16$ if the ith treatment combination has B at the low level, and $c_i = 1/16$ otherwise. There is only one observation on each treatment combination, so a least squares estimate of τ_i is just y_i (the observation on the ith treatment combination). The least squares estimate of the contrast for comparing the high and low levels of B is

$$\sum_i c_i \hat{\tau}_i = -312/16 = -19.5 \,.$$

The contrast estimates for each of the main effects and also the interactions are shown in Table 13.12.

   To test the hypothesis that adding superphosphate has no effect on the yield of mangold roots, we test the null hypothesis $H_0^B : \sum_i c_i \tau_i = 0$ against the alternative hypothesis $H_A^B : \sum_i c_i \tau_i \neq 0$ using the rules of Sect. 7.3; that is,

$$\text{reject } H_0^B \quad \text{if} \quad \frac{ssB}{msE} = \frac{(\sum_i c_i \hat{\tau}_i)^2}{(\sum_i c_i^2) \, msE} = \frac{(\sum_i c_i y_i)^2}{(32/16^2) \, msE} > F_{1,13,\alpha/m} \,.$$

If the 3-, 4-, and 5-factor interactions are assumed to be negligible, the total of their contrast sums of squares, apart from ABD, BCE, and $ACDE$, which are confounded with blocks, forms the error sum of squares with 13 degrees of freedom, ($10 - 2 = 8$ from the 3-factor interactions, $5 - 1 = 4$ from the 4-factor interactions, and 1 from the 5-factor interaction). There are $m = 15$ contrasts of interest, so if each individual contrast is tested at significance level $\alpha/m = 0.005$, the overall significance level would be at most $\alpha = 0.075$. For the main effect of B,

$$\frac{(\sum_i c_i y_i)^2}{(32/16^2) \, msE} = \frac{8(-19.5)^2}{msE} = \frac{3042.00}{6791.23} < F_{1,13,0.005} = t_{13,0.0025}^2 = 11.37 \,,$$

Table 13.12 Analysis of variance for the mangold experiment (ABD, BCE, $ACDE$ are confounded)

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	Contrast estimate	p-values
Block	3	52832	17610.67	–	–	–
A	1	887112	887112.00	130.63	333.0	0.0001
B	1	3042	3042.00	0.45	−19.5	0.5150
C	1	722	722.00	0.11	9.5	0.7496
D	1	144722	144722.00	21.31	134.5	0.0005
E	1	262088	262088.00	38.59	181.0	0.0001
AB	1	338	338.00	0.05	6.5	0.8269
AC	1	48050	48050.00	7.08	77.5	0.0196
AD	1	16562	16562.00	2.44	45.5	0.1424
AE	1	288	288.00	0.04	−5.0	0.8400
BC	1	6272	6272.00	0.92	−28.0	0.3541
BD	1	5832	5832.00	0.86	27.0	0.3710
BE	1	98	98.00	0.01	−3.5	0.9062
CD	1	30752	30752.00	4.53	62.0	0.0530
CE	1	882	882.00	0.13	−10.5	0.7244
DE	1	13778	13778.00	2.03	−41.5	0.1779
Error	13	88286	6791.23			
Total	31	1561656				

so we fail to reject the null hypothesis H_0^B and conclude that there is no evidence to suggest a difference in yield due to the high and low levels of superphosphate. The ratio for each of the other hypothesis tests is given in Table 13.12, and we see that the only hypotheses that would be rejected at overall significance level 0.075 are the hypotheses of no effect of A or D or E.

The block sum of squares is the sum of the ABD, BCE, and $ACDE$ contrast sums of squares, or alternatively, it can be calculated as

$$ss\theta = \frac{1}{8} \sum_h B_h^2 - \frac{1}{32} G^2 = 52832.0$$

as in (13.2.2), where $B_h = y_{h....}$ and $G = y_{......}$. The block mean square is more than twice the size of the error mean square, indicating that the creation of blocks in this experiment was useful for reducing the error variability, assuming that the effects confounded with blocks are negligible.

Since none of the interactions appear to be significantly different from zero, the main effects can be investigated. The contrast estimates of the significant main effects are all positive, suggesting that the high levels of A, D, and E increase the yield of mangold roots significantly. Confidence intervals for all the main-effect contrasts can be obtained via Bonferroni's method using formula (4.4.21), but with error degrees of freedom $df = 13$. If we select the confidence level to match the α level of the hypothesis tests, we obtain a set of simultaneous confidence intervals with overall confidence level of at least 92.5% as follows:

$$
\begin{aligned}
\text{For } A: \ & \left(\overline{y}_{.1...} - \overline{y}_{.0...} \pm t_{13,0.0025} \sqrt{32\hat{\sigma}^2/16^2} \right) \\
& = \left(333.0 \pm 3.372 \sqrt{6791.23/8} \right) \\
& = (333.0 \pm 98.25) = (234.75, \ 431.25) \, ; \\
\text{For } B: \ & (-19.5 \pm 98.25) = (-117.75, \ 78.75) \, ; \\
\text{For } C: \ & \ \ \ (9.5 \pm 98.25) = (-88.75, \ 107.75) \, ; \\
\text{For } D: \ & (134.5 \pm 98.25) = (36.25, \ 232.75) \, ; \\
\text{For } E: \ & (181.0 \pm 98.25) = (82.75, \ 279.25) \, .
\end{aligned}
$$

At a somewhat higher overall significance level, the hypothesis of no interaction between A and C would have been rejected. The AC interaction plot is shown in Fig. 13.4. This plot agrees with the earlier observation that A should be set at its high level. Unless the cost is high, the plot suggests that C should also be set at its high level. Factor B does not seem to affect the yield much and can be set at its low (zero) level. As stated earlier, D and E should be at their high levels.

Fig. 13.4 AC interaction plot for the mangold experiment
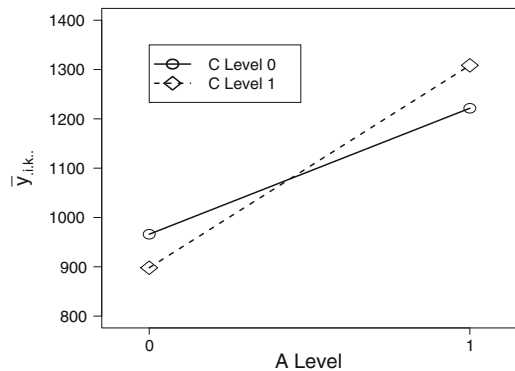
After an experiment of this type, it is good policy to run a *confirmatory experiment* verifying that the selected levels are, indeed, a good combination. Here the recommendation is to use treatment combination 10111. Certainly, in the main experiment, this treatment combination gave the highest yield in Block III, but it cannot easily be compared with the other observations because of the large block differences. A few more observations on this particular treatment combination would help to verify that it is consistently a good choice.

13.6 Plans for Confounded 2^p Experiments

Suggestions for confounding schemes useful for constructing designs for 2^p experiments in $b = 2^s$ blocks of size $k = 2^{p-s}$ are given in Table 13.29 at the end of this chapter. These have been chosen to allow all main effects to be estimable and as many two-factor interactions as possible. The first block of each design can be obtained from the given relations on the factor levels, and the other blocks can be obtained by adding (modulo 2) new treatment combinations to those in Block I. This process was illustrated in Example 13.4.1, p. 446.

If one or more of the listed confounded contrasts is an important contrast in the experiment being designed, the factors should be relabeled. For example, suppose that a design is required for a 2^6 experiment in 8 blocks of 8. The design listed in Table 13.29 confounds *BCD*, *ABE*, *ACDE*, *ADF*, *ABCF*, *BDEF*, and *CEF* (this is also the design of Table 13.10). Suppose that the experimenter wished to estimate the contrasts *ABC* and *BCD*. The design, as listed, confounds *BCD*. However, if the experimenter were to switch the labels *B* and *E* of the actual treatment factors, then the important contrasts would be called *ACE* and *CDE*, both of which can be estimated in the listed design. Equivalently, the experimenter could switch the labels *B* and *E* of the listed design, so as to confound *CDE*, *ABE*, *ABCD*, *ADF*, *ACEF*, *BDEF*, and *BCF*, leaving *ABC* and *BCD* unconfounded.

13.7 Multireplicate Designs

When the experiment is large enough that each treatment combination can be observed $r > 1$ times, we have a choice of several different ways to design the experiment. If the block size can be chosen to be as large as the number of treatment combinations, then a randomized block design (Chap. 10) can be used.

If an incomplete block design is required, we could choose to use one of the standard incomplete block designs described in Chap. 11, such as a balanced incomplete block design. Alternatively, we could take a single-replicate design, confounding one or more interaction contrasts, and repeat the design r times. Alternatively, again, we could piece together r different single-replicate designs, confounding different contrasts in each. The confounding of a contrast in some but not all replicates is called *partial confounding*.

The choice between these three types of incomplete block designs involves tradeoffs concerning what is to be estimated and how accurately. The analysis of a standard incomplete block design of Chap. 11 requires that *every* contrast estimate be adjusted for blocks. Thus, while all contrasts are estimable, their least squares estimators have higher variances than they would if they were unadjusted for blocks—there is some *loss of information* on each contrast. If a balanced incomplete block design is used, then there is the same loss of information on every treatment contrast.

A repeated single-replicate design, on the other hand, loses information completely on the confounded contrasts—they are not estimable and are said to be *completely confounded*—but all contrasts orthogonal to these are unconfounded and can be estimated with no block adjustment and so with no

loss of information. The estimator of any unconfounded contrast is the same as it would be in a complete block design, and with the same variance formula. The use of smaller blocks should, however, result in a smaller error variance σ^2.

The designs with partial confounding fall between these two extremes, allowing all contrasts to be estimable but with different levels of adjustment and loss of information. Complete and partial confounding are illustrated in Sects. 13.8–13.9 and are compared via an example in Sect. 13.10.

13.8   Complete Confounding: Repeated Single-Replicate Designs

If the number of treatment combinations v is divisible by the block size k, the v/k blocks of a single-replicate design can be repeated r times to give an incomplete block design with $b = rv/k$ blocks. The contrasts that are confounded in the single-replicate design are also confounded in the r-replicate design—they cannot be estimated and are said to be *completely confounded*. For all other contrasts, the estimators and corresponding variance formulae are as in a complete block design—no block adjustments are needed.

13.8.1   A Real Experiment—Decontamination Experiment

An experiment was described by M.K. Barnett and F.C. Mead, Jr. in the journal *Applied Statistics* in 1956 to explore the effect of four factors on the efficiency of a decontamination process for the removal of radioactive isotopes from liquid waste. The four treatment factors were:

- A: The amount of aluminum sulphate added to the liquid waste (two levels, 0.4 g and 2.5 g per liter, coded 0, 1).
- B: The amount of barium chloride added to the liquid waste (two levels, 0.4 g and 2.5 g per liter, coded 0, 1).
- C: The amount of carbon added to the liquid waste (two levels, 0.08 g and 0.4 g per liter, coded 0, 1).
- D: Final pH of liquid waste (two levels, 6 and 10, coded 0, 1) achieved by adding sodium hydroxide or hydrochloric acid.

The experimental units were portions of a typical laboratory waste of pH 8.3 and in which the principal radioactivity was attributable to salts of radium, thorium, and actinium. The measurements

Table 13.13 Repeated single-replicate design for a 2^4 experiment in 2×2 blocks of size 8, confounding *ABCD*. Data for the decontamination experiment are shown in parentheses

Block	Treatment combinations (response)							
I	1010	1111	0110	0000	1100	0101	0011	1001
	(183)	(350)	(188)	(881)	(225)	(298)	(1039)	(466)
II	0010	0001	0111	1000	1101	0100	1110	1011
	(650)	(1180)	(238)	(191)	(420)	(289)	(135)	(781)
III	0101	1001	1100	0000	1010	1111	0110	0011
	(273)	(890)	(370)	(834)	(193)	(389)	(163)	(1146)
IV	0001	1110	1000	0100	0111	1011	0010	1101
	(1193)	(156)	(257)	(178)	(254)	(775)	(494)	(429)

Source Barnett and Mead (1956). Copyright © 1956 Blackwell Publishers. Reprinted with permission

taken after the experimental decontamination process were the counts per minute per milliliter of alpha and beta particles. Here, we will only reproduce the data for the alpha particles (see Table 13.13).

Only 8 of the 16 treatment combinations could be examined per day. Four days were available for the experiment, allowing each treatment combination to be measured twice during the course of the experiment. The experimenters anticipated day-to-day variations in the observations and decided to run a block design with $b = 4$ blocks of size $k = 8$. In fact, an unforeseen change of operators became necessary at the end of the first day. Since a block design had been used, any shift in the observations due to the operator change was absorbed into the block differences and did not affect the comparisons of the treatment combinations.

The experimenters first selected a single-replicate design in $b^* = 2$ blocks of size $k = 8$ that confounded the 4-factor interaction contrast $ABCD$ (which they thought unlikely to exist in this experiment). By using this single-replicate design twice, they obtained a design with $r = 2$ observations per treatment combination and with $b = 2b^* = 4$ blocks in which all contrasts except for $ABCD$ could be measured without adjustments for blocks. (The single-replicate design selected is that listed in Table 13.29.) The treatment combinations were randomly ordered within each block, and the final design is shown in Table 13.13.

The chosen model included all main effects and all 2-factor and 3-factor interactions. The block×treatment interaction was assumed to be negligible. Thus, the model was

$$
\begin{aligned}
Y_{hijkl} = {}& \mu + \theta_h + \alpha_i + \beta_j + \gamma_k + \delta_l \\
& + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{il} + (\beta\gamma)_{jk} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} \\
& + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\delta)_{ijl} + (\alpha\gamma\delta)_{ikl} + (\beta\gamma\delta)_{jkl} + \epsilon_{hijkl}, \\
& \epsilon_{hijkl} \sim N(0, \sigma^2) \text{ and mutually independent}, \\
& h = 1, 2, 3, 4; \;\; i = 0, 1; \;\; j = 0, 1; \;\; k = 0, 1; \;\; l = 0, 1; \\
& \qquad (h, i, j, k, l) \text{ in the design.}
\end{aligned}
$$

The analysis of variance table is shown in Table 13.14. There are $b = 4$ blocks in the design, but $ABCD$ is the only treatment contrast confounded. The sum of squares for $ABCD$ is included in the block sum of squares. The sums of squares for the other interactions and main effects can be obtained from rules 4 or 12 in Sect. 7.3. For example, the sum of squares for testing the hypothesis of negligible AC interaction is

$$
\begin{aligned}
ss(AC) &= \sum_i \sum_k \frac{y_{i.k.}^2}{8} - \sum_i \frac{y_{i...}^2}{16} - \sum_k \frac{y_{..k.}^2}{16} + \frac{y_{....}^2}{32} \\
&= \frac{1}{8}(5126^2 + 4172^2 + 3248^2 + 2962^2) - \frac{1}{16}(9298^2 + 6210^2) \\
&\quad - \frac{1}{16}(8374^2 + 7134^2) + \frac{1}{32}(15508^2) \\
&= 13{,}944.5.
\end{aligned}
$$

Alternatively, we can calculate the sum of squares for the AC contrast as follows. The contrast coefficients, which are 1 if the levels of factors A and C are both high or both low, and -1 otherwise, are

$$[1, \;\; 1, \;\; -1, \;\; -1, \;\; 1, \;\; 1, \;\; -1, \;\; -1, \;\; -1, \;\; -1, \;\; 1, \;\; 1, \;\; -1, \;\; -1, \;\; 1, \;\; 1 \,].$$

Table 13.14　Analysis of variance for the decontamination alpha-particle experiment

Source of variation	Degrees of freedom	Sum of squares	Ratio	p-value
Blocks	3	28262.500	–	–
A	1	297992.000	40.25	0.0001
B	1	1444150.125	195.07	0.0001
C	1	48050.000	6.49	0.0232
D	1	700336.125	94.60	0.0001
AB	1	570846.125	77.11	0.0001
AC	1	13944.500	1.88	0.1915
AD	1	22366.125	3.02	0.1041
BC	1	15.125	0.00	0.9646
BD	1	252050.000	34.05	0.0001
CD	1	24531.125	3.31	0.0902
ABC	1	38226.125	5.16	0.0394
ABD	1	144.500	0.02	0.8909
ACD	1	66.125	0.01	0.9260
BCD	1	5618.000	0.76	0.3984
Error	14	103645.000		
Total	31	3550243.500		

Then,

$$ss(AC) = \frac{\left(\sum_i \sum_j \sum_k \sum_l c_{ijkl}\overline{y}_{.ijkl}\right)^2}{\left(\sum_i \sum_j \sum_k \sum_l c_{ijkl}^2/r\right)} = \frac{(334)^2}{16/2} = 13{,}944.5.$$

The error sum of squares is obtained by subtracting the sums of squares for the main effects and interactions from the total sum of squares, where the latter is

$$sstot = \sum_h \sum_i \sum_j \sum_k \sum_l y_{hijkl}^2 - \frac{G^2}{32}.$$

Similarly, the error degrees of freedom are obtained by subtraction.

There are fourteen hypothesis tests to be done. If each is done at a significance level 0.01, the overall level is at most $\alpha = 0.14$. At this level $F_{1,14,.01} = 8.86$, and the significant effects are the AB and BD interactions (and the main effects of A, B, and D averaged over the levels of the other factors). The hypotheses of negligible C main effect and ABC interaction would be rejected at a slightly higher significance level.

Interaction plots of the two most important interactions (AB and BD) are shown in Fig. 13.5. Figure 13.5(a) suggests that B should be set at its high level to achieve a lower radioactivity. The interaction is caused by the fact that the benefit of setting B at its high level is more marked when A is at its low level than at its high level. If B is at its high level, there is a slight preference for A to be at its low level to achieve a lower radioactivity. On the other hand, the system is more stable when A is at its high level, meaning that the radioactivity is not so sensitive to the level of B. So the choice for the setting of A is not completely obvious. Figure 13.5(b) shows a similar picture. Again B should be at its high level with a preference for D at its low level (which also produces the more stable system).

Fig. 13.5 Interaction plots for the decontamination alpha-particle experiment

(a) AB interaction   **(b)** BD interaction
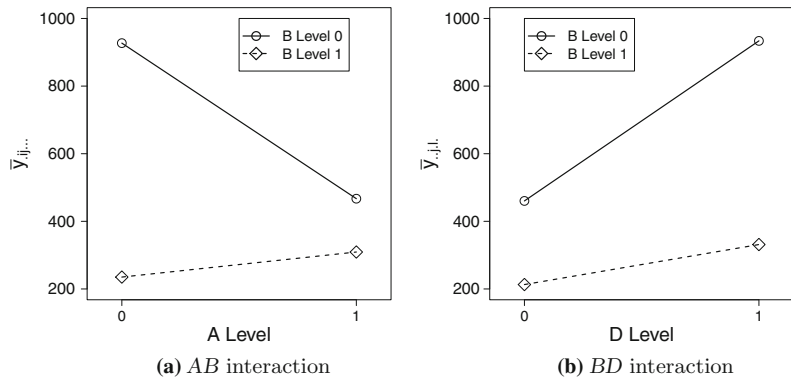
The main effect of C is not significant, but the data suggest that it should be set at its high level unless this increases the cost substantially.

Thus, the results of the analysis of variance table suggest that suitable treatment combinations for the decontamination process are 0110 or 1110. These are the same two treatment combinations that would be selected from a perusal of the data in Table 13.13. The benefit of the investigation of main effects and interactions is that it suggests directions for future experiments (raising the level of B further while keeping D fairly low and perhaps raising A a little). It also allows the chemical analysts to better understand the nature of the chemical system (see the original article of Barnett and Mead). Finally, it helps to ensure that the treatment combination that appears to be the best from the data is not just the result of error variability or a spurious observation (outlier).

13.9   Partial Confounding

Partial confounding is the term applied to a design that is formed from the combination of the blocks from different single-replicate designs with different confounding schemes. A contrast that is confounded in some replicates but not in others is said to be *partially confounded* with blocks. A partially confounded contrast can be estimated using only the data of those replicates in which it is unconfounded. Thus, the variance of the contrast estimator is inversely proportional to the number of replicates in which it is estimable.

For example, the design in Table 13.15 for a 2^3 experiment in 8 blocks of size 4 has four observations on each treatment combination. It is made up of four single-replicate designs; the first confounds the contrast from the ABC interaction, while the second confounds the AB contrast, the third confounds AC, and the fourth BC. This means that the ABC contrast is estimable from the second, third, and fourth single-replicate designs, but not the first, and the AB contrast is estimable from the first, third, and fourth single-replicate designs, but not the second. Similarly, the AC and BC contrasts are estimable from three of the four replicates. The main-effect contrasts A, B, and C are estimable from all four replicates. Consequently, all factorial treatment contrasts can be estimated, but the variance associated with each partially confounded contrast will be larger than that associated with each of the unconfounded contrasts by a factor of four-thirds. The benefit of using a partially confounded design instead of a repeated single-replicate design is that each treatment contrast is estimable, yet all totally unconfounded contrasts are still estimated with the maximum possible precision.

Table 13.15 Design and data (in parentheses) for the coil experiment

Confounded	Block	Treatment combination (response)			
ABC	I	000 (2208)	110 (2133)	101 (2459)	011 (3096)
	II	100 (2196)	010 (2086)	001 (3356)	111 (2776)
AB	III	000 (2004)	110 (2112)	001 (3073)	111 (2631)
	IV	100 (2179)	010 (2073)	101 (3474)	011 (3360)
AC	V	001 (2839)	100 (2189)	011 (3522)	110 (2095)
	VI	000 (1916)	101 (2979)	010 (2151)	111 (2500)
BC	VII	100 (2056)	000 (2010)	011 (3209)	111 (3066)
	VIII	010 (1878)	110 (2156)	001 (3423)	101 (2524)

Source From Fundamental Concepts in the Design of Experiments, Fourth Edition, by Charles R. Hicks. Copyright © 1964, 1973, 1982, 1993 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc

Example 13.9.1 `Coil experiment`

C. R. Hicks, in his textbook *Fundamental Concepts in the Design of Experiments*, describes an experiment to examine the variability of outside diameters of coils of wire. There were three treatment factors of interest, as follows:

- *A*:   Two winding machines, coded 0, 1.
- *B*:   Two wire stocks, coded 0, 1.
- *C*:   Two positions on the coil, coded 0, 1.

Only four of the $v = 8$ treatment combinations could be measured at any one time. Consequently, the experiment was divided into blocks of size $k = 4$. A total of $n = 32$ observations could be taken, and a block design of $b = 8$ blocks of size $k = 4$ was needed.

It is easily verified that no balanced incomplete block design of this size exists (since $r = 4$ and $\lambda = 12/7$). A cyclic design could have been used, but cyclic designs do not have orthogonal factorial structure in general. A partially confounded design was selected, consisting of four single-replicate designs, each confounding a different interaction (*ABC*, *AB*, *AC*, and *BC*). The design and responses are shown in Table 13.15.

We use the standard model for 3 treatment factors and one block factor with no block×treatment interaction.

$$Y_{hijk} = \mu + \theta_h + \tau_{ijk} + \epsilon_{hijk}$$
$$= \mu + \theta_h + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik}$$
$$+ (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{hijk},$$
$$\epsilon_{hijk} \sim N(0, \sigma^2) \text{ and mutually independent},$$
$$h = 1, \ldots, 8; \quad i = 0, 1; \quad j = 0, 1; \quad k = 0, 1;$$
$$(h, i, j, k) \text{ in the design.}$$

Since the main effect contrasts are not confounded in any part of the design, these can be estimated using all of the data. In addition, these contrasts are orthogonal to block contrasts, so no block adjustments are needed. The contrast for comparing the first two levels of factor *B* (averaged over the levels of *A* and *C*), for example, is then

$$\beta_1^* - \beta_0^*, \quad \text{where } \beta_j^* = \beta_j + (\overline{\alpha\beta})_{.j} + (\overline{\beta\gamma})_{j.} + (\overline{\alpha\beta\gamma})_{.j.},$$

and has least squares estimate

$$\hat{\beta}_1^* - \hat{\beta}_0^* = \bar{y}_{..1.} - \bar{y}_{..0.} = \frac{y_{..1.}}{16} - \frac{y_{..0.}}{16} = 2552.75 - 2555.31 = -2.56.$$

To four decimal places, this estimate is -2.5625. Equivalently, in terms of the treatment combinations the contrast is $\sum_i \sum_j \sum_k c_{ijk} \tau_{ijk}$, where the coefficients c_{ijk} in standard order and divisor $v/2$ are given by

$$\frac{1}{4}[-1, -1, 1, 1, -1, -1, 1, 1].$$

with least squares estimate $\sum_i \sum_j \sum_k c_{ijk} \bar{y}_{.ijk} = -2.5625$. The variance associated with this contrast is

$$\frac{\sum_i \sum_j \sum_k c_{ijk}^2 \sigma^2}{4} = \frac{8\sigma^2}{16 \times 4} = \frac{\sigma^2}{8}.$$

The test of the hypothesis $H_0 : \{\beta_1^* - \beta_0^* = 0\}$ against the alternative hypothesis $H_A : \{\beta_1^* - \beta_0^* \neq 0\}$ is similar to (4.3.15) (p. 77); that is,

$$\text{reject } H_0 \text{ if } \frac{ssc_B}{msE} = \frac{(\bar{y}_{..1.} - \bar{y}_{..0.})^2}{msE/8} > F_{1, df, \alpha/m},$$

where df is the error degrees of freedom obtained from the analysis of variance table, which is shown in Table 13.16, m is the number of hypotheses to be tested, and $ssc_B = 8(-2.5625)^2 = 52.53$. To test the equivalent hypothesis $H_0^B : \{\beta_0^* = \beta_1^*\}$ using the rules of Sect. 7.3, we obtain the same value of ssB; that is,

$$ssB = acr \sum_j \bar{y}_{..j.}^2 - abcr \bar{y}_{....}^2 = 52.53.$$

The sum of squares for each of the other main effects can be calculated in a similar fashion. The sum of squares for the interactions can be calculated similarly, except that only three of the four replicates are used. For example, the BC interaction contrast can be estimated only from the first three replicates (that is, from 24 observations, not 32). Thus, the contrast $\frac{1}{2}[(\beta\gamma)_{00}^* - (\beta\gamma)_{01}^* - (\beta\gamma)_{10}^* + (\beta\gamma)_{11}^*]$

Table 13.16 Analysis of variance for the coil experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Blocks (adj)	7	303,414.14	43,344.88	-	
Blocks (unadj)	7	439,777.72	-	-	
A	1	224,282.53	224,282.53	3.97	0.0627
B	1	52.53	52.53	0.00	0.9760
C	1	6,886,688.28	6,886,688.28	121.79	0.0001
AB (adj)	1	737.04	737.04	0.01	0.9104
AC (adj)	1	416,066.67	416,066.67	7.36	0.0148
BC (adj)	1	2,667.04	2,667.04	0.05	0.8307
ABC (adj)	1	70,742.04	70,742.04	1.25	0.2789
Error	17	961,283.11	56,546.07		
Total	31	9,002,296.97			

(where $(\beta\gamma)^*_{jk} = (\beta\gamma)_{jk} + (\overline{\alpha\beta\gamma})_{.jk}$) has least squares estimate

$$\frac{1}{2}\left[\overline{y}_{..00} - \overline{y}_{..01} - \overline{y}_{..10} + \overline{y}_{..11}\right]$$
$$= \frac{1}{2}[2080.86 - 3029.76 - 2099.38 + 3006.11] = -21.085,$$

where only *six* observations y_{hijk} are used in the calculation of $\overline{y}_{..jk}$ for each combination of B and C. (Use of 4 decimal places gives the more accurate value of -21.0833). In terms of the treatment combination effects τ_{ijk}, the contrast coefficients c_{ijk} in standard order and divisor $v/2$ are

$$\frac{1}{4}[1, \ -1, \ -1, \ 1, \ 1, \ -1, \ -1, \ 1 \ ],$$

and the contrast estimate is

$$\sum\sum\sum c_{ijk}(\overline{y}_{.ijk}) = -21.0833,$$

where only the *three* observations from the first three single-replicate designs are used in the calculation of each $\overline{y}_{.ijk}$. The variance associated with this contrast is

$$\frac{\sum_i \sum_j \sum_k c^2_{ijk}\sigma^2}{3} = \frac{8\sigma^2}{16 \times 3} = \frac{\sigma^2}{6}.$$

To test the hypothesis $H_0^{BC} : [(\beta\gamma)^*_{00} - (\beta\gamma)^*_{01} - (\beta\gamma)^*_{10} + (\beta\gamma)^*_{11} = 0]$, we reject H_0^{BC} if

$$\frac{\left(\frac{1}{2}\left[\overline{y}_{..00} - \overline{y}_{..01} - \overline{y}_{..10} + \overline{y}_{..11}\right]\right)^2}{(4/(2^2 \times 6)) \times msE} = \frac{(-21.0833)^2}{msE/6} = \frac{2667.04}{56546.07} < F_{1,17,\alpha/m}.$$

The complete analysis of variance table is shown in Table 13.16. The adjusted mean square for blocks, which is 43,344.88, is smaller than the mean square for error. Thus, blocking was not helpful in this experiment. □

13.10   Comparing the Multireplicate Designs

For some block sizes, we have a choice of possible designs for a multireplicate factorial experiment. For example, suppose that a design is required with blocks of size $k = 4$ for a factorial experiment involving 3 factors, each having two levels (so $v = 8$). Practical considerations dictate that at most $b = 14$ blocks can be used. The contrasts of interest are all of the main-effect and interaction contrasts with coefficients ± 1 as listed in Table 13.2 (p. 436). Three possible ways of designing the experiment with 14 blocks are as follows.

Design Possibility 1

The first possibility is to use a balanced incomplete block design, since one exists with $\lambda = 3$ and $r = 7$, $b = 14$, $v = 8$, $k = 4$. The design (before randomization) is given in Table 13.17, where the blocks are shown as columns. The treatment labels 1–8 of the design are randomly assigned to the $v = 8$ treatment combinations (000, 001, ..., 111) to obtain a balanced incomplete block design suitable for a factorial experiment.

Table 13.17 A balanced incomplete block design with 8 treatment labels and 14 blocks of size 4

						Blocks							
I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
1	5	1	3	1	2	1	2	1	3	1	2	1	2
2	6	2	4	3	4	4	3	2	4	3	4	4	3
3	7	7	5	6	5	6	5	5	7	5	6	5	6
4	8	8	6	8	7	7	8	6	8	7	8	8	7

If we let the effect of treatment combination ijk be denoted by τ_{ijk}, then following Sect. 11.4.3, the least squares estimator of a contrast $\sum\sum\sum c_{ijk}\tau_{ijk}$ is $\sum\sum\sum c_{ijk}\widehat{\tau}_{ijk} = \frac{k}{\lambda v}\sum\sum\sum c_{ijk}Q_{ijk}$, where $Q_{ijk} = T_{ijk} - \frac{1}{k}\sum_h n_{hijk}B_h$, and T_{ijk} is the total of the observations on treatment combination ijk, B_h is the total of the observations in the hth block, and n_{hijk} is 1 if treatment combination ijk is in block h, and otherwise n_{hijk} is 0. Taking all contrast coefficients c_{ijk} equal to $+1$ or -1, the variance of the least squares estimator, $\sum\sum\sum c_{ijk}\widehat{\tau}_{ijk}$, is

$$\text{Var}\left(\sum_i\sum_j\sum_k c_{ijk}\widehat{\tau}_{ijk}\right) = \frac{k}{\lambda v}\sum_i\sum_j\sum_k c_{ijk}^2\,\sigma^2 = \frac{8\sigma^2}{6}, \tag{13.10.3}$$

and this is the same for all main-effect and interaction contrasts for a 2^3 experiment.

Design Possibility 2

For the same experiment, suppose that the 3-factor interaction ABC is expected to be negligible and is of no interest. Then the balanced incomplete block design discussed above is not ideal, because ABC contrast is measured with the same precision as the main-effect and 2-factor interaction contrasts. Suppose, instead, we decide to confound the ABC contrast in each of seven replicates. Using the equations

$$a_1 + a_2 + a_3 = 0 \text{ or } 1 \text{ (mod 2)},$$

the following single-replicate design in two blocks would be obtained:

Block I: 000 011 101 110
Block II: 001 010 100 111

The design in $b = 14$ blocks is obtained by repeating these two blocks $r = 7$ times. Since ABC is confounded in every replicate, it is not estimable—it cannot be measured. ABC is said to be *completely confounded*. All other orthogonal contrasts (including the main-effect and 2-factor interaction contrasts) are unconfounded, so can be estimated without adjusting for blocks.

Let $\sum\sum\sum c_{ijk}\tau_{ijk}$ be a contrast (with all coefficients ± 1) measuring a 2-factor interaction or a main effect. Then its least squares estimator is $\sum\sum\sum c_{ijk}\overline{Y}_{.ijk}$, where the average is taken over the 7 replicates or repeated pairs of blocks. The corresponding variance is

$$\text{Var}\left(\sum\sum\sum c_{ijk}\widehat{\tau}_{ijk}\right) = \text{Var}\left(\sum\sum\sum c_{ijk}\overline{Y}_{.ijk}\right) = \frac{8\sigma^2}{7}. \tag{13.10.4}$$

Table 13.18 Partial confounding of *ABC*, *AB*, *AC*, *BC* contrasts in a design with 14 blocks of size 4 for 3 factors at two levels each

Replicates	Confounded	Blocks	Treatment combinations			
I–IV	*ABC*	I, III, V, VII	000	011	101	110
		II, IV, VI, VIII	001	010	100	111
V	*AB*	IX	000	001	110	111
		X	100	101	010	011
VI	*AC*	XI	000	010	101	111
		XII	001	011	100	110
VII	*BC*	XIII	000	011	100	111
		XIV	001	010	101	110

Comparing (13.10.4) with (13.10.3), we see that the effect of losing all of the information on the *ABC* contrast is to reduce the variance of all other factorial contrasts from $8\sigma^2/6$ to $8\sigma^2/7$.

Design Possibility 3

Instead of repeating the same design seven times as was done above, we could try to spread the loss of information due to confounding across several of the interaction contrasts by using partial confounding. Suppose that we take four copies of the two blocks that confound *ABC*, together with one pair of blocks that confounds *AB*, one pair that confounds *AC*, and one pair that confounds *BC*. This seven-replicate design is shown in Table 13.18.

The *ABC* contrast is confounded in replicates I–IV, but can be estimated (without block adjustments) from replicates V–VII (that is, from Blocks IX–XIV, or three pairs of blocks). The variance of the *ABC* contrast estimator is then $8\sigma^2/3$, compared with $8\sigma^2/6$ in the balanced incomplete block design—it is completely confounded in the second design.

Each 2-factor interaction contrast can be estimated without block adjustments from the six replicates, or pairs of blocks, in which it is not confounded. The variances of their least squares estimators are then $8\sigma^2/6$, the same as in the balanced incomplete block design, but worse than the value $8\sigma^2/7$ for the design with *ABC* completely confounded.

The main effects can be estimated from all seven replicates. The variances of their contrast estimators are all $8\sigma^2/7$, the same as for the second design, but better than the value of $8\sigma^2/6$ for the balanced incomplete block design.

Summary

A summary of the variances of the least squares estimators of the factorial contrasts is given in Table 13.19. No one design is the best for all seven factorial effects. The choice of design would depend upon the importance of estimating the *ABC* contrast relative to the 2-factor and main-effect contrasts.

Table 13.19 Variances of contrast estimators $\Sigma\Sigma\Sigma c_{ijk}\tau_{ijk}$ (with all coefficients ± 1) measuring a 2-factor interaction or a main effect for three design possibilities for $v = 8, r = 7, b = 14, k = 4$

Contrast	Design 1	Design 2	Design 3
	BIBD	Complete Confounding of *ABC*	Partial Confounding of *ABC*, *AB*, *AC*, *BC*
ABC	$\frac{8\sigma^2}{6}$	not estimable	$\frac{8\sigma^2}{3}$
AB, *AC*, *BC*	$\frac{8\sigma^2}{6}$	$\frac{8\sigma^2}{7}$	$\frac{8\sigma^2}{6}$
A, *B*, *C*	$\frac{8\sigma^2}{6}$	$\frac{8\sigma^2}{7}$	$\frac{8\sigma^2}{7}$

We have not exhausted all the possible designs that can be obtained by partial confounding. For example, one could confound the two-factor interactions in two pairs of blocks and the three-factor interaction in one pair of blocks, or alternatively, one could confound each of the seven factorial contrasts in turn in one pair of blocks. (This latter option gives design possibility 1.) In every case, the smaller contrast variances will coincide with the contrasts that are confounded less often, the contrast variance being $8\sigma^2/r$ if the contrast is unconfounded in r replicates.

13.11   Using SAS Software

Analyzing factorial experiments with confounding using the SAS software is straightforward for the types of designs discussed in this chapter. The SAS statements required for the analysis are the same as those outlined in Chaps. 6, 7, and 10. Using the GLM procedure, the blocking factor is listed in the MODEL statement first, so that the Type I sums of squares for factorial effects are appropriately adjusted for blocks.

Any effect that is completely confounded—including effects confounded in a single replicate design—should *not* be included in the MODEL statement. If included after the blocking factor, a completely confounded effect would show zero degrees of freedom under the Type I and Type III sums of squares. The corresponding degree of freedom would already be accounted for under block effects. Partially confounded effects, however, should be included in the model statement as illustrated in the following example.

Table 13.20 SAS program for analysis of an experiment with partial confounding—the coil experiment

```
DATA COIL;
 INPUT BLOCK A B C Y;
 LINES;
  1  0 0 0  2208
  1  1 1 0  2133
  1  1 0 1  2459
  1  0 1 1  3096
  2  1 0 0  2196
  :  : : :   :
  8  1 1 0  2156
  8  0 0 1  3423
  8  1 0 1  2524
;
PROC GLM;
 CLASS BLOCK A B C;
 MODEL Y = BLOCK A B C A*B A*C B*C A*B*C;
 ESTIMATE 'A' A -1 1;
 ESTIMATE 'B' B -1 1;
 ESTIMATE 'C' C -1 1;
 ESTIMATE 'AB' A*B 1 -1 -1 1 / DIVISOR=2;
 ESTIMATE 'AC' A*C 1 -1 -1 1 / DIVISOR=2;
 ESTIMATE 'BC' B*C 1 -1 -1 1 / DIVISOR=2;
 ESTIMATE 'ABC' A*B*C -1 1 1 -1 1 -1 -1 1 / DIVISOR=4;
```

Fig. 13.6 SAS program partial output illustrating partial confounding—the coil experiment

Example 13.11.1 `Partial confounding—Coil experiment, continued`

Table 13.20 contains a SAS program for analysis of the coil experiment data. Corresponding output is given in Fig. 13.6. The coil experiment was a four-replicate 2^3 experiment with partial confounding—each of the four interaction effects was confounded in one of the four replicates. In the GLM procedure, the blocking factor is entered into the MODEL statement first. As a result, the Type I sum of squares for blocks is unadjusted for treatment effects, whereas the Type I sums of squares for each treatment interaction effect is adjusted for block effects. The Type III sum of squares for blocks is adjusted for treatment effects, so can be used to assess the usefulness of blocking in this experiment.

The divisors used in the ESTIMATE statements of the GLM procedure cause use of divisor $v/2 = 4$ for the contrast coefficients c_{ijk} for each contrast. Thus, all contrasts would have been estimated with the same variance had there been no partial confounding. Because each interaction contrast is confounded in one of the four replicates, the variance of each interaction contrast estimator is larger than each main effect contrast estimator by a factor of four-thirds (i.e. $97.079^2/84.073^2$ $= 1.333$). □

13.12   Using R Software

Analyzing factorial experiments with confounding using the R software is straightforward for the types of designs discussed in this chapter. The R statements required for the analysis are the same as those outlined in Chaps. 6, 7, and 10. Using the linear models function `lm`, the blocking factor is entered into the model first, so that the Type I sums of squares for factorial effects are appropriately adjusted for blocks.

Any effect that is completely confounded—including effects confounded in a single replicate design—should *not* be included in the model. If included after the blocking factor, a completely confounded effect would show zero degrees of freedom under the Type I and Type III sums of squares. The corresponding degree of freedom would already be accounted for under block effects. Partially confounded effects, however, should be included in the model statement as illustrated in the following example.

Table 13.21   R program for analysis of an experiment with partial confounding—the coil experiment

```
coil.data = read.table("data/coil.txt", header=T)
head(coil.data, 3)
# Create factor variables
coil.data = within(coil.data,
   {fBlock = factor(Block); fA = factor(A);
    fB = factor(B); fC = factor(C) })

# Analysis of variance
options(contrasts = c("contr.sum", "contr.poly"))
model1 = lm(y ~ fBlock + fA*fB*fC, data=coil.data)
anova(model1)
drop1(model1, ~., test="F")


# Contrast estimates, confidence intervals, and tests
library(lsmeans)
lsmA = lsmeans(model1, ~ fA)
summary(contrast(lsmA, list(A=c(-1, 1))), infer=c(T,T))
lsmB = lsmeans(model1, ~ fB)
summary(contrast(lsmB, list(B=c(-1, 1))), infer=c(T,T))
lsmC = lsmeans(model1, ~ fC)
summary(contrast(lsmC, list(C=c(-1, 1))), infer=c(T,T))
lsmAB = lsmeans(model1, ~ fB:fA)
summary(contrast(lsmAB, list(AB=c(1,-1,-1, 1)/2)), infer=c(T,T))
lsmAC = lsmeans(model1, ~ fC:fA)
summary(contrast(lsmAC, list(AC=c(1,-1,-1, 1)/2)), infer=c(T,T))
lsmBC = lsmeans(model1, ~ fC:fB)
summary(contrast(lsmBC, list(BC=c(1,-1,-1, 1)/2)), infer=c(T,T))
lsmABC = lsmeans(model1, ~ fC:fB:fA)
summary(contrast(lsmABC, list(ABC=c(-1,1,1,-1,1,-1,-1, 1)/4)),
        infer=c(T,T))
```

Table 13.22 R program partial output illustrating partial confounding—the coil experiment

```
> model1 = lm(y ~ fBlock + fA*fB*fC, data=coil.data)
> anova(model1)

Analysis of Variance Table
Response: y
          Df   Sum Sq  Mean Sq  F value   Pr(>F)
fBlock     7   439778    62825     1.11    0.400
fA         1   224283   224283     3.97    0.063
fB         1       53       53     0.00    0.976
fC         1  6886688  6886688   121.79  3.6e-09
fA:fB      1      737      737     0.01    0.910
fA:fC      1   416067   416067     7.36    0.015
fB:fC      1     2667     2667     0.05    0.831
fA:fB:fC   1    70742    70742     1.25    0.279
Residuals 17   961283    56546

> # Contrast estimates
> library(lsmeans)
> lsmB = lsmeans(model1, ~ fB)
> summary(contrast(lsmB, list(B=c(-1, 1))), infer=c(T,T))
 contrast estimate     SE df lower.CL upper.CL t.ratio p.value
 B         -2.5625 84.073 17  -179.94   174.82   -0.03  0.9760

Results are averaged over the levels of: fBlock, fA, fC
Confidence level used: 0.95

> lsmBC = lsmeans(model1, ~ fC:fB)
> summary(contrast(lsmBC, list(BC=c(1,-1,-1, 1)/2)), infer=c(T,T))
 contrast estimate     SE df lower.CL upper.CL t.ratio p.value
 BC        -21.083 97.079 17   -225.9   183.74  -0.217  0.8307

Results are averaged over the levels of: fBlock, fA
Confidence level used: 0.95

> lsmABC = lsmeans(model1, ~ fC:fB:fA)
> summary(contrast(lsmABC, list(ABC=c(-1,1,1,-1,1,-1,-1, 1)/4)), infer=c(T,T))

 contrast estimate     SE df lower.CL upper.CL t.ratio p.value
 ABC       -108.58 97.079 17   -313.4   96.236  -1.119  0.2789

Results are averaged over the levels of: fBlock
Confidence level used: 0.95
```

Example 13.12.1 `Partial confounding—Coil experiment, continued`

Table 13.21 contains an R program for analysis of the coil experiment data. Some of the corresponding output is given in Table 13.22. The coil experiment was a four-replicate 2^3 experiment with partial confounding—each of the four interaction effects was confounded in one of the four replicates. In the second block of code in Table 13.21, the `lm` function fits the standard model involving block effects and factorial effects, saving the results as `model1`. Subsequently, the functions `anova` and `drop1` generate the type 1 and type 3 analysis of variance tables, respectively, the first of which is shown in the top of Table 13.22. The blocking factor is entered into the model first. As a result, the type 1 sum of squares for blocks is unadjusted for treatment effects, whereas the Type I sums of squares for each

Table 13.23   Projectile experiment data, confounding *ABCD*

| | Day 1 | | | Day 2 | |
Run	TC	y_{1ijkl}	Run	TC	y_{2ijkl}
1	0000	97	13	0001	75
7	0011	26	11	0010	39
5	0101	53	9	0100	68
3	0110	15	15	0111	−16
6	1001	145	10	1000	151
4	1010	100	16	1011	97
2	1100	150	14	1101	141
8	1111	54	12	1110	66

Source Johnson and Leone (1977). Copyright © 1977 Johnson and Leone. Reprinted with permission

treatment interaction effect is adjusted for block effects. The Type III sum of squares for blocks (not shown) is adjusted for treatment effects, so can be used to assess the usefulness of blocking in this experiment.

In the third block of code, for each factorial effect contrast, the summary and contrast functions of lsmeans are coupled to compute the contrast estimate, standard error, *t*-test, and 95% confidence interval. Since each main effect contrast is averaged over the four combinations of two other factors, specifying a pair of coefficients ± 1 yields contrast coefficients c_{ijk} with divisor $v/2 = 4$. Likewise, each two-factor interaction contrast is averaged over the two levels of the third factor, so specifying four coefficients $\pm 1/2$ again yields contrast coefficients c_{ijk} with divisor $v/2 = 4$. The three factor interaction contrast coefficients c_{ijk} have divisor $v/2 = 4$ as directly specified. Output for the main effect of *B*, the *BC* interaction, and the *ABC* interaction contrasts are shown in the bottom of Table 13.22.

Since all contrasts considered here have contrast coefficients $c_{ijk} = \pm 1/4$, all contrasts would have been estimated with the same variance had there been no partial confounding. Because each interaction contrast is confounded in one of the four replicates, the variance of each interaction contrast estimator is larger than that of each main effect contrast estimator by a factor of four-thirds. Correspondingly, the squares of the standard errors are in this ratio.   □

Exercises

1. Construct a single-replicate 2^3 design confounding *AB* with blocks. In other words, list the treatment combinations block by block.

2. Construct a single replicate 2^5 design confounding *ABC* and *CDE*. Determine the other effect that is confounded.

3. **Projectile experiment**

   N.L. Johnson and F.C. Leone, in their 1977 book *Statistics and Experimental Design in Engineering and the Physical Sciences*, described a single-replicate 2^4 experiment concerning the performance of a new rifle under test. Under study were the effects on projectile velocity of the factors charge weight (*A*), projectile weight (*B*), propellant web (*C*), and weapon (*D*), where two rifles were used. The design included two blocks each of size eight, corresponding to the two days on which data were collected, confounding *ABCD*. The coded velocity data are given in Table 13.23.

(a) Fit a model including block effects, treatment main effects, and 2-factor interactions. Use residual plots to check the standard model assumptions.

(b) Conduct the analysis of variance, and discuss the results.

(c) Construct simultaneous confidence intervals for any interesting treatment contrasts using an appropriate method of multiple comparisons.

(d) Reanalyze the data using the Voss–Wang method, including all estimable treatment effects in the analysis.

4. **Field experiment, continued**

(a) For the field experiment of Example 13.3.1, p. 437, verify that the sum of squares and the contrast estimate for *BD* are as shown in Table 13.5, p. 438.

(b) Draw the *BD* interaction plot. Does this plot also suggest that *B* and *D* should be at their low levels?

(c) Suppose that the experimenters had expected all of the 3-factor interactions to be negligible and had omitted the corresponding terms from the model (instead of those involving *AD*). Reanalyze the experiment accordingly. What would have been concluded?

(d) Apply the Voss–Wang method to analyze the data of the field experiment. Relevant information is given in Table 13.5, p. 438.

5. Suggest a confounding scheme for a 2^6 experiment in 8 blocks of 8, assuming that all 2-factor interactions are to be estimated, as are the 3-factor interactions involving both *A* and *F*. List all effects confounded. List the treatment combinations in the design block by block.

6. Suggest a confounding scheme for a 2^8 experiment in 16 blocks of 16, assuming that all 2-factor and 3-factor interactions are to be estimated. List all effects confounded. List the treatment combinations in Block I and in two other blocks.

7. **Mangold experiment, continued**

(a) For the mangold experiment of Sect. 13.5, verify that the sum of squares and the contrast estimate for *CD* are as shown in Table 13.12, p. 449.

(b) Draw the *CD* interaction plot. Does this plot agree with the factor levels suggested in Sect. 13.5 for increasing the yield?

(c) Check that the assumption of normality of the error variables is satisfied. Also check that the variances of the errors appear to be equal for each level of the four factors.

(d) Draw a half-normal probability plot of all of the contrast estimates (including the higher-order interactions). Does it appear that the experimenters made the correct assumptions of negligible higher-order interactions?

8. **Decontamination experiment—Beta particles**

An experiment was described by M. K. Barnett and F. C. Mead, Jr. in the journal *Applied Statistics* in 1956 to explore the effect of four factors on the efficiency of a decontamination process for the removal of radioactive isotopes from liquid waste. The measurements taken after the decontamination process were the counts per minute per milliliter of alpha and beta particles. Data for the alpha particles and further description of the experiment were given in Sect. 13.8.1, p. 452. We consider here part of the data for the beta particles, shown in Table 13.24. The four treatment factors were:

Table 13.24 Randomized design for a 2^4 experiment in 2 blocks of size 8, confounding *ABCD*. Data for the decontamination beta-particle experiment are shown in parentheses.

Block	Treatment combinations (response)							
I	1010	1111	0110	0000	1100	0101	0011	1001
	(716)	(686)	(498)	(1437)	(527)	(579)	(1433)	(906)
II	0010	0001	0111	1000	1101	0100	1110	1011
	(1024)	(1364)	(475)	(574)	(664)	(579)	(507)	(1130)

Source Barnett and Mead (1956). Copyright © 1956 Blackwell Publishers. Reprinted with permission

<div style="text-align:center">

A : 0.4 g and 2.5 g per liter of aluminum sulphate (coded 0, 1);

B : 0.4 g and 2.5 g per liter of barium chloride (coded 0, 1);

C : 0.08 g and 0.4 g per liter of carbon (coded 0, 1);

D : Final pH of liquid waste (6 and 10, coded 0, 1).

</div>

The experimenters selected a design in $b = 2$ blocks of $k = 8$ that confounded the four-factor interaction contrast *ABCD*. The fitted model included all main-effects and all 2-factor and 3-factor interactions. The block×treatment interaction was assumed to be negligible.

(a) Use a half-normal probability plot to identify the important contrasts.
(b) Use the method of Voss–Wang to check your selection in part (a).
(c) Draw an interaction plot of any interaction that appears to be nonnegligible by either analysis.
(d) Looking at the results of your analysis, which settings of the factors would you recommend for reducing the beta particle counts?
(e) Suppose that the experimenters had believed before the experiment that the three-factor interactions were all negligible. What would the analysis of variance table have looked like? Would your recommendations have been any different?

9. **Penicillin experiment**

An experiment is described in Example 9.2 of the book *Design and Analysis of Industrial Experiments* edited by O. L. Davies that investigates the effects of various factors on the yield of penicillin in surface culture experiments. The five factors of interest were added to the nutrient medium, which was inoculated with a spore suspension of *P. Chrysogenum*. The spores rise to the surface, causing the growth of mycelium accompanied by the formation of penicillin. The factors and their levels were corn steep liquor (factor *A*, 2% and 3% strength), lactose (factor *B*, 2% and 3% strength), precursor (factor *C*, 0% and 0.05%), sodium nitrate (factor *D*, 0% and 0.3%), and glucose (factor *E*, 0% and 0.5%). Only 16 of the 32 treatment combinations could be carried out at one time, and the experimenters decided to observe 16 treatment combinations in one week and the remaining 16 in the following week. Large week-to-week variations were known to exist, and therefore the experiment was designed as a block design with two blocks, confounding the 5-factor interaction *ABCDE*. The observed yields of penicillin are shown in Table 13.25. Prior to the experiment, it was believed that all that all 3- and 4-factor interactions would be negligible, and also that the *CE* interaction would be important.

(a) Analyze the data, assuming that all 3- and 4-factor interactions are negligible. Do not forget to check the assumptions on the model.

Table 13.25 Data for the penicillin experiment

Block I		Block II	
Treatment combination	Yield	Treatment combination	Yield
00000	142	00001	106
00011	101	00010	148
00101	113	00100	185
00110	200	00111	130
01001	88	01000	129
01010	146	01011	140
01100	200	01101	166
01111	145	01110	215
10001	106	10000	114
10010	108	10011	114
10100	162	10101	88
10111	83	10110	164
11000	109	11001	98
11011	72	11010	195
11101	79	11100	172
11110	118	11111	110

Source Data adapted from *The Design and Analysis of Industrial Experiments* Second edition, 1979. Editor O. L. Davies. Published by Longman Group Ltd

   (b) The experimenters decided to use logarithms of the data. Does your assumption check in part (a) confirm that this should be done? If so, reanalyze the data and state your conclusions.

   (c) Using the logarithms of the data, draw a half-normal probability plot of the contrast estimates without using any knowledge that the higher-order interactions are likely to be negligible. Do your conclusions remain the same? Which analysis do you prefer? Why?

10. **Peas experiment**

The following experiment was run at Biggelswade, in England, and reported by F. Yates in his 1935 paper *Complex Experiments*. The three treatment factors were the standard fertilizers, nitrogen, phosphate, and potassium (factors N, P, and K) each at two levels. The experimental area was divided into $b = 6$ blocks of 1/70 of an acre. Each block was large enough for four plots on which a certain variety of pea was sown, and the fertilizer combinations shown in Table 13.26 were added. The design consists of three identical single-replicate designs each of which confounds the 3-factor interaction NPK. Each block has been separately randomized.

   (a) Estimate the treatment contrasts for all main effects and interactions.

   (b) Calculate the analysis of variance table for this experiment and test all relevant hypotheses. State the overall significance level.

   (c) Draw interaction plots for any important interactions. Give a set of 95% confidence intervals for the main-effect contrasts, if appropriate.

   (d) State your overall recommendations about the fertilizers in this experiment. Would you recommend a followup experiment? If so, what would you investigate?

11. **Field experiment, continued**
The field experiment was described in Example 13.3.1, p. 437. There were four treatment factors

Table 13.26 Data for the peas experiment

Block	Treatment combinations (Yield)		Block	Treatment combinations (Yield)	
I	011 (49.5)	000 (46.8)	II	100 (62.0)	001 (45.5)
	110 (62.8)	101 (57.0)		111 (48.8)	010 (44.2)
III	100 (59.8)	001 (55.5)	IV	110 (52.0)	101 (49.8)
	111 (58.5)	010 (56.0)		000 (51.5)	011 (48.8)
V	010 (62.8)	100 (69.5)	VI	101 (57.2)	011 (53.2)
	111 (55.8)	001 (55.0)		110 (59.0)	000 (56.0)

Source Yates (1935). Copyright © 1935 Blackwell Publishers. Reprinted with permission

Table 13.27 Data for the field experiment, by block and treatment combination (TC)

Block I		Block II		Block III		Block IV	
TC	y_{1ijkl}	TC	y_{2ijkl}	TC	y_{3ijkl}	TC	y_{4ijkl}
0000	58	0001	55	0000	57	0001	50
0011	51	0010	45	0011	56	0010	39
0101	44	0100	42	0101	43	0100	47
0110	50	0111	36	0110	39	0111	43
1001	43	1000	53	1001	52	1000	42
1010	50	1011	55	1010	52	1011	44
1100	41	1101	41	1100	42	1101	34
1111	44	1110	48	1111	54	1110	52

Source Experimental Designs, Second Edition, by W.G. Cochran and G.M. Cox, Copyright © 1957, John Wiley & Sons, New York. Adapted by permission of John Wiley & Sons, Inc

(A, B, C, and D) at two levels each, and the $v = 16$ treatment combinations were observed twice. Each of the $r = 2$ sets of treatment combinations were divided into blocks of size 8. The first two blocks, which confounded the *ABCD* interaction, were shown in Table 13.4, p. 437. The complete design, which is shown in Table 13.27, consisted of two such single-replicate designs.

(a) Calculate the analysis of variance table for this experiment. Now that $r = 2$, there is an estimate for error variability. Test any hypotheses of interest. Are the results similar to those obtained from the first two blocks only?

(b) Draw any interaction plots of interest. If the yield is to be increased, what recommendations would you make about the levels of the factors?

12. Construct a four-replicate 2^3 design in eight blocks of size four, partially confounding each interaction effect. Compare the variance of each interaction contrast with that of each main effect, using divisor $v/2 = 4$ for each contrast.

13. **Catalytic reaction experiment**

J.R. Bainbridge, in his 1951 article in the journal *Industrial and Engineering Chemistry*, described a factorial experiment conducted at a small plant carrying out a catalytic gaseous synthesis reaction to remove the product as a liquid solution. A 2-replicate 2^3 experiment was conducted to study the effects of converter reaction temperature (factor A), throughput rate through the converter (factor B), and the concentration of the active ingredient in the makeup gas (factor C) on each of several response variables, including the strength of the product solution (y_{hijk}). The design was composed of four blocks of size four, with the *ABC* interaction completely confounded. The design

Table 13.28 Data for the catalytic reaction experiment

Run	Block	TC	y_{hijk}	Run	Block	TC	y_{hijk}
1	1	011	89.5	9	3	010	86.2
2	1	101	84.2	10	3	100	81.8
3	1	110	85.2	11	3	001	90.4
4	1	000	89.9	12	3	111	83.6
5	2	010	85.1	13	4	110	75.3
6	2	111	83.5	14	4	000	84.6
7	2	001	90.8	15	4	011	86.7
8	2	100	81.8	16	4	101	82.2

Source Reprinted with permission from Bainbridge (1951). Copyright © 1951 *American Chemical Society*

and data are provided in Table 13.28, including the run order. (The observations in Table 13.28 are "uncoded," each value being 80 plus one-tenth the coded value given by Bainbridge.)

(a) Based on the run order, discuss how the design was probably randomized.
(b) Fit an appropriate model, and use residual plots to check the standard model assumptions.
(c) Conduct the analysis of variance, and discuss the results.
(d) Using a simultaneous confidence level of 95% for all six factorial effects, construct confidence intervals for those effects found to be significant in the analysis of variance.

14. **Catalytic reaction experiment, continued**

In the experiment described in Exercise 13, the covariate "makeup gas purity" was measured. The covariate values were 17, 12, 10, 10, 13, 14, 10, 16, 12, 13, 13, 11, 16, 11, 12, and 11, corresponding to runs 1–16, respectively. Repeat Exercise 13, but for an analysis of covariance.

15. **Design of a follow-up experiment**

An experiment was run in 2007 by Joanne Sklodowski, Josh Svenson, Adam Dallas, Tim Degenero and Paul Cotellesso to examine the compressive strength of various mortar mixes. They examined the effects of four factors: Amount of water (Factor A, 0.75lb and 0.85lb), Sand type (Factor B, play sand and medium grain sand), Temperature of water (Factor C, 58 and 96 °C), Cure time (Factor D, 4 and 6 days). The type and age of cement and the mixing time were held fixed throughout the experiment.
The ingredients were mixed and poured into a cylindrical mold. After the allotted curing time, the cylinder was crushed on a compression machine. The maximum pressure exerted before the cylinder failed was recorded in pounds per square inch (psi).

(a) The analysis of the experiment showed large effects of BCD, B, C, AB. Design a follow-up experiment to examine the interactions AB, BC, BD, CD and BCD, as well as all main effects. You can only take $r = 1$ observation on each treatment combination and you need to run the experiment in 4 blocks of 4. Write out two of the four blocks of your design and state how to find the other two.
(b) Write down a model for this experiment. Write out the "Degrees of Freedom" columns of the analysis of variance table.

16. **Design of an experiment**

   (a) Design an experiment with five factors A, B, C, D, E, each having two levels, in 4 blocks of 8 and $r = 1$ observation per treatment combination. Make sure that you can estimate all main effects, all two-factor interactions, as well as all three-factor interactions that involve *both D and E*.

   (b) Write out the 8 treatment combinations in Block I, and indicate how to find the treatment combinations in the other blocks. Illustrate this with two treatment combinations in the second block.

   (c) Write out the degrees of freedom column for the analysis of variance table corresponding to the model that you would fit.

Table 13.29 Confounding schemes for 2^p experiments in $b = 2^s$ blocks of size $k = 2^{p-s}$. For each design, s independent generators are underlined, and s corresponding equations are given. To obtain Block I of a design, list all k combinations of the first a_i's shown, then use the equations modulo 2 to complete each treatment combination

2^p	b	k	Confounded Contrasts	Block I
2^3	2	4	$\underline{ABC}$	a_1, a_2
				$a_3 = a_1 + a_2$
2^4	2	8	$\underline{ABCD}$	a_1, a_2, a_3
				$a_4 = a_1 + a_2 + a_3$
2^4	4	4	$\underline{AC}$, $\underline{ABD}$, BCD	a_1, a_2
				$a_3 = a_1$
				$a_4 = a_1 + a_2$
2^5	2	16	$\underline{ABCDE}$	a_1, a_2, a_3, a_4
				$a_5 = a_1 + a_2 + a_3 + a_4$
2^5	4	8	$\underline{ABCD}$, $\underline{ABE}$, CDE	a_1, a_2, a_3
				$a_4 = a_1 + a_2 + a_3$
				$a_5 = a_1 + a_2$
2^5	8	4	$\underline{AC}$, $\underline{BD}$, $ABCD$,	a_1, a_2
			$\underline{ABE}$, BCE, ADE, CDE	$a_3 = a_1$
				$a_4 = a_2$
				$a_5 = a_1 + a_2$
2^6	2	32	$\underline{ABCDEF}$	a_1, a_2, a_3, a_4, a_5
				$a_6 = a_1 + a_2 + a_3 + a_4 + a_5$
2^6	4	16	$\underline{ABCD}$, $\underline{CDEF}$, $ABEF$	$a_1, a_2, a_3,\ a_5$
				$a_4 = a_1 + a_2 + a_3$
				$a_6 = a_3 + a_4 + a_5$
2^6	8	8	$\underline{BCD}$, $\underline{ABE}$, $ACDE$,	a_1, a_2, a_3
			$\underline{ABCF}$, ADF, CEF, $BDEF$	$a_4 = a_2 + a_3$
				$a_5 = a_1 + a_2$
				$a_6 = a_1 + a_2 + a_3$
2^7	4	32	$\underline{ABCDE}$, $\underline{ABFG}$, $CDEFG$	$a_1, a_2, a_3, a_4,\ a_6$
				$a_5 = a_1 + a_2 + a_3 + a_4$
				$a_7 = a_1 + a_2 + a_6$
2^7	8	16	$\underline{ABCD}$, $\underline{CDEF}$, $ABEF$,	$a_1, a_2, a_3,\ a_5$
			$\underline{ACEG}$, $BDEG$, $ADFG$, $BCFG$	$a_4 = a_1 + a_2 + a_3$
				$a_6 = a_3 + a_4 + a_5$
				$a_7 = a_1 + a_3 + a_5$
2^7	16	8	$\underline{ABC}$, $\underline{CDE}$, $ABDE$,	$a_1, a_2,\ a_4$
			$\underline{BDF}$, $ACDF$, $BCEF$, AEF,	$a_3 = a_1 + a_2$
			$\underline{ADG}$, $BCDG$, $ACEG$, BEG,	$a_5 = a_3 + a_4$
			$ABFG$, CFG, $ABCDEFG$, $DEFG$	$a_6 = a_2 + a_4$
				$a_7 = a_1 + a_4$

14

14.1 Introduction

In Chap. 13, incomplete block designs for 2^p factorial experiments were obtained by confounding one or more interaction contrasts with block contrasts. In this chapter, we extend the idea of confounding to encompass experiments in which some or all factors have more than two levels. We will code the levels of an m-level factor as $0, 1, \ldots, m-1$.

In Sect. 14.2, we consider single-replicate 3^p experiments arranged in $b = 3^s$ blocks of size $k = 3^{p-s}$. The techniques used in designing these types of experiment can be adapted for m^p experiments in m^s blocks of size m^{p-s} where m is a prime number.

Pseudofactors are introduced in Sect. 14.3 to facilitate confounding in symmetric 4^p experiments and asymmetric $2^p 4^q$ experiments. Then, in Sect. 14.4, we consider asymmetric experiments involving factors or pseudofactors at both two and three levels, allowing us to look at more complicated situations where the treatment factors have a mixture of 2, 3, 4, and 6 levels.

Analysis of a two-replicate 3^3 experiment with partial confounding is illustrated using the SAS and R software packages in Sects. 14.5 and 14.6, respectively.

14.2 Confounding with Factors at Three Levels

14.2.1 Contrasts

In a factorial experiment where all treatment factors have 3 levels, each main effect has 2 degrees of freedom associated with it, each two-factor interaction has $2 \times 2 = 4$ degrees of freedom, etc. (see rule 3 of Sect. 7.3). Therefore, we can find 2 orthogonal contrasts to measure each main effect, 4 orthogonal contrasts to measure each two-factor interaction, and so on.

In a 3^2 experiment, for example, two orthogonal contrasts measuring the main effect of each of factors A and B are the linear and quadratic trend contrasts. Similarly, four orthogonal trend contrasts $A_L B_L$, $A_L B_Q$, $A_Q B_L$, and $A_Q B_Q$ (see Sect. 6.3) measuring the interaction are reproduced in Table 14.1. A different set of four orthogonal contrasts, labeled in pairs as $(AB; A^2 B^2)$ and $(AB^2; A^2 B)$, is also shown in Table 14.1. Although this second set of contrasts is less useful than the set of trend contrasts in measuring details of the interaction for quantitative factors, it will prove extremely useful for confounding purposes. The reader is asked to verify that *any* contrasts that measure the main effects of A and B are orthogonal to all the contrasts in Table 14.1 measuring the interaction.

Table 14.1 Sets of orthogonal contrasts measuring the interaction in a 3^2 experiment

TC	$A_L B_L$	$A_L B_Q$	$A_Q B_L$	$A_Q B_Q$	$(AB; A^2B^2)$		$(AB^2; A^2B)$	
00	1	-1	-1	1	-1	1	-1	1
01	0	2	0	-2	0	-2	1	1
02	-1	-1	1	1	1	1	0	-2
10	0	0	2	-2	0	-2	0	-2
11	0	0	0	4	1	1	-1	1
12	0	0	-2	-2	-1	1	1	1
20	-1	1	-1	1	1	1	1	1
21	0	-2	0	-2	-1	1	0	-2
22	1	1	1	1	0	-2	-1	1

Table 14.2 Groups of treatment combinations corresponding to orthogonal interaction contrasts in a 3^2 experiment

$(AB; A^2B^2)$			$(AB^2; A^2B)$		
00*	01[†]	02[+]	00*	01[†]	02[+]
10[†]	11[+]	12*	10[+]	11*	12[†]
20[+]	21*	22[†]	20[†]	21[+]	22*

Notice that the pair of contrasts labeled $(AB; A^2B^2)$ are two orthogonal contrasts that compare the three groups of treatment combinations (00, 12, 21) and (01, 10, 22) and (02, 11, 20). Any linear combination of this pair of contrasts is also a contrast between these three groups of treatment combinations. We have illustrated these groups of treatment combinations in the left-hand side of Table 14.2, where treatment combinations with the same superscript are in the same group. Notice that each group contains one treatment combination from each row and each column, making sure that each level of each factor is represented once in each group.

Similarly, the pair of contrasts labeled $(AB^2; A^2B)$ comprise two orthogonal contrasts that compare the three groups of treatment combinations (00, 11, 22) and (01, 12, 20) and (02, 10, 21). Any linear combination of this pair of contrasts is also a contrast between these three groups of treatment combinations. The groups are illustrated in the right-hand side of Table 14.2 and also have the property that each group contains one treatment combination from each row and each column.

The reason for the labeling $(AB; A^2B^2)$ and $(AB^2; A^2B)$ is to match the contrasts with the equation method of confounding in Sect. 14.2.3. The contrast names themselves have little meaning, except to acknowledge that each contrast belongs to the AB interaction and, as will be seen, each pair corresponds to a set of equations that partitions the treatment combinations into the three groups represented in Table 14.2.

Many texts list only one of the two labels in each pair, since each is the square of the other. For example, when the exponents are reduced modulo 3, then $A^2B = (AB^2)^2$. The convention is then to list AB^2 rather than A^2B, for example, since the leading exponent is one. However, we will list both labels to aid in identifying a complete set of confounded contrasts in designs with more than three blocks.

14.2.2   Confounding Using Contrasts

In this section we consider the division of treatment combinations into blocks by deliberately confounding negligible contrasts, as in Sect. 13.3.2 for 2^p experiments. For 3^p experiments, we look at designs with 3^s blocks of size 3^{p-s}, starting with 3 blocks of size 3^{p-1}. For a design with $b = 3$ blocks,

two degrees of freedom are used to measure the block differences. Therefore, in a single-replicate design, we must confound a pair of treatment contrasts.

As a simple example, we start with an experiment with two factors A and B in which the interaction is known to be negligible. We will attempt to use two of the interaction contrasts shown in Table 14.1 to divide the treatment combinations into 3 blocks. A pair of trend contrasts, such as $A_L B_Q$ and $A_Q B_Q$ cannot be used to give blocks of equal size, since the values of the coefficients do not fall into 3 groups of 3. However, the pair of contrasts labeled $(AB; A^2B^2)$ have three pairs of coefficients $(-1, \ 1)$, $(0, -2)$, and $(1, \ 1)$ each of which appear three times. If we use these as a guide to dividing the treatment combinations into blocks, we obtain the design in Table 14.3.

Any contrast that is orthogonal to the two confounded contrasts can be estimated without requiring block adjustments. Estimable contrasts include all contrasts measuring the main effects of A and B and the remaining two interaction contrasts labeled $(A^2B; AB^2)$ and linear combinations of these. The trend contrasts in Table 14.1 are not orthogonal to any of the AB, A^2B^2, AB^2, A^2B contrasts, so they do not fall into either the confounded or the estimable category. They are *partly confounded*. In general, interaction trend contrasts can be estimated completely only when no contrasts from the interaction are confounded.

In the present example, the interaction has four degrees of freedom. Two are used to measure blocks. The other two correspond to two estimable contrasts, which are negligible and provide two degrees of freedom to measure σ^2.

If the contrasts labeled $(A^2B; AB^2)$ in Table 14.1 were used instead of the contrasts labeled $(AB; A^2B^2)$ to provide three blocks, the design of Table 14.4 would result. This has the same properties as the design in Table 14.3 in that all main-effect contrasts are estimable and there are two estimable contrasts $(AB; A^2B^2)$ remaining in the interaction. Neither design is better than the other, and a choice can be made at random. Block design randomization should be carried out before the design is used.

As we saw in 2^p experiments, there is a correspondence between the contrasts used for confounding, the contrast names, and the equation method of confounding. In the next section we show how to obtain the design of Table 14.3 by the equation method.

14.2.3   Confounding Using Equations

3^p Experiments in Three Blocks

The design in Table 14.3, which was obtained by confounding the two interaction contrasts labeled $(AB; A^2B^2)$ in Table 14.1, can be obtained by an equation method similar to that of Sect. 13.4. Notice that in Block I the digits of the three treatment combinations add to 0 or 3. In Block II they add to 1 or

Table 14.3 3^2 experiment in 3 blocks of 3, confounding $(AB; A^2B^2)$

Block	Contrast coefficients	Treatment combinations		
I	$(-1, \ 1)$	00	12	21
II	$(0, -2)$	01	10	22
III	$(1, \ 1)$	02	11	20

Table 14.4 3^2 experiment in 3 blocks of 3, confounding $(A^2B; AB^2)$

Block	Contrast coefficients	Treatment combinations		
I	$(-1, \ 1)$	00	11	22
II	$(1, \ 1)$	01	12	20
III	$(0, -2)$	02	10	21

4, and in Block III they add to 2. Now that both factors have three levels, we work modulo 3, which means that we subtract 3 from the sum of the digits until we obtain one of 0, 1, or 2, or equivalently, we take the remainder on division by 3. Writing the treatment combinations as $a_1 a_2$, the blocks can be defined by the confounding equations

$$
\begin{aligned}
&\text{Block I:}\quad \text{Treatment combinations with } L = a_1 + a_2 = 0 \ (\text{mod } 3)\,, \\
&\text{Block II:}\quad \text{Treatment combinations with } L = a_1 + a_2 = 1 \ (\text{mod } 3)\,, \\
&\text{Block III:}\quad \text{Treatment combinations with } L = a_1 + a_2 = 2 \ (\text{mod } 3)\,.
\end{aligned}
$$

Equivalently, the same three blocks can be obtained if the equations are multiplied by 2; that is,

$$
\begin{aligned}
&\text{Block I:}\quad \text{Treatment combinations with } 2L = 2a_1 + 2a_2 = 0 \ (\text{mod } 3)\,, \\
&\text{Block II:}\quad \text{Treatment combinations with } 2L = 2a_1 + 2a_2 = 2 \ (\text{mod } 3)\,, \\
&\text{Block III:}\quad \text{Treatment combinations with } 2L = 2a_1 + 2a_2 = 1 \ (\text{mod } 3)\,.
\end{aligned}
$$

Thus, if the contrasts labeled $(AB; A^2B^2)$ in Table 14.1 are confounded with blocks, the treatment combinations in the three blocks satisfy both

$$
L = a_1 + a_2 = 0, \ 1, \text{ or } 2 \ (\text{mod } 3),
$$

and

$$
2L = 2a_1 + 2a_2 = 0, \ 2, \text{ or } 1 \ (\text{mod } 3).
$$

Alternatively, if the contrasts labeled $(AB^2; A^2B)$ are to be confounded, the equations

$$
L = a_1 + 2a_2 = 0, \ 1, \text{ or } 2 \ (\text{mod } 3)
$$

and, multiplying by 2,

$$
2L = 2a_1 + a_2 = 0, \ 2, \text{ or } 1 \ (\text{mod } 3)
$$

will produce the design in Table 14.4. Notice that the coefficients in the confounding equations correspond to the exponents in the contrast names. A set of equations defines the same set of blocks when it is multiplied by 2. Therefore, the confounded contrast names always come in pairs—one name being the square of the other—$(AB^2)^2 = A^2B^4 = A^2B$, reducing exponents (mod 3).

In general, in a 3^p experiment, if the equations

$$
L = z_1 a_1 + z_2 a_2 + \cdots + z_p a_p = 0, \ 1, \text{ or } 2 \ (\text{mod } 3)
$$

are used to produce three blocks, two contrasts will be confounded that can be labeled $(A^{z_1}B^{z_2} \cdots P^{z_p}; A^{2z_1}B^{2z_2} \cdots P^{2z_p})$, where z_i is 1 or 2 if the factor is present in the interaction, and 0 if it is not, and where the exponent is reduced modulo 3. For example, in a 3^5 experiment, the equations

$$
L = a_1 + 2a_2 + a_4 = 0, \ 1, \text{ or } 2 \ (\text{mod } 3)
$$

will give 3 blocks of size 3^4 confounding AB^2D and $A^2B^4D^2 = A^2BD^2$, which represent two contrasts from the three-factor interaction ABD. It is rarely of importance to identify exactly what the contrasts look like (they are any pair of orthogonal contrasts between the groups of treatment combinations in the

three blocks). What is important is the knowledge that the confounded contrasts belong to a particular interaction and, therefore, that all other main-effect and interaction contrasts are estimable.

3^p Experiments in Nine Blocks

The equation method of confounding can be used to produce $b = 9 = 3^2$ blocks of size 3^{p-2} in a 3^p experiment by selecting two pairs of contrasts to be confounded. If the pair $(A^{z_1}B^{z_2} \cdots P^{z_p}; A^{2z_1}B^{2z_2} \cdots P^{2z_p})$ is chosen for confounding together with the pair $(A^{y_1}B^{y_2} \cdots P^{y_p}; A^{2y_1}B^{2y_2} \cdots P^{2y_p})$, the $b = 9$ blocks are produced from the nine possible pairs of values of the two equations

$$L_1 = z_1 a_1 + z_2 a_2 + \cdots + z_p a_p = 0, \ 1, \text{ or } 2 \ (\text{mod } 3),$$

$$L_2 = y_1 a_1 + y_2 a_2 + \cdots + y_p a_p = 0, \ 1, \text{ or } 2 \ (\text{mod } 3).$$

The $b - 1 = 8$ confounded contrasts are the two pairs originally chosen, together with all possible products. This is most conveniently set out as a table. The selected pairs of contrasts are written in the first row and first column. The table is then filled out by multiplication, and the exponents are reduced modulo 3, as follows:

$$A^{y_1}B^{y_2} \cdots P^{y_p} \qquad\qquad A^{2y_1}B^{2y_2} \cdots P^{2y_p}$$

$$A^{z_1}B^{z_2} \cdots P^{z_p} \quad A^{z_1+y_1}B^{z_2+y_2} \cdots P^{z_p+y_p} \quad A^{z_1+2y_1}B^{z_2+2y_2} \cdots P^{z_p+2y_p}$$

$$A^{2z_1}B^{2z_2} \cdots P^{2z_p} \quad A^{2z_1+y_1}B^{2z_2+y_2} \cdots P^{2z_p+y_p} \quad A^{2z_1+2y_1}B^{2z_2+2y_2} \cdots P^{2z_p+2y_p}$$

If $b = 3^s$ blocks of size 3^{p-s} are required, then s independent pairs of contrast names need to be chosen for confounding. All possible products determine the entire set of $b - 1 = 3^s - 1$ confounded contrasts.

Example 14.2.1 3^4 experiment in 9 blocks of size 9

Suppose that a 3^4 experiment, with factors A, B, C, D, is to be run in $b = 9$ blocks of size 9. Further, suppose that the only interactions thought to be important are the 2-factor interactions and, therefore, these should not be confounded. Now $b = 3^2$ blocks are required, so 2 pairs of contrasts should be chosen for confounding. The $ABCD$ interaction has 16 degrees of freedom, so we can find 16 orthogonal contrasts and label them in pairs as

$$\begin{aligned}
(ABCD; A^2B^2C^2D^2), \quad & (AB^2CD; A^2BC^2D^2), \\
(ABCD^2; A^2B^2C^2D), \quad & (AB^2CD^2; A^2BC^2D), \\
(ABC^2D; A^2B^2CD^2), \quad & (AB^2C^2D; A^2BCD^2), \\
(ABC^2D^2; A^2B^2CD), \quad & (AB^2C^2D^2; A^2BCD).
\end{aligned}$$

Selecting two pairs of contrasts from the 4-factor interaction for confounding contrasts is not a good choice. For example, if $(ABCD^2; A^2B^2C^2D)$ and $(ABCD; A^2B^2C^2D^2)$ were chosen, the set of eight confounded degrees of freedom would be

$$\begin{array}{ccc}
 & ABCD^2 & A^2B^2C^2D \\
ABCD & A^2B^2C^2 & D^2 \\
A^2B^2C^2D^2 & D & ABC
\end{array}$$

Table 14.5 3^4 experiment in 3^2 blocks of 9; confounding $(ABD; A^2B^2D^2)$, $(BCD^2; B^2C^2D)$, $(AB^2C; A^2BC^2)$, and $(AC^2D^2; A^2CD)$

Block	L_1, L_2	Treatment combinations
I	0,0	0000 0112 0221 1022 1101 1210 2011 2120 2202
II	1,2	0001 0110 0222 1020 1102 1211 2012 2121 2200
III	2,1	0002 0111 0220 1021 1100 1212 2010 2122 2201
IV	0,1	0010 0122 0201 1002 1111 1220 2021 2100 2212
V	1,0	0011 0120 0202 1000 1112 1221 2022 2101 2210
VI	2,2	0012 0121 0200 1001 1110 1222 2020 2102 2211
VII	1,1	0100 0212 0021 1122 1201 1010 2111 2220 2002
VIII	2,0	0101 0210 0022 1120 1202 1011 2112 2221 2000
IX	0,2	0102 0211 0020 1121 1200 1012 2110 2222 2001

and we can see that two orthogonal contrasts in the main-effect D would also be confounded. All possible selections of two pairs of contrasts from the $ABCD$ interaction will confound either a main effect or a two-factor interaction. However, in this example, the three-factor interactions are also thought to be negligible, so one possible choice is to confound $(ABD; A^2B^2D^2)$ together with $(BCD^2; B^2C^2D)$. This gives the following set of eight confounded degrees of freedom.

$$BCD^2 \ B^2C^2D$$
$$ABD \quad AB^2C \ AC^2D^2$$
$$A^2B^2D^2 \ A^2CD \ A^2BC$$

Thus, each 3-factor interaction (which has 8 degrees of freedom) has two orthogonal contrasts confounded with blocks and six estimable contrasts, which are assumed to be negligible. This means that there are 24 degrees of freedom from the 3-factor interactions and a further 16 degrees of freedom from the $ABCD$ interaction available for estimating σ^2. The design is obtained by using the linear functions L_1 and L_2, corresponding to the selected confounded contrasts ABD and BCD^2 as follows. For each treatment combination, compute the values of L_1 and L_2 modulo 3:

$$L_1 = a_1 + a_2 \qquad + \ a_4 = 0, \ 1, \text{ or } 2 \ (\text{mod } 3).$$
$$L_2 = \qquad a_2 + a_3 + 2a_4 = 0, \ 1, \text{ or } 2 \ (\text{mod } 3).$$

The design is given in Table 14.5, and it can be verified that the nine blocks are obtained from the nine possible pairs of values of L_1 and L_2.   □

14.2.4 A Real Experiment—Dye Experiment

An experiment is described in the book *Design and Analysis of Industrial Experiments*, edited by O. L. Davies, that investigates three reactants (the base material and two inorganic materials, called here M and N) in the manufacture of a cotton dyestuff. The three factors of interest in the experiment were the concentration of M in the free water in the reaction mixture (factor A at three equally spaced levels), the volume of free water in the reaction mixture (factor B at three equally spaced levels), and the concentration of N in the free water in the reaction mixture (factor C at three equally spaced levels).

Table 14.6 Data for dye experiment

Block I		Block II		Block III	
TC	Volume	TC	Volume	TC	Volume
000	74	020	69	010	13
021	130	011	46	001	112
012	56	002	71	022	125
110	110	100	211	120	199
101	166	121	220	111	218
122	227	112	216	102	201
220	195	210	147	200	74
211	146	201	47	221	198
202	90	222	164	212	102

Source Data adapted from *The Design and Analysis of Industrial Experiments*, Second edition, 1979. Editor O. L. Davies. Published by Longman Group Limited.

Although it was possible to control the conditions in the laboratory fairly accurately, the experimenters divided the treatment combinations into blocks of size 9. This was done as a safeguard against time trends, because the time required to complete the investigation was reasonably long. The experiment involved $r = 2$ replications of each treatment combination, but here we will analyze only the first replicate.

The observations were the volumes of dyestuff resulting from the chemical reactions and are shown in Table 14.6. Looking at the treatment combinations (TC) listed in Block I, we can see that they all satisfy the confounding equation $a_1 + 2a_2 + 2a_3 = 0$ (mod 3). Consequently, the experimenters have confounded two contrasts from the 3-factor interaction, which we can label as $(AB^2C^2; A^2BC)$. Since there are only three blocks, these are the only two contrasts confounded. If the 3-factor interaction can be assumed to be negligible, the remaining six degrees of freedom can be used to measure the error variability. The analysis of variance table is shown in Table 14.7. The sum of squares for testing that the main effect of A (averaged over the levels of the other factors) can be calculated either by using the formulae of Chap. 7 or by adding together the sums of squares for two orthogonal contrasts. For example, rule 4 of Sect. 7.3 gives

$$ssA = 9 \sum_{i=1}^{3} \bar{y}_{.i..}^2 - 27 \, \bar{y}_{....}^2$$
$$= 9(5980.44 + 38,590.42 + 16,698.38) - 27(18,045.44)$$
$$= 64,196.222.$$

Two orthogonal contrasts for A are the linear and quadratic contrasts. From Table A.2, the coefficients for the (nonnormalized) linear contrast are $(-1, \, 0, \, 1)$, and those for the quadratic contrast are $(1, \, -2, \, 1)$. The least squares estimates for these two contrasts are

$$\hat{A}_L = (-\bar{y}_{.0..} + \bar{y}_{.2..}) = 51.889$$

and

$$\hat{A}_Q = (\bar{y}_{.0..} - 2\bar{y}_{.1..} + \bar{y}_{.2..}) = -186.333.$$

Table 14.7 Analysis of variance for the dye experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-values
Block	2	182.00			
A	2	64,196.22	32,098.11	26.60	0.0010
A_L	1	12,116.06	12,116.06	10.04	0.0194
A_Q	1	52,080.17	52,080.17	43.16	0.0006
B	2	16,857.56	8,428.78	6.98	0.0271
B_L	1	12,853.39	12,853.39	10.65	0.0172
B_Q	1	4,004.17	4,004.17	3.32	0.1184
C	2	2,334.89	1,167.44	0.97	0.4324
C_L	1	1,422.22	1,422.22	1.18	0.3193
C_Q	1	912.67	912.67	0.76	0.4179
AB	4	12,512.89	3,128.22	2.59	0.1428
AC	4	4,044.89	1,011.22	0.84	0.5481
BC	4	2,698.89	674.72	0.56	0.7015
Error	6	7,240.67	1,206.78		
Total	26	110,068.00			

To normalize the contrasts, one would divide $\hat{A}_L$ by $\sqrt{\Sigma c_i^2 / (rbc)} = \sqrt{2/9}$ and divide $\hat{A}_Q$ by $\sqrt{\Sigma c_i^2 / (rbc)} = \sqrt{6/9}$.

The sum of squares for testing the hypothesis that the linear contrast for A is negligible is the square of the normalized contrast estimate,

$$ss(A_L) = \frac{(-\bar{y}_{.0..} + \bar{y}_{.2..})^2}{2/9} = \frac{(51.889)^2}{2/9} = 12,116.06;$$

the sum of squares for testing the hypothesis that the quadratic contrast for A is negligible is

$$ss(A_Q) = \frac{(\bar{y}_{.0..} - 2\bar{y}_{.1..} + \bar{y}_{.2..})^2}{6/9} = \frac{(-186.333)^2}{6/9} = 52,080.17;$$

and we see that

$$ss(A_L) + ss(A_Q) = 12,116.06 + 52,080.17 = 64,196.23 = ssA.$$

The other sums of squares in Table 14.7 can be obtained in a similar way.

For testing the hypotheses that the three main effects and the three 2-factor interactions are negligible at individual significance levels $\alpha^* = 0.01$ (an overall level significance level of $\alpha \le 0.06$), we would compare the ratios in the analysis of variance table (Table 14.7) with the critical values from the F-distribution ($F_{2,6,0.01} = 10.9$ for the main effects and $F_{4,6,0.01} = 9.15$ for the 2-factor interactions), and we would reject only the hypothesis that the main effect of A is negligible. Plots for the average response due to A and B are shown in Fig. 14.1. We can see from the plot of the A average responses that, as the levels of the concentration of inorganic material M in the free water increase, the volumes of dyestuff first increase and then begin to decrease. We might expect to see both a significant linear trend and a significant quadratic trend. Testing the two hypotheses that the linear trend in A is negligible

Fig. 14.1 Main-effect plots for the dye experiment

(a) Factor A **(b)** Factor B
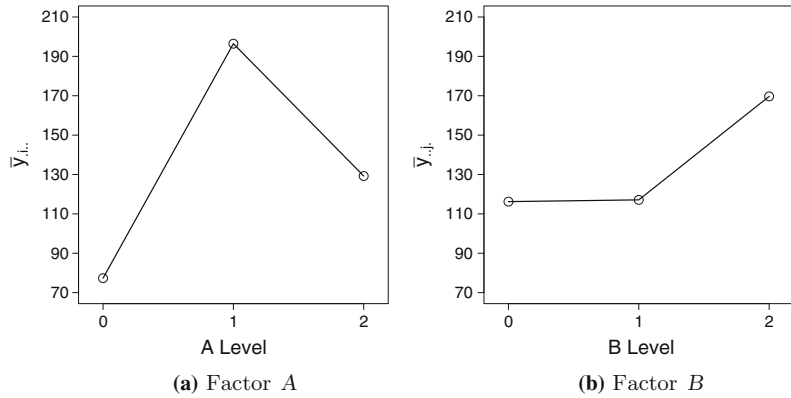
and the quadratic trend in A is negligible, each at level 0.005 (to give an overall significance level of $\alpha^* \leq 0.01$), we have

$$ss(A_L)/msE = 10.04 < F_{1,6,0.005} = 18.6$$

and

$$ss(A_Q)/msE = 43.16 > F_{1,6,0.005} = 18.6,$$

and we conclude that there is a quadratic trend in the levels of A, and that the turning point is towards the center of the range of levels investigated (otherwise, the linear trend would also have been significantly different from zero). Since the objective of the experiment was to boost the volume of dyestuff produced, the results of the experiment suggest that further investigation around the second concentration of inorganic material M might be wise. Although the hypothesis of no effect of B was not rejected, Fig. 14.1 suggests that further experimentation with higher volumes of free water in the reaction mixture is worth consideration.

The above method of testing these two hypotheses uses Bonferroni's method of combining significance levels. An alternative method is to use Scheffé's method of multiple comparisons and test the two hypotheses simultaneously at level 0.01. Since

$$ss(A_L)/msE = 10.04 < 2F_{2,6,0.01} = 21.8$$

and

$$ss(A_Q)/msE = 43.16 > 2F_{2,6,0.01} = 21.8,$$

we arrive at the same conclusion. The more powerful method here is the first since

$$F_{1,6,0.005} < 2F_{2,6,0.01} \, .$$

14.2.5 Plans for Confounded 3^p Experiments

At the end of the chapter, Table 14.20 gives suggested confounding schemes for 3^p experiments in blocks of size 3, 9, or 27. As illustrated in Sect. 13.6, if the design in the table confounds important contrasts in the experiment, then a relabeling of treatment factors should be attempted.

14.3   Designing Using Pseudofactors

14.3.1   Confounding in 4^p Experiments

A treatment factor F with four levels coded 0, 1, 2, 3 can be represented by two factors F_1 and F_2 each having two levels coded 0, 1. The levels of F_1 and F_2 taken together correspond to the levels of the original factor F. One possible correspondence is given below:

F	F_1	F_2
0	0	0
1	0	1
2	1	0
3	1	1

The factors F_1 and F_2 are called *pseudofactors*. All factors in a 4^p experiment can be represented by pseudofactors. Thus, a 4^p experiment in 4^s blocks of size 4^{p-s} can be represented as a 2^{2p} experiment in 2^{2s} blocks of size $2^{2(p-s)}$. The techniques of confounding in a 2^{2p} experiment as discussed in Chap. 13 can therefore be used. The only difference is that an interaction of pseudofactors of the form $F_1G_1G_2$, say, does not represent a 3-factor interaction. It represents one of nine orthogonal contrasts measuring the two-factor interaction, FG. Similarly, F_1F_2 does not represent a contrast in a two-factor interaction. It represents one of three orthogonal contrasts measuring the main effect of factor F.

Example 14.3.1 4^2 experiment in 4 blocks of size 4

Consider a 4^2 experiment with two factors F and G to be run in 4 blocks of size 4. The main effects are to be estimated, but the interaction is thought to be negligible. If F and G are represented by pseudofactors F_1, F_2, G_1, G_2 having two levels each, we can consult Table 13.29 hoping to find a suitable 2^4 experiment in 4 blocks of size 4.

In Table 13.29, we find a design that confounds AC, ABD, and BCD. If we make the correspondence $F_1 = A, F_2 = B, G_1 = C, G_2 = D$, then the design confounds F_1G_1, $F_1F_2G_2$, $F_2G_1G_2$, all three of which belong to the interaction of F and G. All main-effect contrasts of F and G are orthogonal to all interaction contrasts and can therefore be estimated without adjustment for blocks. The design is shown in Table 14.8, with blocks corresponding to combinations of values of $L_1 = a_1 + a_3$ (mod 2) and $L_2 = a_1 + a_2 + a_4$ (mod 2).

If we make a different correspondence, say $F_1 = A, F_2 = D, G_1 = B, G_2 = C$, then a slightly different design is obtained, this time confounding F_1G_2, $F_1F_2G_1$, and $F_2G_1G_2$, which again belong to the interaction of F and G. There is no particular reason to prefer one design over the other. However, a third correspondence, $F_1 = A, F_2 = C, G_1 = B, G_2 = D$, would not be good, since it confounds F_1F_2, $F_1G_1G_2$, $F_2G_1G_2$, and this includes one degree of freedom F_1F_2 from the main effect of F.   □

Table 14.8 4^2 experiment in 4 blocks of 4, confounding three degrees of freedom (F_1G_1, $F_1F_2G_2$, $F_2G_1G_2$) from FG

Block	L_1, L_2	Pseudofactors F_1, F_2, G_1, G_2	Factors F, G
I	0,0	0000 0101 1011 1110	00 11 23 32
II	0,1	0001 0100 1010 1111	01 10 22 33
III	1,0	0010 0111 1001 1100	02 13 21 30
IV	1,1	0011 0110 1000 1101	03 12 20 31

Since two-level pseudofactors are being used, block sizes need only be a power of two, not necessarily a power of four.

14.3.2   Confounding in $2^p \times 4^q$ Experiments

Since factors with 4 levels can be written in terms of pseudofactors having 2 levels each, a $2^p \times 4^q$ experiment can be written in terms of pseudofactors as a $2^{(p+2q)}$ experiment, and no new techniques are needed.

Example 14.3.2 $2^3 \times 4$ `experiment in 4 blocks of size 8`

Suppose that a $2^3 \times 4$ experiment with factors F, G, H, and J is to be run in 4 blocks of size 8. This could be designed using pseudofactors by selecting a design for a 2^5 experiment in 4 blocks from Table 13.29. A design is shown that confounds ABE, CDE, and $ABCD$. If we let the combination of levels of A and C represent the levels of the 4-level factor $J_1 J_2 = J$ with the representation $00 = 0$, $01 = 1$, $10 = 2$, $11 = 3$, and let the levels of B, D, and E respectively represent the levels of F, G, and H, we obtain the design of Table 14.9, that confounds one contrast from each of the 3-factor interactions FHJ, GHJ, and FGJ. All main effects and 2-factor interactions can be estimated. There are 10 degrees of freedom available for estimating σ^2. These come from the two unconfounded degrees of freedom from each of FHJ, GHJ, and FGJ and the one degree of freedom from FGH and the three from $FGHJ$.   □

14.4   Designing Confounded Asymmetric Experiments

A factorial experiment is called an asymmetric experiment when the treatment factors do not all have the same number of levels. For example, $2^2 \times 4^2$, $2^5 \times 3$, $2^2 \times 3^2 \times 4^2$, and 3×6 experiments are all asymmetric experiments. We have already discussed the design of asymmetric $2^p \times 4^q$ experiments in Sect. 14.3.2. We used pseudofactors for the factors with four levels, thus allowing the symmetric designs for 2^{p+2q} experiments to be used. We can use this idea only when the numbers of levels of all factors are powers of the same prime number. For all of the other examples mentioned above, the use of pseudofactors would transform the experiment into a $2^p \times 3^q$ experiment. Consequently, we concentrate on this type of situation in this section.

Since 2 and 3 are relatively prime, the only type of design that can be constructed using the equation method will confound contrasts within the two symmetric parts of the experiment. Consequently, to obtain a design for a $2^p \times 3^q$ experiment in $2^s \times 3^t$ blocks of size $2^{p-s} \times 3^{q-t}$, we combine a design for a 2^p experiment in 2^s blocks with a design for a 3^q experiment in 3^t blocks, using the idea of

Table 14.9 $2^3 \times 4$ experiment in 4 blocks of 8, using pseudofactors and confounding one degree of freedom from interactions FGJ, GHJ, and FHJ

Blocks	Treatment combinations (Factors F, G, H, J)							
I	0000	1002	0101	1103	0013	1011	0112	1110
II	0010	1012	0111	1113	0003	1001	0102	1100
III	0100	1102	0001	1003	0113	1110	0012	1010
IV	0110	1112	0011	1013	0103	1101	0002	1000

a *crossed array* as illustrated in the following example. The total number of blocks created in the combined design is always the product of the numbers of blocks in the original two designs. Likewise, the block sizes in the combined design are products of the block sizes in the original two designs. The confounded contrasts in the combined design are those confounded in the separate designs together with those indicated by all possible products of contrast names.

Example 14.4.1 $2^2 \times 3^2$ `experiment in 6 blocks of size 6`

Suppose that a $2^2 \times 3^2$ experiment is to be run in 6 blocks of size 6. We label the two 2-level factors as A and B and the two 3-level factors as C and D. Since the design must confound within the two symmetric parts of the experiment, one contrast from A, B, or AB must be confounded to divide the 2^2 treatment combinations into two blocks, and one pair of contrasts from C, D, or CD must be confounded to divide the 3^2 treatment combinations into three blocks. The confounded contrasts in the combined design are those confounded in the separate designs together with their products.

For example, we could combine the two designs in Table 14.10. The design labeled d_1 is for a 2^2 experiment in two blocks of size 2 confounding AB, with treatment combinations (TC) grouped into blocks determined by the two values of $L_1 = a_1 + a_2$ (mod 2). The design labeled d_2 is for a 3^2 experiment in three blocks of size 3 confounding the pair of contrasts $(CD^2; C^2D)$, with blocks determined by the three values of $L_2 = a_3 + 2a_4$ (mod 3). The combined array in Table 14.11 divides the treatment combinations into blocks according to the six combinations of values of L_1 and L_2.

A quick way to obtain the design with six blocks is to combine each of the 2 blocks of d_1 with each of the 3 blocks of d_2. For example, to combine the first blocks of d_1 and d_2, each of the combinations 00 and 11 in block I_1 of d_1 is combined with each of the combinations 00, 11, and 22 in block I_2 of d_2 to give the treatment combinations 0000, 1100, 0011, 1111, 0022, 1122 in the first block of the combined design. The other blocks of the combined design are obtained in a similar way.

The $b - 1 = 5$ confounded contrasts are those corresponding to the original confounding schemes, namely the contrast AB and the pair of contrasts represented by $(CD^2; C^2D)$, together with the pair of contrasts represented by the products of these labels—namely $(ABCD^2; ABC^2D)$.   □

Table 14.10 Design d_1 for a 2^2 experiment confounding AB and design d_2 for a 3^2 experiment confounding $(CD^2; C^2D)$

Design	L_1	Block	TC	Design	L_2	Block	TC
d_1	0	I_1	00 11	d_2	0	I_2	00 11 22
	1	II_1	01 10		1	II_2	02 10 21
					2	III_2	01 12 20

Table 14.11 $2^2 \times 3^2$ experiment in 6 blocks of 6, confounding AB, $(CD^2; C^2D)$, $(ABCD^2; ABC^2D)$

L_1, L_2	Block combinations		Treatment combinations
0,0	I_1, I_2	$\rightarrow$ I	0000 0011 0022 1100 1111 1122
0,1	I_1, II_2	$\rightarrow$ II	0002 0010 0021 1102 1110 1121
0,2	I_1, III_2	$\rightarrow$ III	0001 0012 0020 1101 1112 1120
1,0	II_1, I_2	$\rightarrow$ IV	0100 0111 0122 1000 1011 1022
1,1	II_1, II_2	$\rightarrow$ V	0102 0110 0121 1002 1010 1021
1,2	II_1, III_2	$\rightarrow$ VI	0101 0112 0120 1001 1012 1020

Example 14.4.2 `4 × 6 × 3 experiment in 6 blocks of size 12`

Suppose that a $4 \times 6 \times 3$ experiment with factors F, G, H is to be run in 6 blocks of size 12. If we use the pseudofactor labels F_1, F_2, G_1, G_2, and H, then the factors F_1, F_2, and G_1 are in the 2^3 pseudofactor experiment and G_2 and H are in the 3^2 pseudofactor experiment. In the 2^3 experiment, we confound $F_1F_2G_1$ to give the two blocks of the design d_1 of Table 14.12, and in the 3^2 experiment, we confound the pair of contrasts $(G_2H; G_2^2H^2)$ to give the three blocks of the design d_2. Combining each treatment combination in design d_1 with those in d_2 gives the design in Table 14.13, which has $b = 6$ blocks of size 12. The $b - 1 = 5$ confounded degrees of freedom correspond to the original three confounded contrasts together with their products, that is, $F_1F_2G_1$, $(G_2H; G_2^2H^2)$, and $(F_1F_2G_1G_2H; F_1F_2G_1G_2^2H^2)$.

Translating back to the original factors, we can see that one degree of freedom from the interaction FG is confounded, together with two degrees of freedom from each of GH and FGH. This means that all contrasts from the three main effects and also from the interaction FH can be estimated.

If we take the mapping of pseudofactor levels to factor levels as follows, then the design of Table 14.13 is as shown in Table 14.14:

Table 14.12 Design d_1 for a 2^3 experiment confounding $F_1F_2G_1$ and design d_2 for a 3^2 experiment confounding $(G_2H; G_2^2H^2)$

Design	Block	Treatment combinations	Design	Block	Treatment combinations
d_1	I_1	000 011 101 110	d_2	I_2	00 12 21
	II_1	001 010 100 111		II_2	01 10 22
				III_2	02 11 20

Table 14.13 $4 \times 6 \times 3$ experiment in 6 blocks of size 12 using psuedofactors, confounding one degree of freedom from FG and two degrees of freedom from each of GH and FGH

Block combinations			Pseudofactor combinations					
I_1, I_2	$\rightarrow$ I		00000	00012	00021	01100	01112	01121
			10100	10112	10121	11000	11012	11021
I_1, II_2	$\rightarrow$ II		00001	00010	00022	01101	01110	01122
			10101	10110	10122	11001	11010	11022
I_1, III_2	$\rightarrow$ III		00002	00011	00020	01102	01111	01123
			10102	10111	10120	11002	11011	11020
II_1, I_2	$\rightarrow$ IV		00100	00112	00121	01000	01012	01021
			10000	10012	10021	11100	11112	11121
II_1, II_2	$\rightarrow$ V		00101	00110	00122	01001	01010	01022
			10001	10010	10022	11101	11110	11122
II_1, III_2	$\rightarrow$ VI		00102	00111	00120	01002	01011	01023
			10002	10011	10020	11102	11111	11120

Table 14.14 $4 \times 6 \times 3$ experiment in 6 blocks of size 12, confounding one degree of freedom from FG and two degrees of freedom from each of GH and FGH

Block	Treatment combinations
I	000 012 021 130 142 151 230 242 251 300 312 321
II	001 010 022 131 140 152 231 240 252 301 310 322
III	002 011 020 132 141 150 232 241 250 302 311 320
IV	030 042 051 100 112 121 200 212 221 330 342 351
V	031 040 052 101 110 122 201 210 222 331 340 352
VI	032 041 050 102 111 120 202 211 220 332 341 350

F	F_1	F_2		G	G_1	G_2
0	0	0		0	0	0
1	0	1		1	0	1
2	1	0		2	0	2
3	1	1		3	1	0
				4	1	1
				5	1	2

□

14.5   Using SAS Software

In this section we illustrate the use of the SAS software in analyzing a two-replicate factorial experiment with partial confounding. This we do via an example. The analysis is straightforward, as was illustrated in the previous chapter. Along with the correct analysis, we also fit an incorrect model—one without block effects—to illustrate the effect of partial confounding on the analysis.

Example 14.5.1 Dye experiment, continued

The dye experiment was described in Sect. 14.2.4, where part of the data was analyzed as though it came from a single-replicate confounded experiment. In fact, in the original experiment, the design was a partially confounded design made up of two single-replicate 3^3 designs with different confounding schemes. The three factors of interest in the experiment were the concentration of inorganic material M in the free water in the reaction mixture (factor A at three equally spaced levels), the volume of free water in the reaction mixture (factor B at three equally spaced levels), and the concentration of inorganic material N in the free water in the reaction mixture (factor C at three equally spaced levels). The observations were the volumes of dyestuff resulting from the chemical reactions and are shown in Table 14.15 together with the design (prior to randomization). The contrasts $(AB^2C^2; A^2BC)$ are confounded in the first set of three blocks and estimable in the second set, whereas the contrasts $(ABC^2; A^2B^2C)$ are confounded in the second set of three blocks and estimable in the first set.

Since no contrast is completely confounded, no terms need be omitted from the model. Table 14.16 shows the SAS input statements for analyzing this experiment with partial confounding. The statements are exactly as they would be for a replicated experiment with three factors and no confounding. A second run of PROC GLM with no block parameter in the model is included for illustration purposes to show the effect of the partial confounding.

Table 14.15 Data for the dye experiment

| Block I | | Block II | | Block III | |
TC	Volume	TC	Volume	TC	Volume
000	74	020	69	010	13
021	130	011	46	001	112
012	56	002	71	022	125
110	110	100	211	120	199
101	166	121	220	111	218
122	227	112	216	102	201
220	195	210	147	200	74
211	146	201	47	221	198
202	90	222	164	212	102
Block IV		Block V		Block VI	
TC	Volume	TC	Volume	TC	Volume
000	85	010	12	020	115
011	52	021	107	001	148
022	70	002	75	012	47
120	164	100	184	110	145
101	288	111	204	121	142
112	239	122	265	102	216
210	104	220	183	200	75
221	165	201	65	211	124
202	60	212	70	222	114

Source Data adapted from *The Design and Analysis of Industrial Experiments*. Second edition, 1979. Editor O. L. Davies published by Longman Group Ltd.

Table 14.16 SAS program for the dye experiment

```
DATA DYE; INPUT BLK A B C Y;
  LINES;
  1 0 0 0  74
  1 0 2 1 130
  1 0 1 2  56
  1 1 1 0 110
  : : : :  :
  6 2 2 2 114
* Analysis of variance -- correct, with block effect; PROC GLM;
  CLASS  BLK A B C ;
  MODEL Y = BLK A B C A*B A*C B*C A*B*C ;
* Analysis of variance -- without block effect, for comparison;
PROC GLM;
  CLASS  A B C;
  MODEL Y =  A B C A*B A*C B*C A*B*C;
```

Fig. 14.2 Correct analysis of variance for the dye experiment

The GLM Procedure
Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	31	221034.7963	7130.1547	8.25	<.0001
Error	22	19010.8519	864.1296		
Corrected Total	53	240045.6481			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BLK	5	2027.6481	405.5296	0.47	0.7950
A	2	140999.7037	70499.8519	81.58	<.0001
B	2	19447.8148	9723.9074	11.25	0.0004
C	2	4934.4815	2467.2407	2.86	0.0790
A*B	4	27922.6296	6980.6574	8.08	0.0004
A*C	4	13043.6296	3260.9074	3.77	0.0175
B*C	4	2913.1852	728.2963	0.84	0.5130
A*B*C	8	10794.8148	1349.3519	1.56	0.1935

Fig. 14.3 Incorrect analysis of variance, omitting the blocking factor to show the effect of partial confounding

The GLM Procedure
Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	26	219007.1481	8423.3519	10.81	<.0001
Error	27	21038.5000	779.2037		
Corrected Total	53	240045.6481			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	2	140999.7037	70499.8519	90.48	<.0001
B	2	19447.8148	9723.9074	12.48	0.0001
C	2	4934.4815	2467.2407	3.17	0.0582
A*B	4	27922.6296	6980.6574	8.96	<.0001
A*C	4	13043.6296	3260.9074	4.18	0.0092
B*C	4	2913.1852	728.2963	0.93	0.4588
A*B*C	8	9745.7037	1218.2130	1.56	0.1826

All contrasts from the main-effects and 2-factor interactions are orthogonal to the block contrasts and can be estimated without adjustment for blocks. Consequently, the sums of squares for these terms are the same whether or not the block parameter is in the model. This can be verified by comparing the Type III sums of squares for the two runs of PROC GLM shown in Figs. 14.2 and 14.3. Comparing these two analysis of variance tables, we can see that inclusion of the block parameter in the model changes the sum of squares for the three-factor interaction, since the three-factor interaction is partially confounded with blocks. The degrees of freedom for the three-factor interaction remain at 8, as all 8 orthogonal contrasts can be estimated from some portion of the data.

The analysis of variance table (Fig. 14.2) provides *no* evidence that certain contrasts are partially confounded. However, partially confounded contrasts are estimated with larger variance due to the adjustment for blocks. As a result, for the corresponding effects, confidence intervals are wider and tests are less powerful.   □

14.6   Using R Software

In this section we illustrate the use of the R software in analyzing a two-replicate factorial experiment with partial confounding. This we do via an example. The analysis is straightforward, as was illustrated in the previous chapter. Along with the correct analysis, we also fit an incorrect model—one without block effects—to illustrate the effect of partial confounding on the analysis.

Example 14.6.1 `Dye experiment, continued`

The dye experiment was described in Sect. 14.2.4, where part of the data was analyzed as though it came from a single-replicate confounded experiment. In fact, in the original experiment, the design was a partially confounded design made up of two single-replicate 3^3 designs with different confounding schemes. The three factors of interest in the experiment were the concentration of inorganic material M in the free water in the reaction mixture (factor A at three equally spaced levels), the volume of free water in the reaction mixture (factor B at three equally spaced levels), and the concentration of inorganic material N in the free water in the reaction mixture (factor C at three equally spaced levels). The observations were the volumes of dyestuff resulting from the chemical reactions and are shown in Table 14.15 (p. 487) together with the design (prior to randomization). The contrasts $(AB^2C^2; A^2BC)$ are confounded in the first set of three blocks and estimable in the second set, whereas the contrasts $(ABC^2; A^2B^2C)$ are confounded in the second set of three blocks and estimable in the first set.

Since no contrast is completely confounded, no terms need be omitted from the model. Table 14.17 shows the R commands and output for analyzing this experiment with partial confounding. The statements are exactly as they would be for a replicated experiment with three factors and no confounding. A second call of the linear models function `lm` with no blocking factor in the model is included for illustration purposes to show the effect of the partial confounding.

All contrasts from the main-effects and 2-factor interactions are orthogonal to the block contrasts and can be estimated without adjustment for blocks. Consequently, the sums of squares for these terms are the same whether or not the block parameter is in the model. This can be verified by comparing the Type I sums of squares for the two models fit in Table 14.17—the first model with block effects entered first, so for which factorial effects are adjusted for block effects, and the second model without block effects. Inclusion of the block parameter in the model changes the sum of squares for the three-factor interaction, since the three-factor interaction is partially confounded with blocks. The degrees of freedom for the three-factor interaction remain at 8, as all 8 orthogonal contrasts can be estimated from some portion of the data.

The analysis of variance table (the first in Table 14.17) provides no evidence that certain contrasts are partially confounded. However, partially confounded contrasts are estimated with larger variance due to the adjustment for blocks. As a result, for the corresponding effects, confidence intervals (not shown) are wider and tests are less powerful.   □

Table 14.17 R program and output for the dye experiment

```
> dye.data = read.table("data/dye.txt", header=T)
> head(dye.data, 3)

  Blk A B C   y
1   1 0 0 0  74
2   1 0 2 1 130
3   1 0 1 2  56

> # Create factor variables
> dye.data = within(dye.data,
+               {fBlk = factor(Blk); fA = factor(A);
+                 fB = factor(B);   fC = factor(C) })
> # Analysis of variance
> model1 = lm(y ~ fBlk + fA*fB*fC, data=dye.data)
> anova(model1)

Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
fBlk       5    979     196    0.23 0.94703
fA         2 141000   70500   81.58 6.7e-11
fB         2  19448    9724   11.25 0.00043
fC         2   4934    2467    2.86 0.07899
fA:fB      4  27923    6981    8.08 0.00036
fA:fC      4  13044    3261    3.77 0.01749
fB:fC      4   2913     728    0.84 0.51300
fA:fB:fC   8  10795    1349    1.56 0.19353
Residuals 22  19011     864

> # ANOVA without block effects for comparison
> model1 = lm(y ~ fA*fB*fC, data=dye.data)
> anova(model1)

Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
fA         2 141000   70500   90.48 1.1e-12
fB         2  19448    9724   12.48 0.00015
fC         2   4934    2467    3.17 0.05816
fA:fB      4  27923    6981    8.96 9.7e-05
fA:fC      4  13044    3261    4.18 0.00915
fB:fC      4   2913     728    0.93 0.45877
fA:fB:fC   8   9746    1218    1.56 0.18257
Residuals 27  21039     779
```

Exercises

1. Suggest a confounding scheme for a 3^5 experiment in 9 blocks of size 27 if all 2-factor interactions and the 3-factor interaction ABE are to estimated.

2. Suggest a confounding scheme for a 3^5 experiment in 27 blocks of size 9 if all 2-factor interactions and the 3-factor interaction ABE are to be estimated.

3. **Dye experiment, continued**

   (a) For the dye experiment of Sect. 14.2.4, check that the variances of the errors appear to be equal for the different levels of the three factors. Check also that the assumption of normality of the error variables is reasonable.
   (b) Calculate the normalized contrast estimate for Linear $A \times$ Linear B, using the method outlined in Sect. 14.2.4.
   (c) Compute the sum of squares for testing the hypothesis that the Linear $A \times$ Linear B contrast is negligible, using the method outlined in Sect. 14.2.4.
   (d) Test the hypothesis that the Linear $A \times$ Linear B contrast is negligible, using an individual significance level of 0.01.
   (e) Draw an interaction plot for AC and verify that the interaction appears to be negligible.
   (f) Assuming that the contrasts were preplanned, calculate confidence intervals for the pairwise differences in yields due to the three different levels of each of A, B and C. State your overall confidence level.

4. **Dye experiment, continued**

   The experimenters who ran the dye experiment were interested in the linear and quadratic components of the main effects and interactions. Analyze the experiment accordingly. What information have you gathered about the levels of the factors if high yield is of importance?

5. A set of hypothetical data is given in Table 14.18 for a partially confounded 3^2 experiment in 6 blocks of 3. The design is made up of two single-replicate designs: The first confounds the contrasts $(AB; A^2B^2)$ from the interaction, while the second confounds the contrasts $(AB^2; A^2B)$.

   (a) By hand, write out the estimates of the linear and quadratic contrasts for the main effects and their associated variances.
   (b) Using the contrast estimates in part (a), calculate the sums of squares for A and B.

Table 14.18 Partially confounded 3^2 experiment in $b = 6$ blocks of $k = 3$. Hypothetical data are shown in parentheses with corresponding treatment combinations

Replicate	Block	Treatment combinations (Response)		
1	I	00 (53)	12 (59)	21 (80)
Confounds $(AB; A^2B^2)$	II	01 (66)	10 (71)	22 (78)
	III	02 (69)	11 (91)	20 (92)
2	IV	00 (46)	11 (62)	22 (58)
Confounds $(AB^2; A^2B)$	V	01 (65)	12 (61)	20 (76)
	VI	02 (34)	10 (50)	21 (66)

(c) Calculate the least squares estimates of a pair of orthogonal contrasts for $(AB^2; A^2B)$ from the first replicate and the estimates of a pair of orthogonal contrasts for $(AB; A^2B^2)$ from the second replicate. Using these contrast estimates, calculate the sum of squares for the AB interaction (adjusted for blocks).

(d) Prepare an analysis of variance table. Test any hypotheses that you think are of interest and state your conclusions about the two factors and their interaction.

(e) Check your analysis in part (d) using a computer program.

6. **Sugar beet experiment**

F. Yates, in a 1935 paper published in a supplement to the *Journal of the Royal Statistical Society*, describes an agricultural experiment on the yield of sugar beet. The three factors of interest were three standard fertilizers, nitrogen, phosphate, and potassium (factors N, P, and K) each at three equally spaced levels. The experimental field was divided into $b = 3$ blocks and each block subdivided into $k = 9$ 0.1 acre plots. The experiment was designed so that the contrasts $(NP^2K; N^2PK^2)$ were confounded with blocks. The randomized design and yields of sugar beet are shown in Table 14.19.

(a) Prepare an analysis of variance table for the data, assuming that the three-factor interaction is negligible.

(b) Investigate the linear and quadratic trends of the main effects and the two-factor interactions. Yates assumed in his analysis that the only important contrast for each two factor interaction was the linear×linear contrast. Is this assumption supported by your analysis?

(c) Draw any plots that help to illustrate the important features of the analysis.

7. **Example 14.3.2, continued**

In Example 14.3.2, p. 483, we showed one way of associating design factors F, G, H, and J of a $2^3 \times 4$ factorial experiment to the 2-level pseudofactors A–E of a specific design from Table 13.29.

Table 14.19 Data for the sugar beet experiment

Block I		Block II		Block III	
Levels of		Levels of		Levels of	
N, P, K	Yield	N, P, K	Yield	N, P, K	Yield
211	2575	121	2599	202	2189
120	2472	220	2517	020	2093
200	2411	022	2411	210	2354
002	2403	110	2252	111	2268
010	2220	212	2381	001	1926
021	2252	201	2067	122	2152
101	2295	102	2021	221	2349
112	2362	011	1953	012	2025
222	2434	000	1989	100	2106

Source Yates (1935). Copyright © 1935 Blackwell Publishers. Reprinted with permission. (Reprinted in *Experimental Design* (1970), Charles Griffin and Company, Ltd., London. Copyright 1970 Edward Arnold/Hodder & Stoughton Educational. Reprinted with permission.)

There are 10 different ways to make this association (since there are 10 ways of selecting two of A–E to represent J_1 and J_2).

(a) Investigate the confounding schemes for each of the ten possible associations. Specifically, for each association, determine the number of contrasts confounded for each effect, and compare the results.

(b) State under which circumstances you would recommend each design.

8. Consider a $2^2 \times 3^2$ design confounding AB, $(CD^2; C^2D)$, and $(ABCD^2; ABC^2D)$.

(a) Give the design—namely, list the treatment combinations block by block.
(b) Describe how to randomize the design.
(c) Give a set of five orthogonal treatment contrasts that are confounded with blocks.

9. Suggest a confounding scheme for a $2^3 \times 3^3$ experiment in 12 blocks of size 18. Under what circumstances would the design be useful? Write out two blocks of the design.

10. Suggest a confounding scheme for a $2^2 \times 3^2 \times 4$ experiment in 12 blocks of size 12. Under what circumstances would the design be useful? Write out two blocks of the design.

11. Suggest a confounding scheme for a $2^2 \times 3^2 \times 6$ experiment in 9 blocks of size 24. Under what circumstances would the design be useful? Explain how to find the blocks of the design.

12. Suggest a confounding scheme for a $2^2 \times 3^2 \times 6$ experiment in 12 blocks of size 18. Under what circumstances would the design be useful? Write out two blocks of the design.

Table 14.20 Confounding schemes for 3^p experiments in $b = 3^s$ blocks of size $k = 3^{p-s}$. For each design, s independent generators are underlined, and s corresponding equations are given. To obtain Block I of a design, list all k combinations of the first a_i's shown, then use the equations modulo 3 to complete each treatment combination

3^p	b	k	Confounded contrasts	Block I
3^2	3	3	$(\underline{AB}; A^2B^2)$	a_1
				$a_2 = 2a_1$
3^3	3	9	$(\underline{ABC^2}; A^2B^2C)$	a_1, a_2
				$a_3 = a_1 + a_2$
3^3	9	3	$(\underline{AB^2}; A^2B), (\underline{AC}; A^2C^2),$	a_1
			$(BC; B^2C^2), (ABC^2; A^2B^2C)$	$a_2 = a_1$
				$a_3 = 2a_1$
3^4	3	27	$(\underline{ABCD^2}; A^2B^2C^2D)$	a_1, a_2, a_3
				$a_4 = a_1 + a_2 + a_3$
3^4	9	9	$(\underline{AB^2C}; A^2BC^2), (\underline{ABD}; A^2B^2D^2),$	a_1, a_2
			$(AC^2D^2; A^2CD), (BCD^2; B^2C^2D)$	$a_3 = 2a_1 + a_2$
				$a_4 = 2a_1 + 2a_2$
3^5	9	27	$(\underline{ABCD^2}; A^2B^2C^2D), (\underline{AB^2E^2}; A^2BE)$	a_1, a_2, a_3
			$(AC^2DE; A^2CD^2E^2),$	$a_4 = a_1 + a_2 + a_3$
			$(BC^2DE^2; B^2CD^2E)$	$a_5 = a_1 + 2a_2$

Fractional Factorial Experiments

15

15.1 Introduction

Fractional factorial experiments are used frequently in industry, especially in various stages of product development and in process and quality improvement. In a *fractional factorial experiment* only a fraction of the treatment combinations are observed. This has the advantage of saving time and money in running the experiment, but has the disadvantage that each main-effect and interaction contrast will be confounded with one or more other main-effect or interaction contrasts and cannot be estimated separately. Two factorial contrasts that are confounded are referred to as being *aliased*. The term "confounded" is generally reserved for the indistinguishability of a treatment contrast and a block contrast (as described in Chaps. 13 and 14).

We look at two methods of obtaining fractional factorial designs that can be analyzed in a straightforward manner. The first method, described in Sects. 15.2–15.5, is to select one block from one of the single-replicate designs in Chap. 13 as the fraction to be used. The second method of choosing a fraction, described in Sect. 15.6, is popular in industry and uses the concept of an *orthogonal array*. In Sect. 15.7, we aim to identify the treatment factor levels whose effects are least affected by noise variable fluctuations; the intention being to reduce the sensitivity of a product or manufacturing process to uncontrolled variation.

In Sect. 15.8, very small screening designs are introduced. These designs allow one to search among many factors for the few factors which have substantially large main effects. The use of SAS and R software in analyzing fractional factorial experiments is explored in Sects. 15.9 and 15.10, respectively.

15.2 Fractions from Block Designs; Factors with 2 Levels

15.2.1 Half-Fractions of 2^p Experiments; 2^{p-1} Experiments

We start with a very small example to illustrate the ideas of fractional factorial experiments. Suppose that an experiment is to be run with three treatment factors A, B, and C, each having two levels. There are no blocking factors, so the experiment will be run as a completely randomized design. However, only four observations can be taken.

We can obtain a fractional factorial design with just 4 of the 8 total treatment combinations by selecting at random one of the blocks of a single-replicate design with two blocks of size 4. For illustration, consider the block design that confounds the ABC contrast, given in Table 15.1. Suppose

Table 15.1 2^3 experiment in 2 blocks of 4, confounding ABC

Block I	000	011	101	110
Block II	001	010	100	111

Table 15.2 Contrasts for the $\frac{1}{2}$-fraction (001, 010, 100, 111) of a 2^3 experiment, and the corresponding alias scheme

	A	B	C	AB	AC	BC	ABC	Alias scheme
001	-1	-1	1	1	-1	-1	1	$I = ABC$
010	-1	1	-1	-1	1	-1	1	$A = BC$
100	1	-1	-1	-1	-1	1	1	$B = AC$
111	1	1	1	1	1	1	1	$C = AB$

we select the second block, which is (001, 010, 100, 111) and is defined by the equation

$$a_1 + a_2 + a_3 = 1 \bmod 2.$$

The four treatment combinations constitute a $\frac{1}{2}$-*fraction* or $\frac{1}{2}$-*replicate* of a 2^3 experiment, called a 2^{3-1} design, and the ABC contrast is called the *defining contrast* for the fraction. We write

$$I = ABC,$$

which is called the *defining relation* for the fraction.

   With only $n = 4$ observations, there are only $n - 1 = 3$ total degrees of freedom. This means that it is not possible to estimate each of the six remaining contrasts (A, B, C, AB, AC, BC) even if no estimate of σ^2 is required. If we look at the contrasts for a 2^3 experiment (shown in Table 13.2, p. 436) and cross out the rows corresponding to the unobserved treatment combinations, just the contrast coefficients shown in the left part of Table 15.2 remain.

   Table 15.2 shows several interesting features. First, the column corresponding to ABC is not a contrast. The coefficients are the coefficients that one would use in obtaining the sum of the four observations, which is a multiple of the mean. So, ABC is confounded with the mean, and the ABC contrast cannot be measured. The notation $I = ABC$ of the defining relation indicates the equivalence of ABC and the sum of the observations, since I corresponds to a list of coefficients all equal to $+1$.

   Secondly, we see from Table 15.2 that the contrast coefficients for A and BC are identical, the contrasts for B and AC are identical, and the contrasts for C and AB are identical. The main effect of A and the interaction BC are said to be *aliased*, as are B and AC, and C and AB. We write

$$A = BC, \quad B = AC, \quad C = AB.$$

Thus, there are three estimable contrasts in the fraction, but each measures more than one factorial effect. For example, using the cell-means model

$$Y_{ijk} = \mu + \tau_{ijk} + \epsilon_{ijk},$$

the "$A = BC$" contrast with divisor $v/2$ (where v is the number of treatment combinations in the fraction) is obtained by multiplying the τ_{ijk}'s by the coefficients in the column labeled A in Table 15.2, that is,

$$\frac{1}{2}[-\tau_{001} - \tau_{010} + \tau_{100} + \tau_{111}]. \tag{15.2.1}$$

This is an estimable contrast with least squares estimate

$$\frac{1}{2}[-y_{001} - y_{010} + y_{100} + y_{111}].$$

It can be verified that this "$A = BC$" contrast in (15.2.1) can be written as

$$\frac{1}{4}[-\tau_{000} - \tau_{001} - \tau_{010} - \tau_{011} + \tau_{100} + \tau_{101} + \tau_{110} + \tau_{111}]$$
$$+ \frac{1}{4}[+\tau_{000} - \tau_{001} - \tau_{010} + \tau_{011} + \tau_{100} - \tau_{101} - \tau_{110} + \tau_{111}]$$
$$= [-\overline{\tau}_{0..} + \overline{\tau}_{1..}] + \frac{1}{2}[\overline{\tau}_{.00} - \overline{\tau}_{.01} - \overline{\tau}_{.10} + \overline{\tau}_{.11}].$$

Thus, what is being estimated is the sum of the A contrast and the BC contrast, and we could refer to the "$A = BC$" contrast as the $A + BC$ contrast. However, for simplicity, we often refer to it as the A contrast, remembering the role of BC from the list of aliased contrasts.

If in a hypothesis test or half-normal probability plot the main effect of factor A is nonsignificant, then there are two possibilities. One is that neither A nor its alias, BC, is significantly different from zero. The alternative is that the two contrasts have equal and opposite effects and cancel each other out. Since the former is much more likely, this is the assumption that is usually made. If the main effect of A appears to be significant, then it is not clear whether the observed effect is due to the main effect of A or to the BC interaction or the combination of both effects. Because of the aliasing problem, fractional factorial experiments are most often run as *screening experiments*. The word "screening" means that the experimenter is trying to determine which of a large number of factors affect the response (see Sect. 15.8).

The list of aliased contrasts is called the *aliasing scheme* for the design. We generally write this as in the right hand side of Table 15.2, with the first row showing the defining relation and the following rows listing the aliased contrasts. The number of rows in the aliasing scheme is the same as the number of observations in the design.

It can be verified that, if Block I of the single replicate design in Table 15.1 were to be used as the $\frac{1}{2}$-fraction instead of Block II, exactly the same aliasing scheme would result except that, in each pair of aliased contrasts, the coefficients would differ from each other in sign. The fraction would then consist of the four treatment combinations 000, 011, 101, 110, and the A contrast would be

$$\frac{1}{2}[-\tau_{000} - \tau_{011} + \tau_{101} + \tau_{110}]$$
$$= \frac{1}{4}[-\tau_{000} - \tau_{001} - \tau_{010} - \tau_{011} + \tau_{100} + \tau_{101} + \tau_{110} + \tau_{111}]$$
$$- \frac{1}{4}[+\tau_{000} - \tau_{001} - \tau_{010} + \tau_{011} + \tau_{100} - \tau_{101} - \tau_{110} + \tau_{111}]$$
$$= [-\overline{\tau}_{0..} + \overline{\tau}_{1..}] - \frac{1}{2}[\overline{\tau}_{.00} - \overline{\tau}_{.01} - \overline{\tau}_{.10} + \overline{\tau}_{.11}],$$

which we could refer to as the $A - BC$ contrast. This difference in signs can be highlighted by including this information in the aliasing scheme, that is, $I = -ABC$, $A = -BC$, $B = -AC$, and $C = -AB$.

The entire aliasing scheme in the right hand side of Table 15.2 can be deduced from the defining relation without writing out the contrasts. Using the contrast names, we can multiply the defining relation by A to obtain

$$A \times I = A \times ABC.$$

Treating I as a multiplicative identity so that $A \times I = A$, and reducing superscripts modulo 2 so that $A^2BC = BC$, we obtain $A = BC$. The other two rows of the scheme can be obtained in a similar fashion. From now on, we will avoid writing out the contrasts and use the contrast names and the defining relation to obtain the aliasing scheme.

Half-fractions of 2^p experiments have $\frac{1}{2}2^p$ treatment combinations and are called 2^{p-1} experiments. They are almost always obtained by selecting one block from a block design that confounds the highest-order interaction. Thus for $p = 4$, for example, the fraction satisfies either

$$a_1 + a_2 + a_3 + a_4 = 0 \;(\text{mod } 2)$$

and has defining relation $I = ABCD$, or it satisfies

$$a_1 + a_2 + a_3 + a_4 = 1 \;(\text{mod } 2)$$

and has defining relation $I = -ABCD$. For $p = 5$, the fraction satisfies either

$$a_1 + a_2 + a_3 + a_4 + a_5 = 0 \;(\text{mod } 2)$$

and has defining relation $I = -ABCDE$, or it satisfies

$$a_1 + a_2 + a_3 + a_4 + a_5 = 1 \;(\text{mod } 2)$$

and has defining relation $I = ABCDE$. Notice that the sign of the contrast in the defining relation is positive if the equation contains an even number of a_i's and is set equal to 0 (mod 2). It is also positive if the equation contains an odd number of a_i's and also is set equal to 1 (mod 2). Otherwise, the sign is negative. This always holds, even for the more complicated fractions of 2^p experiments discussed in the following sections.

For most purposes, we do not need to know whether the contrasts listed in the defining relation differ in sign. Consequently, unless they are needed, we shall usually ignore the signs in the aliasing scheme.

15.2.2   Resolution and Notation

The defining relation contains a set of contrasts such as AB, ABC, etc. that are aliased with the mean; these contrasts are often called *words*. The number of letters in the shortest word in the defining relation is called the *resolution* of the design.

The design in Table 15.2 is a Resolution III design, since the only word in the defining relation is ABC, which has three letters. In all Resolution III designs, main-effect contrasts are aliased with 2-factor interaction contrasts. In a Resolution IV design, the defining relation contains only words with 4 or more letters. Some main effects are then aliased with 3-factor interactions and 2-factor interactions aliased with other 2-factor interactions. In a Resolution V design, such as that in Table 15.4, some main effects are aliased with 4-factor interactions, and 2-factor interactions are aliased with 3-factor interactions. This is summarized in Table 15.3.

Since the main-effects and low-order interactions are usually the most important factorial effects to be measured, it is generally beneficial to select a design with as high resolution as can be found. The designs in Table 15.60 (at the end of the chapter) all satisfy this requirement.

A $1/2^q$ fraction of a 2^p experiment is usually referred to as a 2^{p-q} *fractional factorial experiment*. The resolution number is sometimes added as a subscript. A resolution III design, for example, can be written as a 2^{p-q}_{III} design.

Table 15.3 Resolution numbers of fractional factorial experiments

Resolution	Main effects aliased with	2-factor interactions aliased with
III	2-factor interactions and/or higher	Main effects and/or interactions
IV	3-factor interactions and/or higher	2-factor interactions and/or higher
V	4-factor interactions and/or higher	3-factor interactions and/or higher

15.2.3   A Real Experiment—Soup Experiment

L.B. Hare, in the (1988) issue of the *Journal of Quality Technology*, described an experiment on a dry soup mix filling process that was run at Thomas J. Lipton, Inc. The company was concerned about keeping the weight of the mix as uniform as possible. They found that most of the variability was due to the uneven flow of the "intermix," which is a mixture of vegetable oil, salt, and other ingredients, during the mixing process. The researchers prepared a list of five treatment factors that they thought might be influential in controlling the mixing process. The factors and their levels (subsequently coded 0, 1) were:

A: number of mixer ports through which vegetable oil was added (two levels, 1 and 3);
B: temperature of mixer jacket (two levels; ambient temperature, presence of cooling water);
C: mixing time (two levels; 60 and 80 sec);
D: batch weight (two levels; 1500 and 2000 lb);
E: delay between mixing and packaging (two levels; 1 day and 7 days).

This was a screening experiment, since the researchers had little idea of which factors were going to turn out to be important in affecting the variability of the soup mix weight. They decided to run a $\frac{1}{2}$-fraction to investigate the five factors, and follow up the experiment with a more detailed study of the important factors later. They chose a Resolution V design with defining relation $I = ABCDE$, which allowed them to include all main effects and two-factor interactions in the model. The corresponding block design, which confounds *ABCDE*, is listed in Table 13.29. The experimenters chose the second block, as it contained the treatment combination that represented the normal operating conditions prior to the experiment. These were 00010, that is, one port, presence of cooling water, 60 s mix, 2000 lb batch weight, and a one day delay before packaging.

The experiment was designed so that it could be run with very little disruption to the daily production routine. Sets of 5 samples were taken every 15 min during the production run for each treatment combination and weighed. The response variable was a measure of variation based on these weights. The randomized design and the responses obtained are shown in Table 15.4. Also shown in the table are the contrasts for the main effects. As in Chap. 13, the contrast has coefficient -1 when the corresponding factor is at its low level and coefficient $+1$ when it is at its high level. The contrast coefficients for the interactions are the products of the corresponding main-effect contrast coefficients.

The experimenters included all the main effects and 2-factor interactions in the model. Since there were $16 - 1 = 15$ degrees of freedom in total and 15 contrasts to estimate (5 main effects and 10 two-factor interactions), there were no degrees of freedom available to estimate the error variability. The experimenters calculated all the contrast estimates and prepared a normal probability plot to find the important contrasts. For example, the contrast estimate for the main effect of E is

$$\hat{E} = (0.78 + 1.10 - 1.70 - 1.28 + 0.97 + 1.47 - 1.85 - 2.10 + 0.76$$
$$+ 0.62 - 1.09 - 1.13 + 1.25 + 0.98 - 1.36 - 1.18)/8 = -0.47.$$

Table 15.4 Design, data (measure of weight variability), and main-effect contrasts for the soup experiment. Defining relation $I = ABCDE$

Levels of A, B, C, D, E	y_{ijklm}	Contrasts				
		A	B	C	D	E
01011	0.78	-1	1	-1	1	1
11111	1.10	1	1	1	1	1
10000	1.70	1	-1	-1	-1	-1
11100	1.28	1	1	1	-1	-1
00001	0.97	-1	-1	-1	-1	1
01101	1.47	-1	1	1	-1	1
00010	1.85	-1	-1	-1	1	-1
10110	2.10	1	-1	1	1	-1
00111	0.76	-1	-1	1	1	1
10011	0.62	1	-1	-1	1	1
01110	1.09	-1	1	1	1	-1
01000	1.13	-1	1	-1	-1	-1
11001	1.25	1	1	-1	-1	1
10101	0.98	1	-1	1	-1	1
11010	1.36	1	1	-1	1	-1
00100	1.18	-1	-1	1	-1	-1

Source Hare (1988). Reprinted with Permission from Journal of Quality Technology © 1998 ASQ, www.asg.org

Table 15.5 Contrast estimates (with divisor $v/2 = 8$) for the soup experiment

Contrasts	A	B	C	D	E	AB	AC	
Estimates	0.145	-0.088	0.038	-0.038	-0.470	-0.015	0.095	
Contrasts	AD	AE	BC	BD	BE	CD	CE	DE
Estimates	0.030	-0.153	0.068	-0.163	0.405	0.073	0.135	-0.315

Fig. 15.1 Soup experiment: half-normal probability plot of contrast absolute estimates (using divisor $v/2 = 8$)
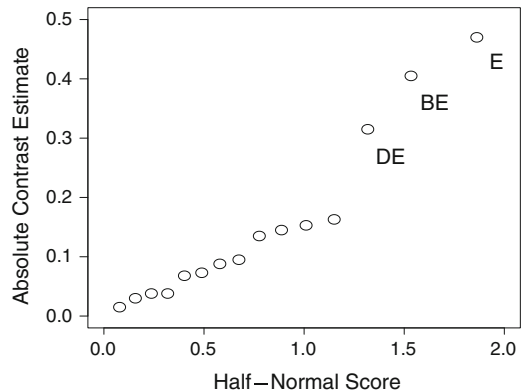
The main effect and two-factor interaction contrast estimates are shown in Table 15.5. In Fig. 15.1, we present a half-normal probability plot (Sect. 7.5.2) which shows the absolute values of the contrast estimates plotted against their half-normal scores. It can be seen that the most important contrasts appear to be E, BE, and DE.
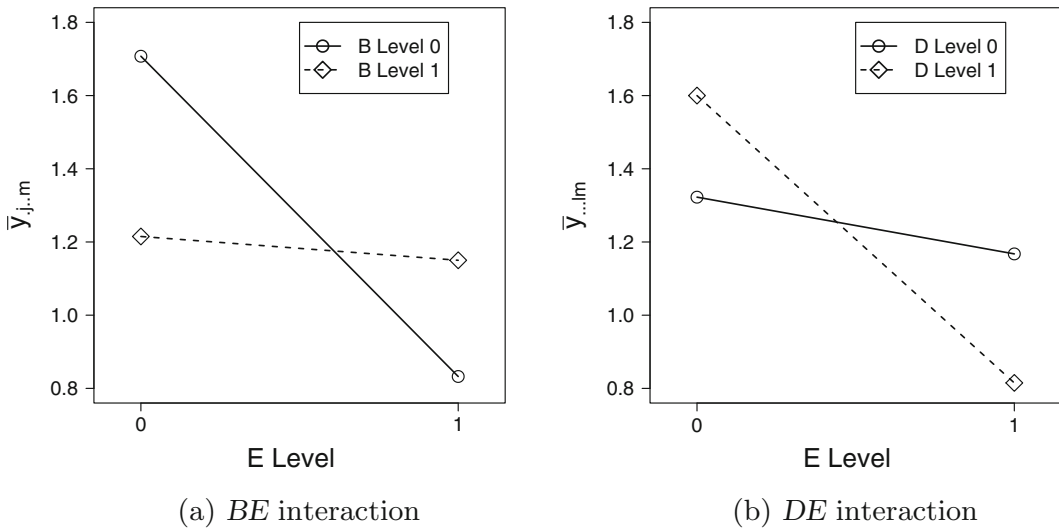
Fig. 15.2   Interaction plots for the soup experiment

Interaction plots of the two interactions BE and DE are shown in Fig. 15.2. The response is a measure of weight variability, and the experimenters wanted to reduce this as much as possible. The estimate of the E contrast is negative, indicating that the low level of E (one-day delay before packaging) is more variable that the high level (seven-day delay). The two interaction plots also indicate that a seven-day delay before packaging would be beneficial using the ambient temperature (low level of B) and the large batch size (2000 lb, high level of D). The interaction plots also indicate that if a seven-day delay is not feasible, then it is better to use cooling water and a small batch size (high B, low D). The production management of the company agreed to a seven-day delay, and the researchers decided to investigate these three factors (B, D, and E) in more detail in a followup experiment, together with factor C, whose interaction with E was the next largest effect.

15.2.4   Quarter-Fractions of 2^p Experiments; 2^{p-2} Experiments

We can obtain a $\frac{1}{4}$-fraction of a 2^p experiment by selecting at random one of the blocks from a single-replicate confounded design with 4 blocks of size 2^{p-2}. The defining relation is then the set of three contrasts that were confounded to obtain the block design.

For example, suppose a 2^5 experiment was to be run as a completely randomized design, but only eight observations could be taken, and the main effects and interaction AE were of particular interest. Table 13.29 (p. 471) lists a design in 4 blocks of 8. Suppose we switch E and D in the listed design, then we obtain a design in 4 blocks of 8 which confounds the three interactions ABD, CDE, and $ABCE$. We can then select one block for the $\frac{1}{4}$-fraction; for example, if we select the block that satisfies

$$a_1 + a_2 + a_4 = 1 \ (\text{mod } 2) \quad \text{and} \quad a_3 + a_4 + a_5 = 0 \ (\text{mod } 2),$$

the treatment combinations in the resulting fraction are

$$00011 \ \ 00110 \ \ 01000 \ \ 01101 \ \ 10000 \ \ 10101 \ \ 11011 \ \ 11110 \ .$$

Table 15.6 Aliasing scheme (ignoring signs) for a $\frac{1}{4}$-fraction of a 2^5 experiment with the defining relation $I = ABD = CDE = ABCE$

I	$=$	ABD	$=$	CDE	$=$	$ABCE$
A	$=$	BD	$=$	$ACDE$	$=$	BCE
B	$=$	AD	$=$	$BCDE$	$=$	ACE
C	$=$	$ABCD$	$=$	DE	$=$	ABE
D	$=$	AB	$=$	CE	$=$	$ABCDE$
E	$=$	$ABDE$	$=$	CD	$=$	ABC
AC	$=$	BCD	$=$	ADE	$=$	BE
AE	$=$	BDE	$=$	ACD	$=$	BC

The defining relation for the fraction is

$$I = ABD = CDE = ABCE\,.$$

(If we work out the contrast coefficients for this fraction, we find that the coefficients for ABD are all $+1$, while those for CDE and $ABCE$ are all -1. Thus, if the signs of the contrasts were taken into account, the defining relation would be $I = ABD = -CDE = -ABCE$.) The other seven rows of the aliasing scheme are obtained by multiplying the defining relation by each of the contrast names in turn. The resulting aliasing scheme (ignoring signs) is shown in Table 15.6.

Only one factorial effect from each row of the aliasing scheme (and none from the defining relation) can be entered into the model for analyzing the experiment. So, for example, we could include all main effects and the two 2-factor interactions AC and AE in the model. (Notice that, if we had not made the switch of labels E and D, then AE would have been aliased with B.)

If the D effect, for example, is insignificant, the interactions AB, CE, and $ABCDE$ are also regarded as insignificant. But if the analysis shows that D has a significant effect on the response, it is unknown whether the effect is due to the main effect of D, or to AB, or to CE, or to $ABCDE$ (although this latter effect is the least likely), or to some combination of all four. The design is useful for screening when it is believed that most main effects and interactions will be negligible but one or two factors will possibly have an important effect on the response.

This design is clearly ideal if all of the interactions are negligible, or if all interactions except exactly one of AC, BE, AE, and BC are thought to be negligible. In the first case, two degrees of freedom are available to estimate σ^2. In the second case, all of the main effects and the one interaction can be measured, and one degree of freedom remains to estimate σ^2. If all main effects and, say, the CD interaction were required to be estimated, then a different block design should be chosen since, in this design $CD = E$. For example, a suitable design could be obtained by interchanging A and D in the list of confounded contrasts. In other words, the design obtained by confounding ABD, ACE, and $BCDE$ will give a $\frac{1}{4}$-fraction in which CD is not aliased with main effects.

A list of useful $\frac{1}{4}$-fractions is given in Table 15.60 at the end of the chapter.

Example 15.2.1 `Sludge experiment`

S.R. Schmidt and R.G. Launsby, in their textbook *Understanding Industrial Designed Experiments*, include an article by J. Brickell and K. Knox on the operation of a biological treatment system (known as an activated sludge system) used in wastewater treatment plants. The details of the system are given in the article. The response, Y, is the removal of "biochemical oxygen demand," which is related to the quality of water. The water quality increases as more biochemical oxygen demand is removed, so the

response Y is to be maximized. The experiment described in the article investigates the effect of five factors on Y:

- A: Reactor biomass concentration (3000 and 6000 mg/l),
- B: Clarifier biomass concentration (8000 and 12000 mg/l),
- C: Waste sludge flow rate (78.5 and 940 m^3/d),
- D: Biological growth rate constant (0.040 and 0.075 d^{-1})
- E: Fraction of food to biomass (0.4 and 0.8 kg/kg).

Since this experiment was to be run in a water treatment plant, it was necessary to keep the number of observations small, and a $\frac{1}{4}$-fraction was selected with defining relation $I = ABD = CDE = ABCE$. This gives the aliasing scheme of Table 15.6.

The experimenters selected the fraction whose treatment combinations, written as $a_1a_2a_3a_4a_5$, satisfied

$$
\begin{aligned}
a_1 + a_2 \quad + a_4 \quad &= 1 \ (\text{mod } 2) \\
a_3 + a_4 + a_5 &= 1 \ (\text{mod } 2)
\end{aligned}
$$

The design, prior to randomization, is shown in Table 15.7 together with the responses obtained.

The experimenters included all main effects and the 2-factor interactions AC and BC in their model. The contrast estimates (with divisors $v/2 = 4$) are listed in Table 15.8, and a half-normal probability plot of the seven contrast estimates is shown in Fig. 15.3. There are too few contrast estimates in total to be able to draw good conclusions from the half-normal probability plot. Nevertheless, the most important effect appears to be the main effect of C and, perhaps to a lesser extent, E. Now, C is aliased with DE, and E is aliased with CD. A followup experiment investigating the effects of C, D, and E would certainly be advisable.

If we try to draw conclusions from the results of the present experiment, and if we are willing to assume that the main effects are the dominant effects in any alias sets, it would seem advisable to set C and possibly B at their high levels in order to maximize the response, and to set E and possibly D at their low levels. On the other hand, if we assume that the interactions in the alias sets might be

Table 15.7 $\frac{1}{4}$-fraction of a 2^5 experiment and data from the sludge experiment

Levels of A, B, C, D, E	y_{ijklm}	Contrasts						
		A	B	C	D	E	AC	BC
00010	195	−1	−1	−1	1	−1	1	1
00111	496	−1	−1	1	1	1	−1	−1
01001	87	−1	1	−1	−1	1	1	−1
01100	1371	−1	1	1	−1	−1	−1	1
10001	102	1	−1	−1	−1	1	−1	1
10100	1001	1	−1	1	−1	−1	1	−1
11010	354	1	1	−1	1	−1	−1	−1
11111	775	1	1	1	1	1	1	1

Source Brickell and Knox (1992). Copyright © 1992 Air Academy Press. Reprinted with permission

Table 15.8   Contrast estimates (with divisor $v/2 = 4$) for the sludge experiment

Contrast	A	B	C	D	E	AC	BC
Estimate	20.75	198.25	726.25	−185.25	−365.25	−66.25	126.25

Fig. 15.3 Half-normal probability plot of contrast absolute estimates for the sludge experiment
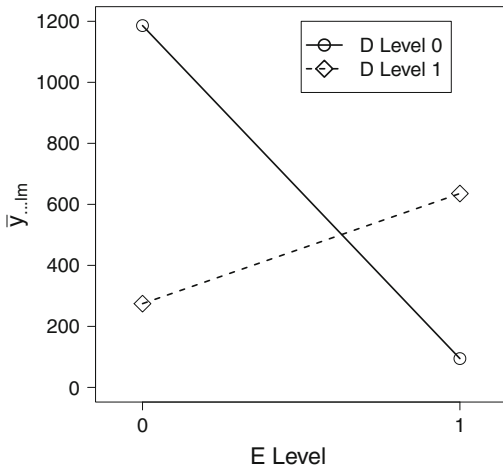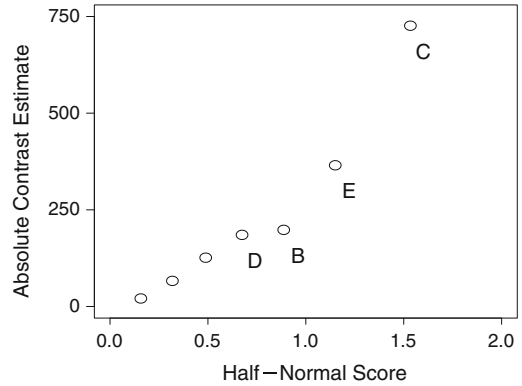
(a) DE interaction
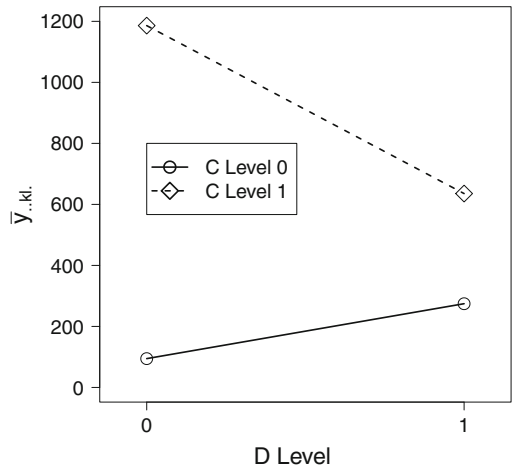
(b) CD interaction

Fig. 15.4 Interaction plots for the sludge experiment

important, we would examine the DE and CD interaction plots (see Fig. 15.4). The DE plot suggests that D and E should both be at their low levels, and the CD plot suggests that C should be at its high level and D at its low level. Since the recommendations from the interaction plots agree with those from the main effect comparisons, we would feel comfortable in recommending that the process be set at the cheaper of 01100 or 11100. Notice that the first of these was included among the experimental runs (and happened to give rise to the largest observed yield), whereas the second was not included. In either case, the experiment should be re-run at the chosen setting to confirm the results.

The authors of the article point out that other considerations, such as cost, come into play before any system can be changed. In an actual water treatment plant, it would be expensive to change the levels of D (biological growth rate constant) and E (fraction of food to biomass) from their current settings. Also, increasing the waste sludge flow rate (C) increases cost. A followup experiment could verify that the above recommendations were correct, and also could examine intermediate values of C. □

15.2.5   Smaller Fractions of 2^p Experiments

Smaller fractions of a 2^p experiment can be obtained in exactly the same way as the $\frac{1}{2}$-fractions and $\frac{1}{4}$-fractions of the preceding subsections. For a $1/2^s$ fraction, the first step is to find a design with 2^s blocks of size 2^{p-s} that confounds negligible interactions. One block is selected at random. The aliasing scheme is then checked to ensure that as few important contrasts as possible are aliased with each other. If the aliasing scheme is not suitable, then an attempt is made to obtain a better design by interchanging letters in the confounding scheme, or by investigating different confounding schemes. A list of useful $\frac{1}{4}$-, $\frac{1}{8}$- and $\frac{1}{16}$-fractions of 2^p experiments is given in Table 15.60 at the end of the chapter.

Example 15.2.2 `Welding experiment`

An experiment was discussed by A.K. Shahani in *The Statistician* in 1970 which involved a (1/1024) fraction of a 2^{21} experiment (that is, a 2^{21-16} experiment). The experiment, which required only $v = 2^5 = 32$ observations, was designed by Dr. Shahani for Bristol Aerojet Ltd and concerned the "pull strength" of welds resulting from a certain welding process. The company wished to discover which settings of the 21 factors would give welds with pull strength exceeding a given size. Of the 21 factors, only a few were expected to have important effects on the pull strength, and this allowed the use of such a highly fractionated design.

All 21 factors were easy to manipulate, and the engineers selected two reasonable settings for each factor (coded 0 and 1). For some of the factors, the two levels chosen were at equal distances on each side of the current operating conditions. For others, such as factors A, D, and W, the low levels were at the current operating conditions and could not be lowered further. If we label the factors A, B, ..., W omitting I and O, the contrasts selected for confounding were as follows:

$$
\begin{array}{llll}
ABV & ACW & ADT & AES \\
BCU & ABEN & ACDQ & ACEP \\
ADEM & BCER & BDEL & CDEK \\
ABCEH & ABDEJ & ACDEG & BCDEF
\end{array}
$$

The defining relation consists of these 16 contrasts together with all their possible products. Since the shortest word in the defining relation is of length 3, the design is Resolution III. Although each main effect is aliased with several two-factor interactions, the main effects are not aliased with each other.

The 32 treatment combinations and their responses are shown in Table 15.9. (We have corrected typing errors that occurred in the original paper in the two treatment combinations in the second row of our table). The responses are in coded units, details of which were not given in the original paper.

This experiment has too many factors to be able to analyze it easily by hand. The main-effect contrast estimates (with divisors $v/2 = 16$), obtained from a computer package, are shown in Table 15.10. Under the current operating conditions, it was known that the error standard deviation σ was about 60 units. The experimenters were willing to assume that this would not change appreciably under different operating conditions and therefore calculated the standard error of a main effect contrast $\Sigma_i c_i \tau_i$ to be

$$
\sqrt{\mathrm{Var}(\Sigma_i c_i \hat{\tau}_i)} = \sigma \sqrt{\Sigma c_i^2} = 60\sqrt{32/16^2} = 21.21.
$$

Without assuming that the coded responses follow a normal distribution, the experimenters then deemed any contrast whose estimated absolute value turned out to be several times larger than 21.21 to be important.

Table 15.9 Treatment combinations and responses for the 2^{21-16} welding experiment

Treatment combination	Response	Treatment combination	Response
00000111100000011111	430	10000100001111000100	422
01000010001010011001	336	11000001101101100010	380
00100001100011111010	438	10100110101100100001	96
01100101011011011100	394	11100110111100000111	319
00010001011001010111	334	10010010110110001100	202
01010100110110110001	322	11010111010001010101 0	238
00110110001101011001 0	184	10110101101010110100 1	188
01110011100111001010 0	348	11110000000000100111 1	−234
00001000011111010111 1	384	10001011111000111010 0	338
01001101110101000100 1	404	11001110010010101001 0	370
00101111001110100101 0	542	10101100101001001000 1	114
01101010100100110110 0	316	11101001000011011011 1	432
00011110100011110011 1	256	10011101000100011110 0	206
01011011001001100000 1	82	11011000101110001101 0	106
00111001110010000001 0	528	10111010010101101100 1	110
01111100011000010010 0	35	11111111111111111111 1	370

Source Shahani (1970). Copyright © 1970 Blackwell Publishers. Reprinted with permission

Table 15.10 Contrast estimates for the welding experiment (with divisor 16)

Contrast	A	B	C	D	E	F	G
Estimate	−104.8	−34.6	−39.4	−152.5	12.3	37.4	5.3
Contrast	H	J	K	L	M	N	P
Estimate	101.9	70.5	48.4	−23.4	43.5	100.1	32.9
Contrast	Q	R	S	T	U	V	W
Estimate	16.6	3.0	42.1	−8.1	7.1	72.8	−69.0

The contrast estimates whose absolute values exceed 63.63 are those for the main effects of A, D, H, J, N, V, and W. The estimates for main effects of M, K, and S are all around 2 standard errors, with those for C and F a little smaller.

Since there are 32 observations, a total of 31 orthogonal contrasts can be measured. Thus there are 10 sets of confounded interaction contrasts that can be measured in addition to the 21 main-effect contrasts. The identification of these contrast sets requires writing out the entire aliasing scheme—a daunting task! A proper analysis of the main effects also requires knowledge about which interactions are aliased with which main effects. A followup experiment to separate out the most likely aliased effects would be needed.

Assuming, temporarily, that the process can be improved by considering the main effects only, the contrast estimates (high level minus low level) suggest that factors A, D, and W (whose contrast estimates are negative) should be set at their low levels and factors H, J, K, N, V, M, and S (whose contrast estimates are positive) should be set at their high levels. As mentioned above, A, D, and W were already set at the lowest possible values in the original process, and therefore further experimentation with these factors is unnecessary. The other seven factors were discussed by the engineers and new (higher) settings selected for these, resulting in an improved process that met the pull strength requirements. The author of the article commented that the research and development department should give consideration to a further experiment involving these seven factors in which main effects and two-factor

interactions could all be measured. He suggested the use of a 2^{7-1} experiment, which would require 64 observations. Fewer observations would require aliasing some of the 2-factor interactions (see Table 15.60). □

15.3   Fractions from Block Designs; Factors with 3 Levels

15.3.1   One-Third Fractions of 3^p Experiments; 3^{p-1} Experiments

To obtain a fraction of a 3^p experiment, we use the same idea that we used for 2^p experiments. We select one block at random from a confounded single-replicate design with 3^s blocks of size 3^{p-s} with a suitable confounding scheme. For example, suppose a $\frac{1}{3}$-fraction of a 3^4 experiment is required (that is, a total of $3^{4-1} = 27$ observations). The highest-order interaction in a 3^4 experiment that can be confounded is the 4-factor interaction. Therefore, the maximum number of letters in a word in the defining relation of the fraction is also four. For a Resolution IV design, when 3- and 4-factor interactions are negligible, the main-effect contrasts can be estimated, but the 2-factor interactions will be aliased. This is the best design available and, unless a larger budget can be obtained to allow more observations, some aliasing among the low-order interactions will have to be tolerated.

Suppose the selected single-replicate confounded design is that which confounds the pair of contrasts $(ABCD; A^2B^2C^2D^2)$ from the 4-factor interaction. The block design is constructed using the equations

$$a_1 + a_2 + a_3 + a_4 = 0, \ 1, \text{ or } 2 \ (\text{mod } 3)$$

as in Sect. 14.2.3, and one block is selected at random for the $\frac{1}{3}$-fraction. Since there are 27 treatment combinations to be observed, the aliasing scheme has 27 rows. Seven rows from the aliasing scheme are given in Table 15.11. The remaining rows contain main effects or 2-factor interaction contrasts (such as AB^2 or A^2B) that are aliased only with higher-order interactions. The 27 rows of the aliasing scheme include one for effects aliased with the mean, together with 13 additional pairs of rows, such as the pair of rows involving AB and A^2B^2 which represent contrasts from the same two-factor interaction. The two rows containing AB and A^2B^2 indicate, for example, that the pair of contrasts $(AB; A^2B^2)$ is aliased with the pairs of contrasts $(CD; C^2D^2)$ and $(ABC^2D^2; A^2B^2CD)$. Use of this design is illustrated in Example 15.3.1.

Example 15.3.1 `Refinery experiment`

John (1971) describes an experiment of Vance (1962) to find a set of operating conditions to optimize the quality of lube oil treated at a refinery. There were four factors of interest, called here A, B, C, and D, and three equally spaced levels were selected for each of these so that quadratic trends could be measured.

Table 15.11 Seven rows from the aliasing scheme for a $\frac{1}{3}$-fraction of a 3^4 experiment with the defining relation $I = ABCD = A^2B^2C^2D^2$

I	$=$	$ABCD$	$=$	$A^2B^2C^2D^2$				
AB	$=$	A^2B^2CD	$=$	C^2D^2	A^2B^2	$=$	CD	$=$	ABC^2D^2
AC	$=$	A^2BC^2D	$=$	B^2D^2	A^2C^2	$=$	BD	$=$	AB^2CD^2
AD	$=$	A^2BCD^2	$=$	B^2C^2	A^2D^2	$=$	BC	$=$	AB^2C^2D

Table 15.12 $\frac{1}{3}$-fraction of a 3^4 experiment and data from the refinery experiment

Treatment combination	y_{ijkl}	A_L	A_Q	B_L	B_Q	C_L	C_Q	D_L	D_Q
0000	4.2	−1	1	−1	1	−1	1	−1	1
0012	5.9	−1	1	−1	1	0	−2	1	1
0021	8.2	−1	1	−1	1	1	1	0	−2
0102	13.1	−1	1	0	−2	−1	1	1	1
0111	16.4	−1	1	0	−2	0	−2	0	−2
0120	30.7	−1	1	0	−2	1	1	−1	1
0201	9.5	−1	1	1	1	−1	1	0	−2
0210	22.2	−1	1	1	1	0	−2	−1	1
0222	31.0	−1	1	1	1	1	1	1	1
1002	7.7	0	−2	−1	1	−1	1	1	1
1011	16.5	0	−2	−1	1	0	−2	0	−2
1020	14.3	0	−2	−1	1	1	1	−1	1
1101	11.0	0	−2	0	−2	−1	1	0	−2
1110	29.0	0	−2	0	−2	0	−2	−1	1
1122	55.0	0	−2	0	−2	1	1	1	1
1200	8.5	0	−2	1	1	−1	1	−1	1
1212	37.4	0	−2	1	1	0	−2	1	1
1221	66.3	0	−2	1	1	1	1	0	−2
2001	11.4	1	1	−1	1	−1	1	0	−2
2010	21.1	1	1	−1	1	0	−2	−1	1
2022	57.9	1	1	−1	1	1	1	1	1
2100	13.5	1	1	0	−2	−1	1	−1	1
2112	51.6	1	1	0	−2	0	−2	1	1
2121	76.5	1	1	0	−2	1	1	0	−2
2202	31.0	1	1	1	1	−1	1	1	1
2211	74.5	1	1	1	1	0	−2	0	−2
2220	85.1	1	1	1	1	1	1	−1	1

Sources John (1971). Copyright © 1971 P.W.M. John. Reprinted with permission

Since this was a preliminary experiment, a $\frac{1}{3}$-fraction of Resolution IV was thought to be adequate. The experimenters used a design with defining relation $I = ABCD = A^2B^2C^2D^2$. Part of the aliasing scheme is shown in Table 15.11. We see from row 2 that two degrees of freedom from the AB interaction are aliased with two degrees of freedom from each of the CD and $ABCD$ interactions. The other two degrees of freedom from each of these interactions are aliased with 3-factor interactions. (For example, the pair $(AB^2; A^2B)$ is aliased with the pairs $(A^2CD; AC^2D^2)$ and $(BC^2D^2; B^2CD)$). A similar confounding pattern occurs with AC and BD and also with AD and BC.

The treatment combinations can be obtained from the equation

$$a_1 + a_2 + a_3 + a_4 = 0 \ (\text{mod } 3)$$

and are shown in Table 15.12, prior to randomization, together with the data collected. Also shown are the linear and quadratic contrast coefficients for the main effects. The objective of the experiment was to select factor levels that would increase the response (a measure of quality).

The analysis of variance is complicated by the aliasing of pairs of degrees of freedom for two-factor interactions. We have not tried to separate these but have listed the contributions of the pairs of

Table 15.13 Analysis of variance for the refinery experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square
A	2	4496.29	2248.14
A_L	1	4399.22	4399.22
A_Q	1	97.07	97.07
B	2	2768.69	1384.35
B_L	1	2647.49	2647.49
B_Q	1	121.20	121.20
C	2	5519.79	2759.89
C_L	1	5516.00	5516.00
C_Q	1	3.79	3.79
D	2	283.37	141.68
D_L	1	213.56	213.56
D_Q	1	69.81	69.81
AB, CD	6	339.00	56.50
AC, BD	6	1384.24	230.71
AD, BC	6	753.38	125.56
Total	26	15,544.66	

interactions on the same line of the analysis of variance table shown in Table 15.13. Without information concerning negligible interactions, we are unable to obtain an estimate for the error variance. The most important interactions appear to be the AC or BD interactions, and the BC or AD interactions; the corresponding interaction plots are shown in Fig. 15.5. In each case, the plots indicate that in order to increase the response, factors A, B, and C should all be set at their high levels, cost permitting, and factor D should be set at its middle level. They also indicate that since the lines are not too far from parallel, it would be reasonable to examine the main-effect contrasts.

Normalized linear and quadratic main-effect contrast estimates are obtained as

$$\frac{1}{d} \sum_i \sum_j \sum_k \sum_l c_{ijkl} \hat{\tau}_{ijkl} = \frac{1}{d} \sum_i \sum_j \sum_k \sum_l c_{ijkl} y_{ijkl} \, ,$$

where the c_{ijkl}'s are the contrast coefficients in Table 15.12 and the divisor d is the square root of the sum of squares of the coefficients (that is, $\sqrt{18}$ for the linear contrasts and $\sqrt{54}$ for the quadratic contrasts). These estimates are listed in Table 15.14, and a half-normal probability plot of the estimates is shown in Fig. 15.6. Eight estimates are too few to make a good judgment, but the most important effects appear to be the linear trends in C, A, and B (in that order). All of these contrast estimates are positive, suggesting that the high levels should be selected in order to increase the response. This agrees with the conclusions from the interaction plots as well as the observed data in Table 15.12. Note that we could have examined interactions more closely by including individual interaction contrast estimates in the half-normal probability plot. We have not done this because of the complicated confounding of the interactions.

Since we have no estimate for error, we are unable to test any hypotheses. However, had the experimenters believed, prior to the experiment, that some or all of the interactions were negligible, then tests would have been done for the remaining interactions and the main effects. The sums of squares for testing the linear and quadratic main-effect contrasts are the squares of the corresponding normalized contrast estimates in Table 15.14. For example, the sums of squares for testing the

(a) AC interaction

(b) BD interaction

(c) AD interaction
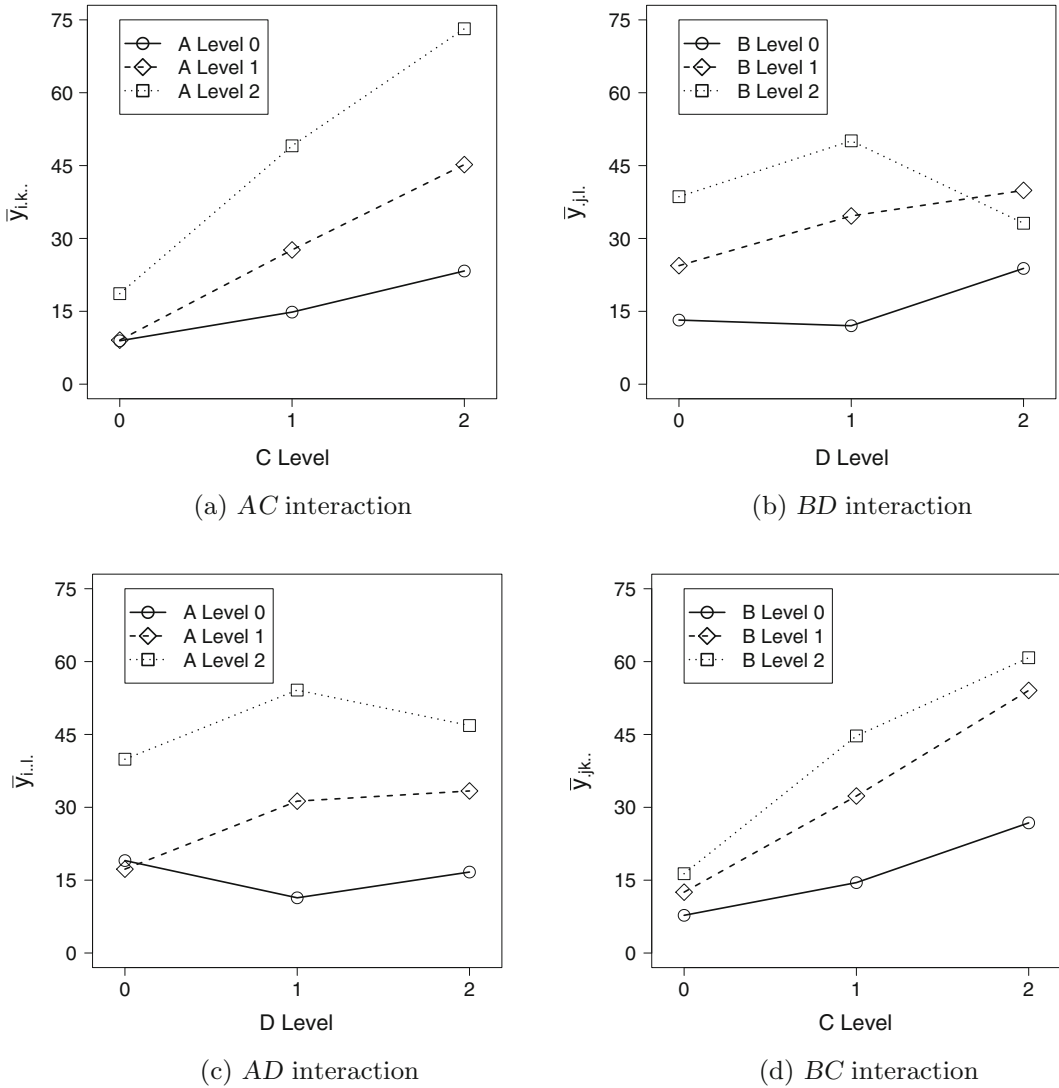
(d) BC interaction

Fig. 15.5 Interaction plots for the refinery experiment

hypothesis that the linear trend of factor A is negligible, against the alternative hypothesis that it is not negligible, is

$$ss(A_{\mathrm{L}}) = 66.327^2 = 4399.22\,.$$

The sums of squares for the main effects of factors A, B, C, and D can be obtained either by adding their respective linear and quadratic contrast sums of squares, or by using the rules of Chap. 7 with $r = 1/3$ (since this is a one-third fraction). For example,

$$ssA = ss(A_{\mathrm{L}}) + ss(A_{\mathrm{Q}}) = 66.327^2 + 9.852^2 = 4399.22 + 97.07 = 4496.29\,,$$

or

$$ssA = 9\Sigma \bar{y}_{i\dots}^2 - 27\bar{y}_{\dots}^2 = 28766.30 - 24270.01 = 4496.29\,. \qquad \square$$

Table 15.14 Normalized contrast estimates for the refinery experiment

	$\hat{A}_L$	$\hat{A}_Q$	$\hat{B}_L$	$\hat{B}_Q$	$\hat{C}_L$	$\hat{C}_Q$	$\hat{D}_L$	$\hat{D}_Q$
Estimate	66.33	9.85	51.45	-11.01	74.27	-1.95	14.61	-8.36

Fig. 15.6 Half-normal probability plot for the refinery experiment contrast absolute estimates
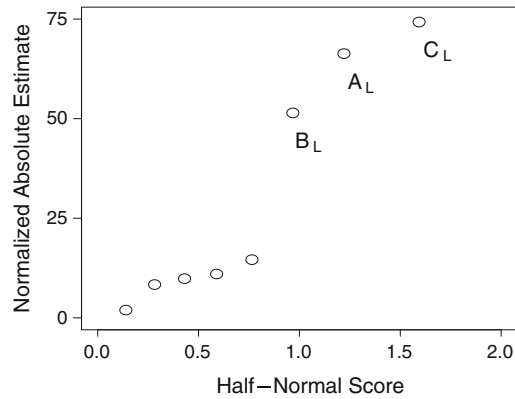
15.3.2   One-Ninth Fractions of 3^p Experiments; 3^{p-2} Experiments

As an example of a $\frac{1}{9}$-fraction, we take the sixth block of the 3^4 single-replicate confounded design shown in Table 14.5. The list of confounded interactions in the block design provides the list of interactions in the defining relation for the 3^{4-2} fractional factorial design. In the block design of Table 14.5, the confounded contrasts are $(AB^2C; A^2BC^2)$, $(ABD; A^2B^2D^2)$, $(AC^2D^2; A^2CD)$, and $(BCD^2; B^2C^2D)$ so, in the fraction, which consists of 9 treatment combinations, the defining relation is

$$
\begin{aligned}
I \quad &= AB^2C \ = A^2BC^2 \\
= ABD \ &= A^2CD \ = B^2C^2D \\
= A^2B^2D^2 \ &= BCD^2 \ = AC^2D^2 \,.
\end{aligned}
$$

This design has Resolution III (since the shortest word has 3 letters), and main-effect contrasts will be aliased with 2-factor interaction contrasts. The nine observations provide 8 degrees of freedom, which is sufficient to estimate the four main effects (with two degrees of freedom each). Therefore, the design would be useful if all two-factor interactions were believed to be negligible. Since there are no degrees of freedom available for estimating σ^2, a half-normal probability plot of normalized contrast estimates would be drawn as was done in Fig. 15.6.

15.4   Fractions from Block Designs; Other Experiments

15.4.1 $2^p \times 4^q$ Experiments

The simplest way to design a fractional factorial experiment when all factors have four levels, or when some factors have two levels and the others have four levels, is to use pseudofactors. For example, suppose we require a design for a $2^3 \times 4$ experiment with eight observations. A design in four blocks of size 8 is shown in Table 14.9 (p. 483). The confounded contrasts are FGJ_1J_2, GHJ_2, and FHJ_1, where J_1 and J_2 are the two 2-level pseudofactors making up the 4-level factor J. Suppose Block I is

Table 15.15 Aliasing scheme for a $\frac{1}{4}$-fraction of a $2^3 \times 4$ experiment

I	$=$	FGJ_1J_2	$=$	GHJ_2	$=$	FHJ_1
F	$=$	GJ_1J_2	$=$	$FGHJ_2$	$=$	HJ_1
G	$=$	FJ_1J_2	$=$	HJ_2	$=$	$FGHJ_1$
H	$=$	$FGHJ_1J_2$	$=$	GJ_2	$=$	FJ_1
J_1	$=$	FGJ_2	$=$	GHJ_1J_2	$=$	FH
J_2	$=$	FGJ_1	$=$	GH	$=$	FHJ_1J_2
J_1J_2	$=$	FG	$=$	GHJ_1	$=$	FHJ_2
FJ_2	$=$	GJ_1	$=$	FGH	$=$	HJ_1J_2

selected from the design to give a $\frac{1}{4}$-fraction of a $2^3 \times 4$ experiment, then the defining relation is

$$I = FGJ_1J_2 = GHJ_2 = FHJ_1$$

and the design is Resolution III. The aliasing scheme, shown in Table 15.15, indicates that the F contrast, for example, is aliased with one contrast from each of the GJ, $FGHJ$ and HJ interactions. There are 3 contrasts (J_1, J_2 and J_1J_2) for the 4-level factor J. The aliasing scheme shows that J is aliased with contrasts from the FGJ, GHJ, FH, GH, FHJ, FG interactions.

An experiment involving pseudofactors will be illustrated in Example 15.5.1 in Sect. 15.5.

15.4.2 $2^p \times 3^q$ Experiments

Suppose that a $\frac{1}{6}$-fraction of a $2^3 \times 3^3$ experiment is required, which has a total of 36 observations. Again, we follow the idea of selecting one block from a block design. So we first need a block design in $b = 6$ blocks, each of size 36. Following the procedure of Example 14.4.1, p. 484, we select, for example, (i) a block design in two blocks of size 4 from the 2^3 experiment with factors A, B, C, confounding the ABC interaction, and (ii) a block design in three blocks of size 9 from the 3^3 experiment with factors E, F, G, confounding the pair of contrasts $(DE^2F; D^2EF^2)$. Combining the treatment combinations in the blocks of these designs as in Example 14.4.1 leads to a design with six blocks in which the five contrasts ABC, $(DE^2F; D^2EF^2)$, and $(ABCDE^2F; ABCD^2EF^2)$ are confounded. Then, if one of the six blocks is selected for the $\frac{1}{6}$-fraction, we have a Resolution III design with defining relation

$$I = ABC = DE^2F = D^2EF^2 = ABCDE^2F = ABCD^2EF^2 \,,$$

and the contrasts ABC, $(DE^2F; D^2EF^2)$, and $(ABCDE^2F; ABCD^2EF^2)$ are aliased with the mean.

The aliasing scheme for the fraction has 36 rows and includes the following three rows:

$$A = BC = ADE^2F = AD^2EF^2 = BCDE^2F = BCD^2EF^2 \,,$$
$$B = AC = BDE^2F = BD^2EF^2 = ACDE^2F = ACD^2EF^2 \,,$$
$$C = AB = CDE^2F = CD^2EF^2 = ABDE^2F = ABD^2EF^2 \,.$$

Thus, the 2-level factors A, B, and C are aliased with 2-factor interactions between the 2-level factors plus some higher-order interactions. For example, the A contrast is aliased with the contrasts BC, $(ADE^2F; AD^2EF^2)$, and $(BCDE^2F; BCD^2EF^2)$.

A similar aliasing happens for the 3-level factors. For example,

$$D \; = ABCD \; = D^2E^2F = EF^2 \;\; = ABCD^2E^2F = ABCEF^2 \,,$$
$$D^2 = ABCD^2 = E^2F \;\;\; = DEF^2 = ABCE^2F \;\;\; = ABCDEF^2 \,,$$

so the pairs of contrasts $(D; D^2)$, $(ABCD; ABCD^2)$, $(DEF^2; D^2E^2F)$, $(EF^2; E^2F)$, and $(ABCEF^2; ABCE^2F)$ are aliased with one another.

Finally, there is aliasing of interactions involving both 2- and 3-level factors, for example

$$AD \; = BCD \; = AD^2E^2F = AEF^2 \;\;\; = BCD^2E^2F = BCEF^2 \,,$$
$$AD^2 = BCD^2 = AE^2F \;\;\; = ADEF^2 = BCE^2F \;\;\;\; = BCDEF^2 \,,$$

so the pairs of contrasts $(AD; AD^2)$, $(BCD; BCD^2)$, $(ADEF^2; AD^2E^2F)$, $(AEF^2; AE^2F)$, and $(BCEF^2; BCE^2F)$ are aliased with one another.

The design would be useful mainly when most of the interactions were expected to be negligible.

15.5   Blocked Fractional Factorial Experiments

If experimental conditions are not constant over the entire experiment, it may be necessary to arrange a fractional factorial experiment in blocks. For example, consider the soup experiment in Sect. 15.2.3 (p. 499), for which the experimenters used the resolution V 2^{5-1} fraction with defining relation $I = ABCDE$. Suppose the experimenters had decided that the experimental conditions could be kept fairly stable over the course of 8 observations but not 16. The treatment combinations would then have been divided into two blocks of size 8. If the fraction is divided into $b = 2$ blocks, then $b - 1 = 1$ contrast *and its alias* must be confounded. If CDE, for example, is selected for confounding, then the aliased pair of contrasts $CDE = AB$ is confounded with blocks, and neither of these contrasts can be measured. Rather than confound a 2-factor interaction, an alternative might be to select the Resolution IV design with defining relation $I = ABDE$ and to confound the aliased pair of contrasts $BCE = ACD$. Then, all two-factor interactions can be estimated, although six of them will be in aliased pairs. The choice between these two designs is the choice of losing information on one 2-factor interaction completely while aliasing the others with high-order interactions, or aliasing three pairs of 2-factor interactions.

For each fractional factorial design listed Table 15.60, at the end of the chapter, a suggestion (shown in parentheses) is given for selecting an interaction to be confounded when running the corresponding fraction in two blocks. (The aliases of this interaction can be obtained by multiplication with the interaction names in the defining relation, as usual).

Example 15.5.1 `Flour experiment`

M.G. Tuck, S.M. Lewis, and J.I.L. Cottrell describe a series of four experiments in the 1993 issue of *Applied Statistics* that were carried out at Spillers Milling Ltd. in order to identify a flour that would give a "high loaf volume and be tolerant to fluctuations in the bread making process". In the the third experiment in the series, four flour formulations were investigated (four levels of factor A, coded 0, 1, 2, 3), together with four noise factors each at two levels. The noise factors were amount of yeast (factor N, low or high), proof time (factor S, short or long), degree of mixing and moulding (factor Q, "undermixing, little water, heavy pressure" or "overmixing, much water, little pressure"), and dough time delay (factor T, short or long). Thus, this was a 4×2^4 experiment. A $\frac{1}{2}$ fraction with $v = 32$ treatment combinations was selected and divided into two blocks of size 16, representing the number of observations that could be taken per day.

The 4-level factor A can be written in terms of two pseudofactors A_1 and A_2, with the level correspondence $0 = 00$, $1 = 01$, $2 = 10$, $3 = 11$. The researchers selected the first block of the

Table 15.16 A blocked $\frac{1}{2}$-fraction of a 4×2^4 experiment for the flour experiment

Block (day)	Treatment combination	Av. specific volume	A_1	A_2	A_1A_2	N	S	Q	T
I	000011	436	-1	-1	1	-1	-1	1	1
	000110	507	-1	-1	1	-1	1	1	-1
	001001	434	-1	-1	1	1	-1	-1	1
	001100	508	-1	-1	1	1	1	-1	-1
	010010	436	-1	1	-1	-1	-1	1	-1
	010111	508	-1	1	-1	-1	1	1	1
	011000	404	-1	1	-1	1	-1	-1	-1
	011101	510	-1	1	-1	1	1	-1	1
	100010	440	1	-1	-1	-1	-1	1	-1
	100111	517	1	-1	-1	-1	1	1	1
	101000	442	1	-1	-1	1	-1	-1	-1
	101101	501	1	-1	-1	1	1	-1	1
	110011	458	1	1	1	-1	-1	1	1
	110110	536	1	1	1	-1	1	1	-1
	111001	464	1	1	1	1	-1	-1	1
	111100	532	1	1	1	1	1	-1	-1
II	000000	567	-1	-1	1	-1	-1	-1	-1
	000101	549	-1	-1	1	-1	1	-1	1
	001010	391	-1	-1	1	1	-1	1	-1
	001111	418	-1	-1	1	1	1	1	1
	010001	458	-1	1	-1	-1	-1	-1	1
	010100	499	-1	1	-1	-1	1	-1	-1
	011011	381	-1	1	-1	1	-1	1	1
	011110	451	-1	1	-1	1	1	1	-1
	100001	499	1	-1	-1	-1	-1	-1	1
	100100	483	1	-1	-1	-1	1	-1	-1
	101011	368	1	-1	-1	1	-1	1	1
	101110	456	1	-1	-1	1	1	1	-1
	110000	475	1	1	1	-1	-1	-1	-1
	110101	597	1	1	1	-1	1	-1	1
	111010	414	1	1	1	1	-1	1	-1
	111111	452	1	1	1	1	1	1	1

Sources Tuck et al. (1993). Copyright © 1993 Blackwell Publishers. Reprinted with permission

$\frac{1}{2}$-fraction with defining relation $I = A_1A_2NSQT$. The aliased pair of contrasts selected for confounding to provide two blocks were $NQ = A_1A_2ST$. Provided that the 16 combinations of design and noise factor levels were observed in a completely random order in each block, we can analyze this as an ordinary blocked fractional factorial experiment, without distinguishing between design and noise factors. The distinction between these two types of factors in made in Example 15.7.1, where flour sensitivity to the noise variable fluctuations is investigated.

The treatment combinations in each block are shown, prior to randomization, in Table 15.16, together with the main-effect contrast coefficients. All of the contrasts, apart from the confounded contrasts NQ and A_1A_2ST, are orthogonal to the block contrast, and the estimates of their aliased pairs can be

Table 15.17   Contrast estimates (with divisor 16) for the flour experiment

$\widehat{A_1}$	$\widehat{A_2}$	$\widehat{A_1A_2}$	$\widehat{N}$	$\widehat{S}$	$\widehat{Q}$	$\widehat{T}$
11.06	3.69	24.06	−52.44	59.81	−47.06	0.56

$\widehat{A_1N}$	$\widehat{A_2N}$	$\widehat{A_1A_2N}$	$\widehat{A_1S}$	$\widehat{A_2S}$	$\widehat{A_1A_2S}$	
5.44	7.56	−11.56	4.44	14.56	−2.31	

$\widehat{A_1Q}$	$\widehat{A_2Q}$	$\widehat{A_1A_2Q}$	$\widehat{A_1T}$	$\widehat{A_2T}$	$\widehat{A_1A_2T}$	
3.06	9.19	−17.19	9.19	9.56	−15.81	

Table 15.18   Analysis of variance for the flour experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Block	1	957.03	−		
A	3	5719.84	1906.62	3.06	0.0737
A_1	1	979.03	979.03	1.57	0.2363
A_2	1	108.78	108.78	0.17	0.6843
A_1A_2	1	4632.03	4632.03	7.42	0.0198
N	1	21997.53	21997.53	35.26	0.0001
S	1	28620.28	28620.28	45.88	0.0001
Q	1	17719.03	17719.03	28.40	0.0002
T	1	2.53	2.53	0.00	0.9504
AN	3	1763.59	587.86	0.94	0.4533
AS	3	1896.84	632.28	1.01	0.4235
AQ	3	3113.59	1037.87	1.66	0.2318
AT	3	3407.09	1135.69	1.82	0.2017
Error	11	6862.34	623.85		
Total	31	92059.72			

calculated without block adjustments. The response variable for each treatment combination was the average specific volume of three loaves, and the observed values (ml/100g) are listed in Table 15.16.

Multiplying the responses by the A_1 contrast coefficients and then dividing by $v/2 = 16$, we obtain the estimate of the difference in the effect of the high and low levels of A_1. Translating back to the original levels of factor A, this contrast compares the average of the flours 2 and 3 with the average of flours 0 and 1. The contrast estimate is $(7634 − 7457)/16 = 11.0625$ units.

The contrast for the interaction of the noise variable N with the pseudofactor A_1 is obtained by multiplying the A_1 and N contrast coefficients in Table 15.16 and dividing by $v/2 = 16$ to obtain the same standard error as the main-effect contrasts. Thus the contrast has coefficients

$$[\, 1,\ \ 1, -1, -1,\ \ 1,\ \ 1, -1, -1, -1, -1,\ \ 1,\ \ 1, -1, -1,\ \ 1,\ \ 1, \\ 1,\ \ 1, -1, -1,\ \ 1,\ \ 1, -1, -1, -1, -1,\ \ 1,\ \ 1, -1, -1,\ \ 1,\ \ 1\,]$$

with divisor 16, and the contrast estimate is $(7589 − 7502)/16 = 5.4375$. All of the main-effect and 2-factor interaction contrast estimates are shown in Table 15.17, and the analysis of variance table is shown in Table 15.18.

Selecting individual significance levels of $\alpha^* = 0.01$ for each hypothesis test, (for an overall Type I error probability of at most $\alpha = 0.09$), we compare the F-ratios in Table 15.18 with either

$F_{1,11,0.01} = 9.65$ or $F_{3,11,0.01} = 6.22$ as appropriate. The interactions of the flour formulations with the noise variables are not significantly different from zero, but the noise variables N, S, Q themselves do have a large effect on the specific volume. Although the flours are not significantly different in terms of average specific volume, the contrast A_1A_2 appears to be the most important of the three flour contrasts investigated. This contrast compares the average of flours 0 and 3 with the average of flours 1 and 2. The first pair give the higher average specific volume. Before the experiment took place, the experimenters had expected flour 3 (coded 11) to be the best. The difference of averages contrast, which compares flour 3 with the average of the other three flours, has least squares estimate

$$\begin{aligned} \bar{y}_{11...} &- \tfrac{1}{3}(\bar{y}_{00...} + \bar{y}_{01...} + \bar{y}_{10...}) \\ &= 491.000 - \tfrac{1}{3}(476.250 + 455.875 + 463.250) \\ &= 25.875 \, . \end{aligned}$$

A preplanned 95% confidence interval for this contrast is given by

$$\begin{aligned} \bar{y}_{11...} - \frac{1}{3}(\bar{y}_{00...} + \bar{y}_{01...} + \bar{y}_{10...}) &\pm t_{11,0.025}\sqrt{mse\left(\frac{1}{8} + 3\left(\frac{1}{9}\right)\left(\frac{1}{8}\right)\right)} \\ &= 25.875 \pm 2.201\sqrt{(623.849)(0.1667)} \\ &= 25.875 \pm 22.443 \\ &= (3.432, 48.318). \end{aligned}$$

At the 95% confidence level, it does appear that flour 3 (coded 11) has specific volume between 3.4 and 48.3 units larger than the average of the other three flours. (We can draw this conclusion only because the contrast was preplanned. Otherwise, we would need to use Scheffé's method of multiple comparisons with $t_{11,0.025}$ replaced by $\sqrt{3F_{3,11,0.05}} = 3.24$, and the interval would include zero). □

15.6   Fractions from Orthogonal Arrays

15.6.1 2^p Orthogonal Arrays

The simplest type of orthogonal array is that shown in Table 15.19, consisting of a set of $2^p - 1$ orthogonal contrasts. The first column has the first half of its 8 entries equal to -1 and the second half equal to $+1$. The second column has the first quarter of its entries equal to -1, the second quarter equal to $+1$, the third quarter equal to -1 again and the fourth quarter equal to $+1$ again. The third column is divided into eighths with alternating -1's and $+1$'s. If the columns had been longer, the next column would have been divided into sixteenths, and so on. These are the "independent" columns. The fourth, fifth and sixth columns of Table 15.19 are the products of corresponding coefficients in the first three columns taken in pairs, and the last column is the triple product of the first three columns. The result is a table with $2^p = 8$ rows and $2^p - 1 = 7$ columns in which any pair of columns are orthogonal.

The p independent columns of an orthogonal array define the treatment combinations for a 2^p design. As usual, a contrast coefficient of -1 in a column corresponds to level 0 in the corresponding factor, and a contrast coefficient of $+1$ in a column corresponds to level 1. If all eight treatment combinations of Table 15.19 are used for a 2^3 experiment, then the orthogonal array defines a full factorial experiment, and no aliasing of contrasts occurs.

Now suppose that only 4 observations can be taken in a 2^3 experiment. Instead of proceeding as in Sect. 15.2.1 and choosing a defining relation, we could first construct an orthogonal array with $n = 4$

Table 15.19 Contrasts for a 2^3 experiment

	A	B	C	AB	AC	BC	ABC
000	-1	-1	-1	1	1	1	-1
001	-1	-1	1	1	-1	-1	1
010	-1	1	-1	-1	1	-1	1
011	-1	1	1	-1	-1	1	-1
100	1	-1	-1	-1	-1	1	1
101	1	-1	1	-1	1	-1	-1
110	1	1	-1	1	-1	-1	-1
111	1	1	1	1	1	1	1

Table 15.20 An orthogonal array for four observations

-1	-1	1
-1	1	-1
1	-1	-1
1	1	1

Table 15.21 $\frac{1}{2}$-fractions of a 2^3 experiment obtained from orthogonal arrays

	Design d_1				Design d_2		
TC	A	B	C	TC	A	B	C
001	-1	-1	1	011	-1	1	1
010	-1	1	-1	000	-1	-1	-1
100	1	-1	-1	110	1	1	-1
111	1	1	1	101	1	-1	1

rows and $n - 1 = 3$ columns. One is shown in Table 15.20, where the first column has the first half of its entries -1, and the second half $+1$, the second column is divided into quarters, and the third column is the product of the first two. Since we have 3 factors, suppose we label the columns in order as A, B, C. The three columns then show the parts of the A, B, and C contrasts corresponding to a $\frac{1}{2}$-fraction. However, the third column is also the product of the first two columns, so it not only represents the C contrast but also the interaction between A and B. Consequently, C is aliased with AB. Similarly, the first column is the product of the last two columns, so A is aliased with BC. Similarly, again, B is aliased with AC. The defining relation must be $I = ABC$ in order to produce this aliasing scheme.

The coefficients in the contrasts tell us which treatment combinations are represented, and the design is "Design d_1" shown in Table 15.21. Notice that this is the same design that would have been produced from the equation $a_1 + a_2 + a_3 = 1 \pmod 2$. We could obtain the $\frac{1}{2}$-fraction corresponding to $a_1 + a_2 + a_3 = 0 \pmod 2$, by multiplying any one of the columns by -1 (see Design d_2 in Table 15.21, where the second column has been multiplied by -1).

Thus, we have arrived back at the same type of design that we studied in Sect. 15.2, and this will often (but not always) be the case. The main difference in procedure is that when we start with an orthogonal array, we are starting with an unlabeled list of contrasts which can be labeled in any way we please. The labeling then determines the defining relation and the design.

Any columns in an orthogonal array can be multiplied by -1 and we still obtain an orthogonal array, although the treatment combinations may not be identical, or they may be identical but in a different order (try multiplying the B and C columns for the designs in Table 15.21 by -1 and see whether the same design results).

Now we return to the orthogonal array of Table 15.19, which is reproduced in Table 15.22 with column headings indicating which columns are products of which other columns. For example, column

Table 15.22 An orthogonal array for 8 observations

			Columns			
1	2	3	12	13	23	123
−1	−1	−1	1	1	1	−1
−1	−1	1	1	−1	−1	1
−1	1	−1	−1	1	−1	1
−1	1	1	−1	−1	1	−1
1	−1	−1	−1	−1	1	1
1	−1	1	−1	1	−1	−1
1	1	−1	1	−1	−1	−1
1	1	1	1	1	1	1

7 is the product of columns 1, 2, and 3. We consider using this array for a 2^5 experiment instead of a 2^3 experiment. Since there are only 8 rows, we will be looking for a $\frac{1}{4}$-replicate (that is, a 2^{5-2} fractional factorial experiment).

Suppose that we label the first 5 columns as A, B, C, D, E. Since the product of the first two columns gives column 4 and the product of the first and third columns gives column 5, aliasing would occur between D and AB, and between E and AC. Consequently, ABD and ACE must be in the defining relation, together with their product, so we have

$$I = ABD = ACE = BCDE .$$

The rest of the aliasing scheme can be written out also, and we would see that A is aliased with BD and CE, that B is aliased with AD, and that C is aliased with AE. The sixth column, which is the product of columns 2 and 3, and also of columns 4 and 5, can be labeled BC or DE, and these two interactions are aliased. The seventh column is $ABC = CD = BE = ADE$. The eight treatment combinations are deduced from the −1's and +1's in the first five columns; that is,

$$00011, \ 00110, \ 01001, \ 01100, \ 10000, \ 10101, \ 11010, \ 11111 .$$

Different sets of treatment combinations corresponding to the same defining relation (but with different signs in the aliasing scheme) can be obtained by multiplying one or more columns of Table 15.22 by −1.

There is nothing special about labeling the first five columns of Table 15.22 as A, B, C, D, E. Any five columns could have been chosen. Different choices may lead to different aliasing schemes, and sometimes these aliasing schemes may not be equally good. Table 15.62 at the end of the chapter lists orthogonal arrays for various-sized experiments. Some useful column labelings for various fractional factorial experiments are suggested in the table.

The standard notation, used by industrial statisticians and engineers, for an orthogonal array is the letter L with subscript equal to the number of runs. Sometimes, the largest Resolution III design that can be used with the array is added in brackets. The orthogonal array in Table 15.20 provides a Resolution III design with 4 observations for 3 or fewer two-level factors and would be written as $L_4(2^3)$. Similarly, the design of Table 15.22 would be written as $L_8(2^7)$. Occasionally, an orthogonal array for 2^p experiments will be written using factor levels rather than contrast coefficients. The orthogonality could then be checked by ensuring that in every pair of columns, all possible pairs of factor levels (00, 01, 10, and 11) appear the same number of times (see, for example, the factor levels shown together with designs d_1 and d_2 of Table 15.21).

Table 15.23   Treatment factors and their levels for the wafer experiment

Factors	Prior level	Experimental levels	
		Low (0)	High (1)
A (rotation method)	Oscillating	Continuous	Oscillating
B (wafer batch)		668G4	678D4
C (deposition temperature)	1215 °C	1210 °C	1220 °C
D (deposition time)	Low	High	Low
E (arsenic flow rate)	57%	55%	59%
F (acid etch temp.)	1200 °C	1180 °C	1215 °C
G (acid flow rate)	12%	10%	14%
H (nozzle position)	4	2	6

Example 15.6.1 `Wafer experiment`

R. Kackar and A. Shoemaker (*AT&T Technical Journal*, 1986) describe an experiment that they helped to run at AT&T to try to reduce the variability of the thickness of an "epitaxial layer" deposited onto silicon wafers during the manufacture of integrated circuit devices.

The wafers were mounted on a seven-sided "susceptor" with two wafers (one above the other) on each side. The susceptor rotated inside a heated bell jar as chemical vapors were introduced via a nozzle near the top of the jar. The chemicals were deposited on the wafers, and the bell jar was cooled when the thickness of the deposited layer was close to the target of $14.5\,\mu$m.

The engineers identified eight factors that might affect the variability of the thickness of the epitaxial layer. These are shown in Table 15.23 together with the operating factor levels prior to the experiment and the levels selected for the experiment.

The experimenters decided to take 16 observations. The 16 treatment combinations were selected via the orthogonal array $L_{16}(2^{15})$ shown in Table 15.24. The orthogonal array is constructed as described earlier in this section. The labels in the row headings of the table identify which columns are products of which other columns. The assignment of factors to columns chosen by the experimenters is indicated in the foot of the table. The experiment is a 2^{8-4} experiment, and the defining relation is generated by 4 confounded interactions. Notice, from the heading and the foot of Table 15.24, that D must be aliased with ABC, F must be aliased with ABE, G with ACE, and H with BCE. Thus, the defining relation includes $ABCD$, $ABEF$, $ACEG$, $BCEH$, and all their possible products (a total of $2^4 = 16$ terms in the defining relation):

$$
\begin{aligned}
I \quad &= ABCD \ = ABEF \ = CDEF \\
&= ACEG \ = BDEG \ = BCFG \ = ADFG \\
&= BCEH \ = ADEH \ = ACFH \ = BDFH \\
&= ABGH \ = CDGH \ = EFGH \ = ABCDEFGH
\end{aligned}
$$

This is a Resolution IV design, and there is considerable aliasing between 2-factor interactions. For example, the contrast listed in column labelled 12 in Table 15.24 not only measures the 2-factor interaction AB, but also measures its aliased 2-factor interactions CD, EF, GH (and some higher-order interactions).

There were 70 measurements taken for each treatment combination (5 measurements on each of the 2 wafers on the 7 sides of the receptor). From these, two different response variables were calculated—the average of the 70 measurements (which we denote by x) and the log sample variance of the 70

Table 15.24 An orthogonal array for 16 observations: An $L_{16}(2^{15})$

							Columns							
1	2	12	3	13	23	123	4	14	24	124	34	134	234	1234
−1	−1	1	−1	1	1	−1	−1	1	1	−1	1	−1	−1	1
−1	−1	1	−1	1	1	−1	1	−1	−1	1	−1	1	1	−1
−1	−1	1	1	−1	−1	1	−1	1	1	−1	−1	1	1	−1
−1	−1	1	1	−1	−1	1	1	−1	−1	1	1	−1	−1	1
−1	1	−1	−1	1	−1	1	−1	1	−1	1	1	−1	1	−1
−1	1	−1	−1	1	−1	1	1	−1	1	−1	−1	1	−1	1
−1	1	−1	1	−1	1	−1	−1	1	−1	1	−1	1	−1	1
−1	1	−1	1	−1	1	−1	1	−1	1	−1	1	−1	1	−1
1	−1	−1	−1	−1	1	1	−1	−1	1	1	1	1	−1	−1
1	−1	−1	−1	−1	1	1	1	1	−1	−1	−1	−1	1	1
1	−1	−1	1	1	−1	−1	−1	−1	1	1	−1	−1	1	1
1	−1	−1	1	1	−1	−1	1	1	−1	−1	1	1	−1	−1
1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	1	1	1	1
1	1	1	−1	−1	−1	−1	1	1	1	1	−1	−1	−1	−1
1	1	1	1	1	1	1	−1	−1	−1	−1	−1	−1	−1	−1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
A	B		C				D	E			F		G	H

Table 15.25 Treatment combinations and response variables for the wafer experiment

Treatment combination	Average response $x_{ijklmnpq}$	Log variance response $v_{ijklmnpq}$
0 0 0 0 0 0 0 0	14.821	−0.4425
0 0 0 0 1 1 1 1	14.888	−1.1989
0 0 1 1 0 0 1 1	14.037	−1.4307
0 0 1 1 1 1 0 0	13.880	−0.6505
0 1 0 1 0 1 0 1	14.165	−1.4230
0 1 0 1 1 0 1 0	13.860	−0.4969
0 1 1 0 0 1 1 0	14.757	−0.3267
0 1 1 0 1 0 0 1	14.921	−0.6270
1 0 0 1 0 1 1 0	13.972	−0.3467
1 0 0 1 1 0 0 1	14.032	−0.8563
1 0 1 0 0 1 0 1	14.843	−0.4369
1 0 1 0 1 0 1 0	14.415	−0.3131
1 1 0 0 0 0 1 1	14.878	−0.6154
1 1 0 0 1 1 0 0	14.932	−0.2292
1 1 1 1 0 0 0 0	13.907	−0.1190
1 1 1 1 1 1 1 1	13.914	−0.8625

measurements (which we denote by v). The treatment combinations, corresponding to the orthogonal array in Table 15.24, together with the two response variables, are shown in Table 15.25.

The experimenters first analyzed the log variance response. The contrast estimates (high level minus low level) for this response variable are shown in Table 15.26. The contrast estimates for factors A and H are considerably larger in absolute value than those for the other factors. Consequently, A and H

Table 15.26 Contrast estimates for log sample variance response variable

Contrast	A	B	C	D	E	F	G	H
Estimate	0.352	0.122	0.105	-0.249	-0.012	-0.072	-0.101	-0.566

Table 15.27 Contrast estimates for the mean response variable

Contrast	A	B	C	D	E	F	G	H
Estimate	-0.055	0.056	-0.109	-0.836	-0.067	0.060	-0.098	0.142

should be investigated for reducing variability in the response. Since the log variance response is to be reduced, and the contrast estimate for A is positive while that for H is negative, we would want to set A at its low level (continuous rotation) and H at its high level (position 6). All other factors can be set at their current operating conditions.

The second requirement of the experimenters was to achieve an average thickness of 14.5 μm. Contrast estimates for the average response are shown in Table 15.27. Not surprisingly, factor D, deposition time, has by far the largest effect on the mean response, and the experimenters were able to adjust this factor in order to meet the target.

As with any good experiment, the experimenters wished to confirm their results. Their confirmation experiment investigated two treatment combinations. The first treatment combination consisted of the prior operating levels of factors A and C–H, with factor B at level 1, and the second treatment combination was the same except that the levels of A and H were changed as discussed above. The confirmation experiment showed that the variance of the thickness had been reduced by a factor 2.5— quite a remarkable result! □

15.6.2 $2^p \times 4^q$ Orthogonal Arrays

The orthogonal arrays of Sect. 15.6.1 can be used when one or more factors have 4 levels. Each 4-level factor requires 3 independent columns to represent 3 orthogonal contrasts. For example, the orthogonal array in Table 15.22 could be used for a $2^3 \times 4$ experiment as follows. The first three columns (which are independent) could be labeled A, B, and C. If the 4th column is labeled D_1, then D_1 is aliased with AB. If the 7th column is labeled D_2, then D_2 is aliased with ABC. The product of the coefficients in the 4th and 7th columns gives the coefficients in the 3rd column, so D_1D_2, the remaining contrast for the 4-level factor, is aliased with C, and the defining relation is

$$I = ABD_1 = ABCD_2 = CD_1D_2 .$$

This is a Resolution II design, which should be avoided if possible, since it confounds two main effects (C and D). A better design is to assign D_2 to the 5th column, where it is aliased with AC. The product of the 4th and 5th columns gives the 6th column, so that D_1D_2 is aliased with BC. The defining relation is

$$I = ABD_1 = ACD_2 = BCD_1D_2 ,$$

which is Resolution III. The 7th column of Table 15.22 corresponds to the ABC contrast, which is aliased with AD_1D_2, CD_1, and BD_2 contrasts. The complete aliasing scheme is

$$
\begin{aligned}
I &= ABD_1 &&= ACD_2 &&= BCD_1D_2 \\
A &= BD_1 &&= CD_2 &&= ABCD_1D_2 \\
B &= AD_1 &&= ABCD_2 &&= CD_1D_2 \\
C &= ABCD_1 &&= AD_2 &&= BD_1D_2 \\
D_1 &= AB &&= ACD_1D_2 &&= BCD_2 \\
D_2 &= ABD_1D_2 &&= AC &&= BCD_1 \\
D_1D_2 &= ABD_2 &&= ACD_1 &&= BC \\
ABC &= CD_1 &&= BD_2 &&= AD_1D_2
\end{aligned}
$$

The design would be useful for an experiment where all interactions were expected to be negligible, in which case one degree of freedom would be available to estimate σ^2.

15.6.3 3^p Orthogonal Arrays

The orthogonal arrays for 2^p experiments introduced in Sect. 15.6.1 have the property that any pair of columns in the array are orthogonal (that is, the sum of the products of corresponding coefficients is zero). An examination of the arrays in Tables 15.21, 15.22 and 15.24 reveals that this orthogonality arises because every pair of coefficients $(-1, -1)$, $(-1, 1)$, $(1, -1)$ and $(1, 1)$ occurs equally often in every pair of columns. We could rewrite the array to contain the factor labels 0, 1 instead of the contrast coefficients -1, 1, and every pair of labels would occur the same number of times in every pair of columns. This is the way that orthogonal arrays are defined for 3^p experiments.

An orthogonal array with 9 treatment combinations is shown as columns 1–4 in Table 15.28 for four factors, each having 3 levels. If any pair of columns is selected, it can be verified that each of the nine pairs of levels $(0, 0)$, $(0, 1)$, $(0, 2)$, $(1, 0)$, $(1, 1)$, $(1, 2)$, $(2, 0)$, $(2, 1)$, $(2, 2)$ occurs once. The first column consists of three copies of each of 0, 1, and 2. The second column consists of 0, 1, and 2, in order, repeated three times. The third column is obtained from the sum of the coefficients in the first two columns reduced modulo 3 (thereby ensuring that any factor assigned to the 3rd column will be aliased with the interaction between the first two factors). The fourth column is obtained from twice the sum of columns 2 and 3 (ensuring that any factor assigned to the fourth column will be aliased with the interaction of factors assigned to columns 2 and 3 and with the interaction of the first two factors). It is not possible to find more than four orthogonal columns with only 9 observations.

Table 15.28   A 3^p orthogonal array for 9 observations

Columns				Contrasts							
1	2	3	4								
0	0	0	0	-1	1	-1	1	-1	1	-1	1
0	1	1	1	-1	1	0	-2	0	-2	0	-2
0	2	2	2	-1	1	1	1	1	1	1	1
1	0	1	2	0	-2	-1	1	0	-2	1	1
1	1	2	0	0	-2	0	-2	1	1	-1	1
1	2	0	1	0	-2	1	1	-1	1	0	-2
2	0	2	1	1	1	-1	1	1	1	0	-2
2	1	0	2	1	1	0	-2	-1	1	1	1
2	2	1	0	1	1	1	1	0	-2	-1	1
A	B	C	D	A_L	A_Q	B_L	B_Q	C_L	C_Q	D_L	D_Q

In Table 15.28, a pair of orthogonal contrasts is given corresponding to each of the four columns in the orthogonal array. It can be verified that this set of 8 contrasts is orthogonal. As for 2^p experiments, a 3^p orthogonal array with n rows can have at most $n - 1$ orthogonal columns of contrast coefficients, and therefore can accommodate at most $(n - 1)/2$ three-level factors. An experiment is discussed in Sect. 15.7.1 that uses part of the orthogonal array for three-level factors and 27 observations listed in Table 15.65.

15.7   Design for the Control of Noise Variability

Design for the control of noise variability is sometimes known as *robust design* or *parameter design* and refers to the procedure of developing or designing a product in such a way that it performs consistently, and as intended, under the variety of conditions of its use throughout its life. The ideas apply equally well to the design of high quality manufacturing systems and other organizational processes. Experimentation involves *design factors* (also known as *control factors*) which are possible inexpensive to control in the design of the product, and *noise factors* which may affect the performance of a product but which are difficult or impossible to control when the product is in use.

Experiments that involve both design and noise factors are often known colloquially as *Taguchi experiments*. Dr. Taguchi was a Japanese quality consultant who advocated the use of quality improvement techniques, including the design of experiments, to the Japanese engineering and industrial communities from the 1950s. One of his fundamental contributions is the principle that reduction of variation is generally the most difficult task from an engineering perspective and so should be the focus of attention during the design of a product.

There are two different types of designs for such experiments—"mixed arrays" and "product arrays". *Mixed arrays* are ordinary fractional factorial designs in which the difference between the design and noise factors is ignored at the design stage except to ensure that the design-by-noise interactions are estimable and not confounded with blocks or part of the defining relation of a fraction. For a mixed array, all of the treatment combinations (composed of both design and noise factors) are observed in a random order. This complete randomization allows the design×noise interactions to be studied and to try to identify the particular levels of the design factors that are least affected by changing the levels of the noise factors. An example of a mixed array is the design used for the flour experiment of Example 15.5.1 which had a single design factor at 4 levels and four two-level noise factors. Analysis of the design×noise interactions for this experiment is illustrated in Example 15.7.1 assuming a complete randomization of the order of treatment combinations.

Product arrays, on the other hand, are designed and analyzed differently from the usual factorial experiments. They are composed of two fractional factorial or full factorial experiments, one for the design factors and one for the noise factors. Then every combination of design factors is observed in conjunction with every combination of noise factors. In product arrays, first the order of the design factor combinations is randomized. Then, for each design factor combination in turn, observations are taken across all of the noise factor combinations in a random or non-random order. (Occasionally, the order of randomization is reversed). This restricted randomization means that the usual analysis of design×noise interactions is not valid. Instead, for each design factor combination, the average and log sample variance of the responses are calculated across the different noise factor combinations. The average response and the log variance response are then taken as two totally separate sets of data values and analyzed separately. The objective of the experiment is to find out which factors affect the log sample variance response the most, and which factors most affect the average response. Design factor combinations are then sought that give a low variance across the noise factor combinations and also that give an average response close to the target value. Finally, confirmatory observations are taken.

(a) AQ interaction   (b) AT interaction
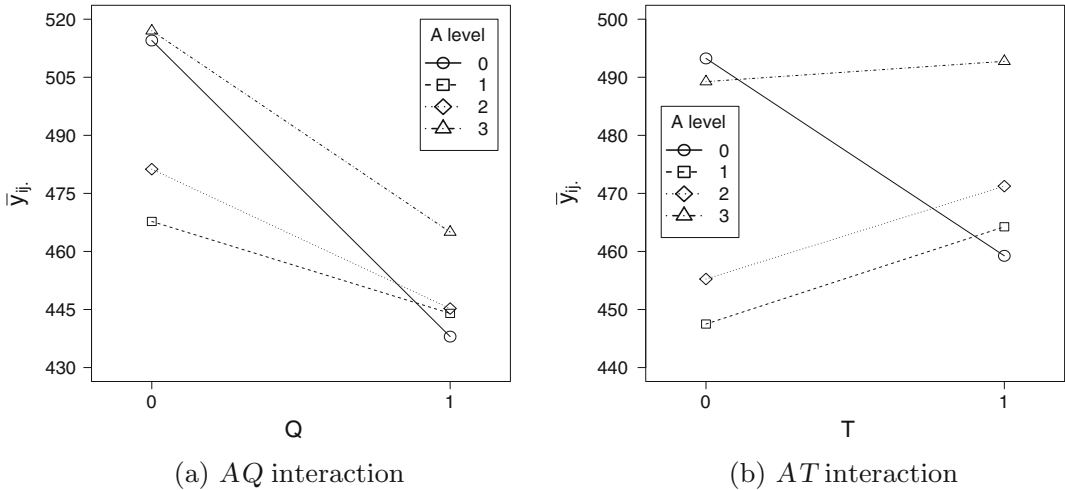
Fig. 15.7 Design×noise interaction plots for the flour experiment

The reason for taking the log of the sample variance before analysis is that the assumptions of the linear model with $\ln(s^2)$ as the response are more closely satisfied than taking s^2 itself as the response. (See Sect. 5.6.2 for a discussion of transformations).

The design of the wafer experiment of Example 15.6.1 was a product array, where the levels of the single noise factor "position" were the 14 locations on the susceptor and the two responses, log sample variance response and average response, were analyzed separately. An example of a product array with factors at three levels will be illustrated in Sect. 15.7.1 and the computer analysis discussed in Sects. 15.9.2 and 15.10.2 using SAS and R software, respectively.

Example 15.7.1 `Flour experiment, continued`

One purpose of the flour experiment in Example 15.5.1, p. 513, was to find which of the four flours (factor A, coded 0, 1, 2, 3) was least variable under the different levels of the four noise variables: amount of yeast (factor N), proof time (factor S), degree of mixing and moulding (factor Q), and dough time delay (factor T), each at 2 levels. If the treatment combinations in Table 15.16 were observed in a completely random order within each block, this experiment can be analyzed as a mixed array. The analysis of variance is shown in Table 15.18. The interactions of A with the noise variables can provide information on which flours are least sensitive to noise fluctuations.

Although none of the design×noise factor interactions were significantly different from zero, we illustrate the search for robust design factor levels by looking at the two largest such interactions (AQ and AT). If we plot the average response (specific volume) for each level of A versus the levels of Q or T on the horizontal axis, we obtain the interaction plots in Fig. 15.7. From the AT interaction plot, we can see that flours 1, 2, and 3 are much more stable than flour 0 in terms of the resulting average specific volume of loaves. This is also apparent, to a lesser extent, in the AQ interaction plot. Thus, flour 0 is not as robust as the other three flours and should probably be ruled out of consideration for general use.

If we now examine the average specific volume of loaves baked over the two levels of T, we see that flour 3 seems to be the best. a similar result is obtained by averaging over the levels of Q. In this experiment, the AQ and AT interactions were not significantly different from zero, but if they had been larger, this analysis would have pointed to flour 3 being preferable both in terms of robustness to fluctuating noise factors and of leading to loaves with high specific volume.

15.7.1   A Real Experiment—Inclinometer Experiment

A collaborative study involving statisticians and mechanical engineers was described by S. Lewis, B. Hodgson, R. New, and C. Sexton in the 1989 *Proceedings of the Institute of Mechanical Engineers International Conference on Engineering Design*. The experiment sought to improve the performance of an inclinometer, which is an instrument that records the angle of tilt of an object such as a crane jib. The design of the inclinometer is described in the article as follows:

"The basic design of the product is composed in four parts: a bob-weight and flexure, a flanged flywheel and a copper-plated disc (PCB). All are attached to a shaft supported in low-friction bearings. When the object to which the flywheel is attached is tilted, the bob-weight assembly moves to stay perpendicular to the earth, causing the PCB to rotate relative to the casing. The main performance difficulty of the inclinometer is that it does not immediately register the true angle of tilt. Spurious swing of the disc is produced by movement of the object."

The purpose of the experiment was to vary the relative sizes of the parts of the inclinometer to find a combination of factors that would reduce the swing. The engineers identified 7 design factors (A–G) that could be altered and that might affect the swing. Three levels were selected for each factor so that linear and quadratic trends could be investigated. The levels of factors A–F were equally spaced. The factors were:

A: Flexure length (30.00, 31.25, 32.5) B: Flexure thickness (0.05, 0.275, 0.5)
C: Flexure width (4.0, 5.0, 6.0) D: Flange thickness (1.0, 3.5, 6.0)
E: Flange width (6.0, 10.5, 15.0) F: Bob-weight length (12.0, 20.0, 28.0)
                    G: Copper plating thickness (0.0175, 0.035, 0.07)

All measurements are in millimeters, and the levels of all factors are coded 0, 1, and 2. For the experiment, it was possible to produce the factor levels exactly as specified, but in mass production variability naturally creeps in. The experimenters decided to build the production variability into the experiment as noise factors as follows (measured in mm, except where stated):

H: Flexure length (-0.25, $+0.25$) P: Flexure thickness (-0.005, $+0.005$)

J: Flange thickness (-0.025, $+0.025$) K: Flange width (-0.025, $+0.025$)

                    L: Copper plating thickness (-0.005, $+0.005$)

            M: Tolerance on bob weight mass (-9.0, $+9.0 \times (1/100)g$)

                N: Maximum horizontal amplitude of vibration (5, 25)

The two levels of each noise factor were coded as 0 and 1. Thus, the entire experiment was a $3^7 \times 2^7$ factorial experiment, where the 3-level factors were the design factors and the 2-level factors were the noise factors. The treatment combination 0000000 of the design factors in conjunction with the combination 0000000 of the noise factors would have flexure length (A) of $(30.00 - 0.25)$ mm $=$ 29.75 mm, flexure thickness of $(0.050 - 0.005)$ mm $= 0.045$ mm, and so on.

The objective of the experiment was to select the combinations of the design factors that gave the least amount of swing. In terms of producing a product of consistently high quality, it was also important that the variability of the amount of swing also remain low across the different noise combinations.

The experimenters selected a product array formed from a $\frac{1}{81}$ fraction of the 3^7 design-treatment combinations and a $\frac{1}{16}$ fraction of the 2^7 noise combinations. This gave a total of $27 \times 8 = 216$ observations. For the 3^{7-4} fractional factorial experiment, seven columns of the orthogonal array $L_{27}(3^{13})$ were selected. These are indicated in Table 15.65 (p. 562). For the 2^{7-4} fractional factorial experiment, the orthogonal array $L_8(2^7)$ shown in Table 15.22 (p. 518) was used with the noise factors

Table 15.29 Maximum angle of swing for the inclinometer experiment. Combinations of design factors *A–G* are in rows, and combinations of noise factors *H–N* are in columns

Noise factors	H	0	0	0	0	1	1	1	1		
	P	0	0	1	1	0	0	1	1		
	J	0	0	1	1	1	1	0	0		
	K	0	1	0	1	0	1	0	1		
	L	0	1	0	1	1	0	1	0		
	M	0	1	1	0	0	1	1	0		
	N	0	1	1	0	1	0	0	1		
Design factors											
ABCDEFG										Mean	$\ln(s^2)$
0 0 0 0 0 0 0		0.62	3.54	3.56	0.62	3.09	0.71	0.73	3.20	2.01	0.73
0 0 1 1 1 1 1		0.59	3.11	3.11	0.59	2.98	0.63	0.64	3.02	1.83	0.53
0 0 2 2 2 2 2		0.59	3.01	3.02	0.59	2.97	0.61	0.62	3.00	1.80	0.50
0 1 0 1 1 2 2		0.51	2.65	2.65	0.50	2.53	0.53	0.54	2.56	1.56	0.21
0 1 1 2 2 0 0		0.18	0.96	0.96	0.18	0.89	0.19	0.20	0.90	0.56	−1.85
0 1 2 0 0 1 1		1.88	9.58	9.55	1.85	9.30	1.92	1.94	9.48	5.69	2.80
0 2 0 2 2 1 1		0.19	1.03	1.03	0.19	0.97	0.21	0.21	0.93	0.60	−1.72
0 2 1 0 0 2 2		1.85	9.46	9.42	1.82	9.19	1.90	1.92	9.35	5.61	2.77
0 2 2 1 1 0 0		0.52	2.73	2.72	0.51	2.61	0.55	0.56	2.64	1.61	0.27
1 0 0 1 2 1 2		0.29	1.56	1.56	0.29	1.45	0.31	0.32	1.47	0.91	−0.87
1 0 1 2 0 2 0		0.95	4.98	4.93	0.94	4.79	0.99	1.00	4.82	2.93	1.48
1 0 2 0 1 0 1		1.16	6.09	6.09	1.13	5.70	1.21	1.26	5.93	3.57	1.87
1 1 0 2 0 0 1		0.26	1.45	1.45	0.25	1.30	0.29	0.30	1.30	0.83	−1.05
1 1 1 0 1 1 2		1.15	5.99	5.92	1.13	5.69	1.19	1.22	5.84	3.51	1.84
1 1 2 1 2 2 0		0.85	4.31	4.30	0.84	4.23	0.86	0.88	4.28	2.57	1.21
1 2 0 0 1 2 0		1.10	5.74	5.67	1.07	5.43	1.14	1.17	5.57	3.36	1.75
1 2 1 1 2 0 1		0.29	1.55	1.55	0.28	1.45	0.31	0.32	1.47	0.90	−0.88
1 2 2 2 0 1 2		0.91	4.64	4.66	0.90	4.56	0.94	0.95	4.57	2.77	1.35
2 0 0 2 1 2 1		0.39	2.05	2.06	0.39	1.96	0.41	0.42	1.97	1.21	−0.30
2 0 1 0 2 0 2		0.67	3.61	3.57	0.65	3.27	0.72	0.74	3.41	2.08	0.79
2 0 2 1 0 1 0		1.42	7.31	7.38	1.41	7.14	1.48	1.51	7.24	4.36	2.27
2 1 0 0 2 1 0		0.69	3.66	3.60	0.67	3.37	0.73	0.74	3.47	2.12	0.82
2 1 1 1 0 2 1		1.18	6.04	6.06	1.17	5.90	1.21	1.23	5.95	3.59	1.88
2 1 2 2 1 0 2		0.37	1.95	1.95	0.37	1.87	0.39	0.40	1.88	1.15	−0.40
2 2 0 1 0 0 2		0.39	2.15	2.16	0.38	1.94	0.44	0.44	1.96	1.23	−0.25
2 2 1 2 1 1 0		0.44	2.29	2.29	0.43	2.21	0.46	0.47	2.22	1.35	−0.07
2 2 2 0 2 2 1		1.84	9.35	9.19	1.79	9.06	1.85	1.89	9.28	5.53	2.74

Sources Lewis, Hodgson, New, and Sexton (1989). Copyright © 1989 Mechanical Engineering Publications. Reprinted with permission

Table 15.30   Analysis of variance for $\ln(s^2)$ response for the inclinometer experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
A	2	0.6316	0.3158		
Linear A	1	0.5798	0.5798	22.73	0.0005
Quadratic A	1	0.0519	0.0519	2.03	0.1794
B	2	0.1358	0.0679		
Linear B	1	0.0581	0.0581	2.28	0.1571
Quadratic B	1	0.0777	0.0777	3.05	0.1064
C	2	9.8448	4.9224		
Linear C	1	9.8241	9.8241	385.18	0.0001
Quadratic C	1	0.0207	0.0207	0.81	0.3852
D	2	18.8987	9.4493		
Linear D	1	18.3769	18.3769	720.53	0.0001
Quadratic D	1	0.5217	0.5217	20.46	0.0007
E	2	7.0366	3.5183		
Linear E	1	7.0044	7.0044	274.63	0.0001
Quadratic E	1	0.0322	0.0322	1.26	0.2829
F	2	9.5150	4.7575		
Linear F	1	9.4043	9.4043	368.73	0.0001
Quadratic F	1	0.1106	0.1106	4.34	0.0593
G	2	0.0354	0.0177	0.69	0.5184
Error	12	0.3061	0.0255		
Total	26	46.4039			

Table 15.31   Contrast estimates (log var response) for the nonnegligible contrasts

Lin A	Lin C	Lin D	Quad D	Lin E	Lin F
0.359	1.478	−2.021	0.590	−1.248	1.446

assigned to the columns in the order $H, P, K, -J, -L, -M, N$, where the minus signs indicate that the column was multiplied by -1 (thus reversing the high and low levels).

The maximum absolute angle of swing was ascertained for each of the selected combinations of design- and noise- factor levels, and these are shown in Table 15.29. The noise-factor combinations label the columns, and the design-factor combinations label the rows. The last two columns of the table show the average and log sample variance of the observations for the design combinations calculated across the noise combinations.

Consider first using the log sample variance $\ln(s^2)$ of the observations as the response variable. The analysis of variance table is shown in Table 15.30. We have included the information needed for testing the hypotheses of negligible linear and quadratic trends in each of the factors except for G. The levels of G are not equally spaced, and therefore the correct trend contrast coefficients are not those shown in Table A.2.

If we test the hypotheses of negligible contrasts for each trend contrast shown in Table 15.30 at individual significance levels $\alpha^* = 0.01$ and test the hypothesis of no effect of factor G at level $\alpha^* = 0.01$ (for an overall level of at most $\alpha = 0.13$), we reject the hypotheses of negligible linear trends in factors A, C, D, E, F and of a negligible quadratic trend in factor D. Factors B and G show very little effect on log variance response, so these factors (flexure thickness and copper plating thickness) cannot be employed to achieve less variability in the swing in the inclinometer. The contrast estimates for the nonnegligible contrasts are shown in Table 15.31. From the signs on the contrast estimates we

Table 15.32   Analysis of variance for average response for the inclinometer experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
A	2	0.1288	0.0644		
Linear A	1	0.1023	0.1023	0.53	0.4813
Quadratic A	1	0.0264	0.0264	0.14	0.7183
B	2	0.2899	0.1449		
Linear B	1	0.2850	0.2850	1.47	0.2486
Quadratic B	1	0.0049	0.0049	0.03	0.8768
C	2	12.9528	6.4764		
Linear C	1	12.8863	12.8863	66.48	0.0001
Quadratic C	1	0.0665	0.0665	0.34	0.5689
D	2	24.6042	12.3021		
Linear D	1	22.9193	22.9193	118.23	0.0001
Quadratic D	1	1.6850	1.6850	8.69	0.0122
E	2	9.0561	4.5280		
Linear E	1	7.9385	7.9385	40.95	0.0001
Quadratic E	1	1.1177	1.1177	5.77	0.0334
F	2	11.5710	5.7855		
Linear F	1	11.2476	11.2476	58.02	0.0001
Quadratic F	1	0.3234	0.3234	1.67	0.2208
G	2	0.6725	0.3362	1.73	0.2179
Error	12	2.3262	0.1938		
Total	26	61.6014			

Table 15.33   Contrast estimates (average response)

Lin C	Lin D	Quad D	Lin E	Quad E	Lin F
1.692	−2.257	1.060	−1.328	0.863	1.581

see that in order to reduce the variability of the swing, factors A, C, and F should be set at their low levels, while E should be set at its high level and, from the main effect plot in Fig. 15.8(a) D should also be set at its high level.

In order to reduce the size of the swing, we need to use as response variable the average swing for each design combination (averaged over the noise combinations). These are listed in Table 15.29. The analysis of variance (shown in Table 15.32) identifies the linear trends of factors C, D, E, and F as having large effects on the swing. The contrast estimates are shown in Table 15.33. The signs of the estimates suggest that factors D and E should be set at their high levels and factors C and F at their low levels. Since this agrees with the conclusions of the analysis of variability, it is possible to reduce the size and the variability of the swing simultaneously.

Plots of the least squares estimates of the effect of the levels of factor D for both log sample variance and average response are shown in Fig. 15.8. The conclusions of the experiment are that the dimensions of the flexure and bob-weight (A, C, F) should be decreased, while the dimensions of the flange (D, E) should be increased. The experimenters comment in the article that the results match what would be expected by engineering principles. The SAS and R commands for analyzing this product array experiment are discussed in Examples 15.9.1 and 15.10.1, respectively.

(a) Log sample variance response   (b) Average response
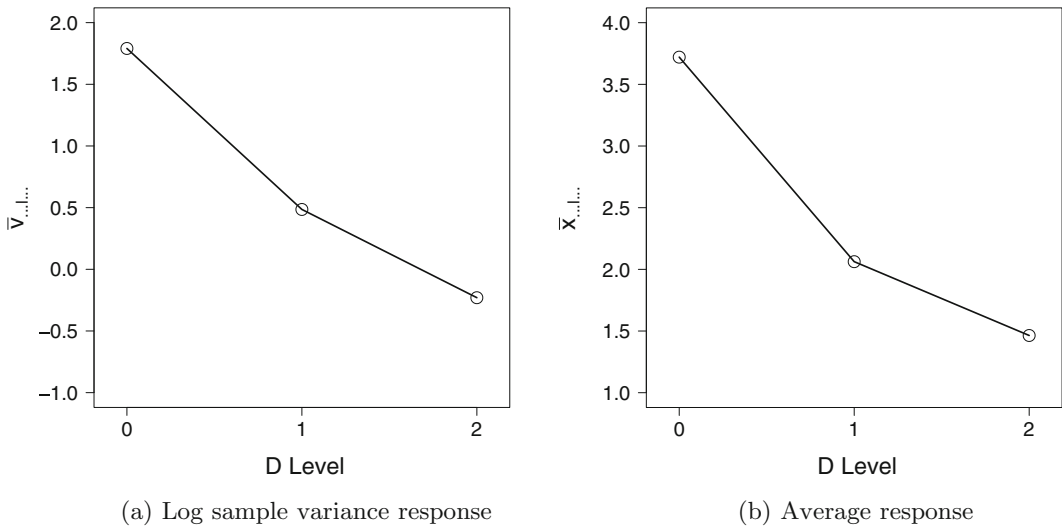
Fig. 15.8 Plots of the effect of the levels of factor D for the inclinometer experiment, where $x_{ijklmnp}$ denotes average response and $v_{ijklmnp}$ denotes the log sample variance $\ln(s^2)$ for corresponding design factor combinations

15.8   Small Screening Designs: Orthogonal Main Effect Plans

15.8.1   Saturated Designs

A design is called *saturated* if it uses only n treatment combinations, observed once each, to estimate $n - 1$ factorial contrasts plus a mean. Saturated designs were studied in Sect. 7.5, p. 219, under the heading "one observation per cell". Due to the lack of degrees of freedom for estimating σ^2, half-normal probability plots and the Voss-Wang method of simultaneous confidence intervals were used to identify contrasts with large effects. Similarly, the soup fractional factorial experiment in Sect. 15.2.3, p. 499, used a saturated design with $n = 16$ observations to measure a mean, 5 main effects and 10 two-factor interactions, using a half-normal probability plot for identifying unusually large contrasts.

In the extreme case, a saturated design may have only n observations for measuring the main effects of $p = n - 1$ factors plus a mean, in which case interactions cannot be measured separately from the main effects. If the $n - 1$ main effect contrasts are orthogonal, such designs are known as *orthogonal main-effect plans*. For example, the two designs in Table 15.21, p. 517, are saturated orthogonal main-effect plans for measuring the main effects of three factors, and the design of Table 15.22 is a saturated orthogonal main effect plan for measuring the main effects of 7 two-level factors. The design discussed in Sect. 15.3.2, p. 511, is an orthogonal main effect plan for four 3-level factors and $n = 9$ observations, and so is the design of Table 15.28, p. 523.

Orthogonal main-effect plans can be used in the early stages of experimentation with the objective of finding the factors with large main effects, and with the intention of investigating their interactions later. Implicitly, such a strategy assumes that any factors which interact will also have large main effects and so will not be screened out in the initial experiment. This assumption may not, of course, be true. For example, in the soup experiment of Sect. 15.2.3, the largest effects by far were the BE interaction and the E main effect, but the main effect of B was extremely small. So, if only the main effects had been estimated in a screening experiment, factor B would not have been selected for follow up and the large BE interaction would not have been detected. Nevertheless, it does appear that in many experiments both factor main effects do tend to appear large when the corresponding two factors

interact. So, in an initial screening experiment with very few observations, experimenters will often be content to measure main effects only.

Until now, for two-level factors, and the number of observations being a power of 2, we have constructed orthogonal main effect plans as in Sect. 15.6.1, p. 516. For $n = 2^q$, we wrote down q columns where the cth column has $n/2^{q-c}$ alternating sets of 2^{q-c} entries of -1s and $+1$s ($c = 1, \ldots, q$) as, for example, in Table 15.22, p. 518. The remaining columns were formed as products of corresponding coefficients in the first q columns taken in pairs, then in triples, and so on. We then labeled the columns in terms of main effect and interaction contrasts and determined the aliasing scheme from the relationship among the contrast coefficients. If all $n - 1$ columns are used to measure main effects, the design is a saturated orthogonal main effect plan.

If n is not a power of 2, the above strategy cannot be used for obtaining an orthogonal main effect plan. In a paper in *Biometrika* in 1946, R.L. Plackett and J.P. Burman provided a different method of constructing saturated orthogonal main effect plans with n observations for $p = n - 1$ two-level factors. This method can be used for many values of n which are multiples of 4 (including powers of 2). When n is a power of 2, these designs are likely to have similar aliasing properties to those discussed above. Otherwise, they are very different and, although all pairs of main-effect contrasts are orthogonal, we cannot write down a defining relation. Fractional factorial designs without defining relations are called *non-regular*, while designs that do have defining relations are called *regular*. In non-regular designs, main effect contrasts are correlated with interaction contrasts, but they may not be completely aliased with them.

Plackett and Burman's construction of orthogonal main effect plans uses a cyclic method. They listed the first row of each orthogonal array, called the *generator*. The entire orthogonal array is obtained from the generator by repeatedly cycling it to the right to obtain subsequent rows, and then appending a row of -1's. Such designs are known as *Plackett–Burman designs*. Some generators for cyclically generated orthogonal main-effect plans for $n = 8, 12, 16, 20$ and 24 are listed in Table 15.63, and further generators are given in Plackett and Burman's paper. Interestingly, for $n = 28$ (and some larger sizes), there is no generator that can construct an orthogonal main effect plan by cycling in this way, and Plackett and Burman used a different method of construction for these sizes.

As an example of a Plackett–Burman design, if we take the generator for $n = 12$ from Table 15.63, we have

$$1 \quad 1 \quad 1 \ -1 \quad 1 \quad 1 \ -1 \quad 1 \ -1 \ -1 \ -1,$$

and this gives us the first row of the orthogonal array. Cycling this to the right, and wrapping the end round to the beginning, we get the second row of the array as

$$-1 \quad 1 \quad 1 \quad 1 \ -1 \quad 1 \quad 1 \ -1 \quad 1 \ -1 \ -1 \,.$$

By repeatedly cycling row by row, we obtain 11 distinct rows. The 12th row of -1's is then added so that each factor appears $n/2$ times at its high level and $n/2$ times at its low level. The resulting array is shown in Table 15.34. The rows should be randomly ordered before the design is used.

We can check that all pairs of columns are orthogonal, so main effects can be estimated independently of each other. But the design is non-regular and has no defining relation. (Notice that we can also create the array in Table 15.34 by cycling to the left; rows 2–10 will just be in reverse order).

Some non-regular orthogonal main effect plans also exist for factors with more than 2 levels. For example, the orthogonal array $L_{18}(3^6 \times 6)$ of Table 15.64, at the end of the chapter, is a saturated orthogonal main effects plan for $n = 18$ observations for 6 three-level factors (whose main effects require 2 degrees of freedom each) and one six-level factor (whose main effect requires 5 degrees of

Table 15.34 A Plackett–Burman saturated orthogonal main-effect plan for 11 factors and 12 observations

1	1	1	−1	1	1	−1	1	−1	−1	−1
−1	1	1	1	−1	1	1	−1	1	−1	−1
−1	−1	1	1	1	−1	1	1	−1	1	−1
−1	−1	−1	1	1	1	−1	1	1	−1	1
1	−1	−1	−1	1	1	1	−1	1	1	−1
−1	1	−1	−1	−1	1	1	1	−1	1	1
1	−1	1	−1	−1	−1	1	1	1	−1	1
1	1	−1	1	−1	−1	−1	1	1	1	−1
−1	1	1	−1	1	−1	−1	−1	1	1	1
1	−1	1	1	−1	1	−1	−1	−1	1	1
1	1	−1	1	1	−1	1	−1	−1	−1	1
−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

freedom). Also, the orthogonal array $L_{27}(3^{13})$ shown in Table 15.65 is a saturated orthogonal main effects plan with $n = 27$ observations for 13 three-level factors.

Example 15.8.1 `Lactic acid experiment`

B. Naveena, M. Altaf, K. Bhadriah, and G. Reddy, in their 2005 paper in *Bioresource Technology* used a Plackett–Burman design with $n = 16$ observations to screen the main effects of 15 factors. They wanted to find out which factors most affect the amount of "L(+) lactic acid" produced from wheat bran using microbial metabolism in solid state fermentation. They explained that L(+) lactic acid is used widely, including in foods, anti-inflammatory drugs, and synthesis of biodegradable products. The purpose of the experiment was to take a step towards the use of renewable cheap raw material

Table 15.35 Plackett–Burman design and data (g. lactic acid per 10g. wheat bran) for the lactic acid experiment

A	B	C	D	E	F	G	H	I	J	K	L	M	N	P	y
1	1	1	1	−1	1	−1	1	1	−1	−1	1	−1	−1	−1	1.4659
−1	1	1	1	1	−1	1	−1	1	1	−1	−1	1	−1	−1	1.4284
−1	−1	1	1	1	1	−1	1	−1	1	1	−1	−1	1	−1	1.1367
−1	−1	−1	1	1	1	1	−1	1	−1	1	1	−1	−1	1	2.3041
1	−1	−1	−1	1	1	1	1	−1	1	−1	1	1	−1	−1	2.2765
−1	1	−1	−1	−1	1	1	1	1	−1	1	−1	1	1	−1	1.2114
−1	−1	1	−1	−1	−1	1	1	1	1	−1	1	−1	1	1	1.8786
1	−1	−1	1	−1	−1	−1	1	1	1	1	−1	1	−1	1	0.9304
1	1	−1	−1	1	−1	−1	−1	1	1	1	1	−1	1	−1	0.9730
−1	1	1	−1	−1	1	−1	−1	−1	1	1	1	1	−1	1	1.1137
1	−1	1	1	−1	−1	1	−1	−1	−1	1	1	1	1	−1	0.9961
−1	1	−1	1	1	−1	−1	1	−1	−1	−1	1	1	1	1	1.0268
1	−1	1	−1	1	1	−1	−1	1	−1	−1	−1	1	1	1	1.7966
1	1	−1	1	−1	1	1	−1	−1	1	−1	−1	−1	1	1	1.2928
1	1	1	−1	1	−1	1	1	−1	−1	1	−1	−1	−1	1	0.8937
−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1.0449

Sources Naveena et al., *Bioresource Technology*, 2005, published by Elsevier

Table 15.36 Contrast estimates for the lactic acid experiment

A	B	C	D	E	F	G	H
−0.065	−0.370	−0.044	−0.076	0.238	0.428	0.349	−0.016

I	J	K	L	M	N	P
0.276	0.036	−0.331	0.287	−0.026	−0.143	0.088
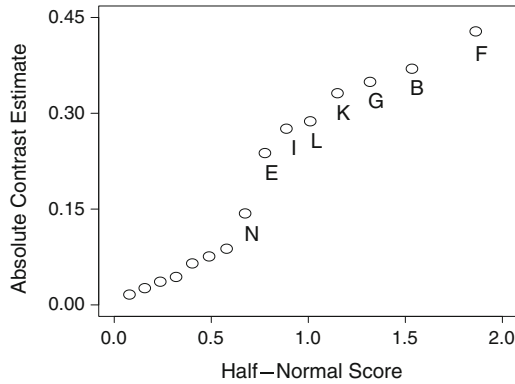
Fig. 15.9 Half-normal probability plot of normalized contrast absolute estimates for the lactic acid experiment

(here, wheat bran) for large scale production of L(+) lactic acid. The fifteen factors that were studied consisted of three physical factors (which we label A–C), a buffer (D), and eleven nutrients (E–N, P).

The Plackett–Burman design that was used can be obtained by cycling from the generator of Table 15.63. The design and the responses (grams of lactic acid per 10 grams of wheat bran) are shown in Table 15.35 before randomization. The 15 main effect contrast estimates (with divisor 1.0) and a corresponding half-normal probability plot are shown in Table 15.36 and Fig. 15.9.

We can see that with only $n = 16$ observations, it is possible to identify 7 (or possibly 8) of the 15 factors as likely to be the most influential in the lactic acid production. In order of size of contrast estimate, the 8 factors with the largest main effects are those labeled F, B, G, K, L, I, E, and N, which are all nutrients except for B which is a physical factor. These 8 factors can then be followed up in a later experiment, and their interactions examined. Since the main effect estimates for the five factors F, G, L, I, and E are all positive, the experimenters suggested that these should be examined at higher levels in the next experiment.

The estimate for the main effect of nutrient K is negative, and the experimenters commented that the wheat bran is already rich in this nutrient, so this nutrient may not need to be added in future experiments. The low level of factor N had already been set at zero, so a negative main effect estimate suggests that this nutrient, too, need not be added in future. The factor B main effect estimate is also negative and it was suggested that this be retained at its low level. Thus, main effects of only 5 factors and their interactions need to be followed up in future and this could be done in a full factorial experiment with 32 observations, or a resolution V half fraction with 16 observations. □

It is not necessary to assign all of the columns of an orthogonal main effect plan to factors. For example, the design of Table 15.35 could have been used to measure the main effects of factors A–L only and, if so, this would no longer be a saturated design—there would be 3 degrees of freedom for error. Similarly, seven of the 13 columns of the orthogonal main effect plan $L_{27}(3^{13})$ of Table 15.65 was used as a non-saturated orthogonal main effect plan for the seven 3-level design factors in the

inclinometer experiment of Sect. 15.7.1. The orthogonal array $L_{18}(3^7 \times 2)$ indicated in the right hand side of Table 15.64 is a non-saturated orthogonal main effect plan with two degrees of freedom for error.

15.8.2 Supersaturated Designs

A design is called *supersaturated* if the number of factorial effects to be estimated, plus the mean, exceeds the number of observations. In some sense, all the regular fractional factorial designs in the earlier sections are supersaturated if an insufficient number of interactions can be assumed to be negligible. For instance, in the sludge experiment of Example 15.2.1, p. 502, there were only $n = 8$ observations but, ideally, 5 main effects, 10 interactions and a mean were of interest. Similarly, in the refinery experiment of Example 15.3.1, p. 507, there were 4 main effects (requiring 2 degrees of freedom each), 6 two-factor interactions (requiring 4 degrees of freedom each) and a mean of interest – a total of 33 factorial effects, but only $n = 27$ observations could be taken. In both of these examples, and others like them, contrasts were either in the defining relation and could not be measured at all, or were measurable within a set of aliased contrasts. Aliased contrast estimators were completely correlated and non-aliased contrast estimators were independent.

However, the word *supersaturated* is not usually applied to fractions with alias schemes. Rather, the term is usually reserved for designs in which, although there are fewer observations than contrasts of interest plus the mean, *some information can be gained on all contrasts*. To achieve this, one must give up the idea of independent estimates and allow some or all contrast estimators to be correlated. In the extreme case of fewer observations than the number of factors ($n < p$), it is usually necessary to estimate main effects only and to postpone consideration of interactions among important factors to a later date.

Since the contrast estimates will be correlated, we would like the correlations to be as small as possible. The correlation between contrast estimators can be calculated using the information about the covariance of two contrast estimators from Sect. 6.7.2, and dividing by the square root of the product of their variances to obtain the formula for correlation. Following Sect. 6.7.2, p. 172, the estimators of the two contrasts $\Sigma c_i \tau_i$ and $\Sigma k_s \tau_s$, with one observation per treatment combination in the design, have correlation

$$\text{Corr}\left(\sum_{i=1}^{n} c_i Y_i \sum_{s=1}^{n} k_s Y_s \right) = \frac{\sum_{i=1}^{n} \sum_{s=1}^{n} c_i k_s}{\sqrt{\sum c_i^2 \sum k_s^2}} \ .$$

For two-level factors, each of the n contrast coefficients is -1 or $+1$, so

$$\text{Corr}\left(\sum_{i=1}^{n} c_i Y_i \sum_{s=1}^{n} k_s Y_s \right) = \frac{1}{n} \sum_{i=1}^{n} \sum_{s=1}^{n} c_i k_s \ . \tag{15.8.1}$$

In general, our recommendation is only to consider using a supersaturated design if the largest correlation between two columns is at most 1/3 and if there are likely to be very few large main effects, say at most $n/3$ (effect sparsity), (see Example 15.8.2 for problems that may be encountered in supersaturated designs when there are many large effects.)

Table 15.37 A supersaturated design with $n_{ssd} = n/2 = 6$ observations for measuring main effects of up to $n - 2 = 10$ factors, obtained from a Plackett–Burman design with $n = 12$ rows

−1	−1	−1	1	1	1	−1	1	1	−1
−1	1	−1	−1	−1	1	1	1	−1	1
1	−1	1	−1	−1	−1	1	1	1	−1
−1	1	1	−1	1	−1	−1	−1	1	1
1	−1	1	1	−1	1	−1	−1	−1	1
1	1	−1	1	1	−1	1	−1	−1	−1

One Method of Construction

There have been a few simple methods of construction proposed for supersaturated designs with $n < p$, and we describe one here that selects rows from a non-regular orthogonal main effect plan. This method was proposed by D. Lin in his article in *Technometrics*, 1993, and results in $n_{ssd} = n/2$ observations for investigating main effects of $p = n - 2$ two-level factors. The method uses an approach similar to that used in forming regular fractions from orthogonal arrays in Sect. 15.6. It starts with an orthogonal main effect plan with n rows and $n - 1$ columns containing an equal number of -1 and $+1$, with n being a multiple of 4. It then selects one column, called a *branching column*, and takes the $n_{ssd} = n/2$ rows which have $+1$ in the branching column. The $n_{ssd} = n/2$ rows which have -1 could be taken instead.

In passing, we note that, when a column of $+1$s (representing the mean) is appended to a saturated orthogonal main effect plan with n observations and $n - 1$ two-level factors, the resulting $n \times n$ array is often referred to as a *Hadamard matrix*. So, a supersaturated design constructed as above is often referred to as being constructed from a Hadamard matrix using a branching column.

All of the orthogonal main effect plans in Table 15.63 with number of rows n being a multiple of 4, but not a power of 2, can be used to construct supersaturated designs using a branching column in this way. When n is a power of 2, this technique may lead to complete aliasing between main effect contrasts. For example, in the design of the lactic acid experiment of Table 15.35, if we were select the last column (labeled P) as a branching column, and keep just those rows which have -1 in the branching column, then the main effect contrast of A would be identical to that of L, and the main effect contrast of C would be identical to that of D, and so on, giving a Resolution II design.

Complete aliasing between main effect contrasts does not occur using the cyclic generators of Table 15.63 when n is a multiple of 4 but not a power of 2. For example, suppose we generate the Plackett–Burman design in Table 15.34 using the generator for $n = 12$ from Table 15.63. If we select the last column as the branching column, and keep only the rows that have $+1$ in this column, we obtain the supersaturated design in Table 15.37 with $n_{ssd} = 6$ observations for measuring main effects of up to $n - 2 = 10$ factors. The treatment combinations in the supersaturated design can be identified from the rows of Table 15.37, with -1 representing the low level and $+1$ representing the high level of each factor. The rows should be randomly ordered before the design is used.

Using (15.8.1), we can verify that all pairs of contrast estimators have correlation 0.33 or -0.33. However, there is some hidden aliasing which is not easy to notice. If we take the contrasts for B, D and H and multiply the corresponding coefficients together, we obtain contrast C. There are several other hidden aliases of this type, too. Thus, this design should only be used in a setting where at most $n_{ssd}/3 = 2$ main effects are expected to be large. Otherwise, there will be problems with masking of effects. Example 15.8.2 shows that this design is successful in identifying two large main effects, but not three. Similar issues hold for all very small designs.

Analysis of a Supersaturated Design

In a supersaturated design, there are too many factors (and too few observations) to fit a model by least squares that contains all main effects so, typically, smaller models are investigated. But, even so, contrast estimates are not likely to be independent which means that half normal probability plots cannot be used here. Even when there are few large main effects, any analysis of supersaturated designs is tricky due to the hidden aliases. Various sophisticated methods of analysis have been researched, including certain "penalized regression" techniques which are outside the scope of this book. One simple method is to compare the fit of all possible regression models (Chap. 8) containing first one variable, then two variables, then three variables, and so on, up to $p = n - 1$ variables; a technique called *all-subsets regression*. There are still decisions to be made on how to select the best of these models, and most software packages will present options for these. Here, we will compare only the R^2 values for models (see (8.6.1) in Sect. 8.6.1), but alternative more sophisticated methods would be preferable. Research is still continuing today on the best methods of selecting influential factors (or variables) when the number of observations is so small. With so few observations, the selected model should be taken only as a possible selection of the most influential factors, and not as a model to explain the data. The potentially influential factors that are identified must be followed up in a future experiment; some of the identified factors may have only appeared to have had large effects due to the hidden aliasing.

Example 15.8.2 `Identifying influential factors`

In this example, we examine how the contrast correlations, that are an unavoidable part of a supersaturated design, might affect the identification of the factors that influence the response. Let us start with an example using the orthogonal main effects plan with $n = 12$ observations in Table 15.34 and label the columns to be the main effect contrasts of the ten factors $A, B, \ldots, J$. This same design is shown in columns 3–12 of Table 15.39, p. 537, together with two sets of responses listed as y_{BD} and y_{BDH} in the first two columns. The data y_{BD} in the first column were created by assuming that the contrast main effect values for B and D are 16 and 24, respectively, and all other main effect values are randomly selected from a $N(0, 2)$ distribution and random errors from a $N(0,1)$ distribution. Using data y_{BD}, we calculate the contrast estimates by multiplying the contrast coefficient by the corresponding response and dividing by $v/2 = 6$, we obtain the estimates in Table 15.38. Since we used an orthogonal main effects plan (Plackett–Burman design), these are independent estimates. It can be seen that factors B and D clearly have the largest main effects and these would be selected as the only two influential factors. An all subsets regression will also identify these two factors as the most influential, and the regression model is

$$\hat{y} = 34.70 + 7.62x_2 + 12.62x_4 \,,$$

where x_2 and x_4 are the coded levels ± 1 of factors B and D and, when multiplied by 2, the corresponding parameter estimates match the contrast estimates in Table 15.38. (The difference of a factor of 2 is due to the fact that the levels ± 1 are 2 apart and are treated as actual values rather than coded levels for the regression model).

Table 15.38 Main effect contrast estimates obtained from the Plackett–Burman design in Example 15.8.2

A	B	C	D	E	F	G	H	J	K
−1.84	15.24	1.37	25.24	−0.23	0.08	−1.30	−0.78	1.55	−0.05

Now, taking the column labeled K in Table 15.39 as the branching column and keeping only those rows corresponding to $+1$ in this column, we obtain the supersaturated design with $n_{ssd} = 6$ observations in Table 15.37.

An all-subsets regression would again select factors B and D. A regression model containing these two factors has R^2, so accounts for 97.7% of the variability in the data. The fitted model is

$$\hat{y} = 34.62 + 7.08x_2 + 11.97x_4 \,,$$

which is very similar to that obtained from the orthogonal main effects plan. However, with so few observations, the model should not be used for prediction; it should only be used as a guide to which factors are to be followed up in a future experiment. At the follow-up stage, a predictive model can be fitted, and the interaction between B and D can be examined. Also, notice that the number of large main effects in this example is, as recommended, not more than $n_{ssd}/3 = 2$.

If we try to use this design in a situation where there are, say, three large main effects, the contrast correlations will most likely prevent the correct factors from being selected. For example, the data set y_{BDH} in Table 15.39 was created to correspond to true values of the main effects of B, D and H equal to 16, 24, 14, respectively, and these influential factors will be selected when using the orthogonal main effects plan. But from the supersaturated design, only factor D can be detected, and an all-subsets regression will identify C as a second influential factor. This is due to the hidden aliasing in which the main effect contrast of C is defined by the main effect contrasts of B, D and H, as described above.   □

Example 15.8.2 illustrates that supersaturated designs can be successful in detecting the important factors *provided that there is only a small number* of these (say, at most $n_{ssd}/3$) compared with the number of observations, n_{ssd}. Analysis of the data in Example 15.8.2 is illustrated in Sects. 15.9.3 and 15.10.3, for the SAS and R software respectively.

15.8.3   Saturated Orthogonal Main Effect Plans Plus Interactions

Saturated orthogonal main effect plans which are equivalent to regular fractions do not give scope for estimating interactions independently of main effects. One can draw interaction plots as we did in Chaps. 6 and 7, but since the interactions are aliased with main effects, one cannot separate out the information. However, non-regular saturated orthogonal main effect plans for n not a power of 2 do, in general, offer the ability to gain some information on a few 2-factor interactions.

For example, if we take the Plackett–Burman design of Table 15.34, p. 531, and add the AB and AC interaction contrasts, we obtain the set of contrasts in Table 15.39. Notice that neither interaction column is identical to any main effect column and all correlations of interaction columns with main effect columns are 0.333 or -0.333. After adding interaction contrasts to the plan, such designs are supersaturated and so the same cautions and methods of analysis discussed in the previous section apply. Much more has been written on the topic of estimating interactions for non-regular fractions. A summary can be found in the paper of Xu et al. (2009).

An alternative possibility of measuring not only added interactions but also quadratic trends in main effects is given by the definitive screening designs outlined in the next section.

Table 15.39 Main effect, *AB*, and *AC* interaction contrasts for the saturated orthogonal main-effect plan of Table 15.34, together with the responses for Example 15.8.2

y_{BD}	y_{BDH}	A	B	C	D	E	F	G	H	I	J	K	AB	AC
28.99	35.99	1	1	1	−1	1	1	−1	1	−1	−1	−1	1	1
57.32	50.33	−1	1	1	1	−1	1	1	−1	1	−1	−1	−1	−1
39.40	46.40	−1	−1	1	1	1	−1	1	1	−1	1	−1	1	−1
40.83	47.84	−1	−1	−1	1	1	1	−1	1	1	−1	1	1	1
13.35	6.35	1	−1	−1	−1	1	1	1	−1	1	1	−1	−1	−1
28.17	35.17	−1	1	−1	−1	−1	1	1	1	−1	1	1	−1	1
13.98	20.98	1	−1	1	−1	−1	−1	1	1	1	−1	1	−1	1
54.51	61.51	1	1	−1	1	−1	−1	−1	1	1	1	−1	1	−1
32.85	25.85	−1	1	1	−1	1	−1	−1	−1	1	1	1	−1	−1
39.78	32.78	1	−1	1	1	−1	1	−1	−1	−1	1	1	−1	1
52.08	45.08	1	1	−1	1	1	−1	1	−1	−1	−1	1	1	−1
15.15	8.15	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1	1

Table 15.40 Design and data for the vaccine experiment

A_L	B_L	C_L	D_L	E_L	F_L	Response
0	1	−1	−1	−1	−1	3.6
0	−1	1	1	1	1	28.2
1	0	−1	1	1	−1	22.0
−1	0	1	−1	−1	1	4.2
−1	−1	0	1	−1	−1	2.7
1	1	0	−1	1	1	27.5
−1	1	1	0	1	−1	2.5
1	−1	−1	0	−1	1	31.9
1	−1	1	−1	0	−1	14.9
−1	1	−1	1	0	1	8.5
1	1	1	1	−1	0	30.7
−1	−1	−1	−1	1	0	2.9
0	0	0	0	0	0	14.4

Source Erler et al. (2013), *Biotechnology Letters*, © 2012 Springer Science + Business Media Dordrecht

15.8.4 Definitive Screening Designs

Definitive screening designs were introduced by Jones and Nachtsheim in the *Journal of Quality Technology* in 2011. These designs have three-level factors, and allow estimation of linear and quadratic main effect trends as well as linear×linear trends in the interactions.

Jones and Nachtsheim list definitive screening design plans for $p = 4, \ldots, 12$ factors (for example, the design for $p = 6$ is shown in Table 15.40). The rows need to be randomly ordered before the designs are used. The designs given for $p = 4$, 6, 8, and 10 allow the main effect linear trends to be estimated independently of all linear and quadratic main effect trends, and independently of the linear×linear interaction contrasts. For other cases, the linear trend contrasts are not quite orthogonal, but they can still be estimated independently of the quadratic main effects and the linear×linear interaction trend contrasts. The quadratic trends are not orthogonal to each other nor to the linear×linear trends, but they are not completely confounded (—in Table 15.40, for example, each linear×linear contrast is a function of the quadratic trends for all factors not involved in the interaction). Because of the correlations, all contrast estimates must be least squares estimates which are adjusted for other effects in the model.

Definitive screening designs are supersaturated in the sense that, for p factors, they use only $2p + 1$ observations to measure a total of $p(p + 3)/2$ trend contrasts plus a mean (e.g. if there are $p = 4$ factors, a definitive screening design uses 9 observations to measure 8 linear and quadratic main effect trend contrasts, 6 interaction linear×linear trend contrasts and a mean). Like supersaturated designs, definitive screening designs may be analyzed using all-subsets regression or a more sophisticated "penalized regression" technique. As in Chap. 7, if the selected model contains linear×linear interaction terms, the constituent linear main effect terms should be included too.

Example 15.8.3 `Vaccine experiment`

An experiment, described by Erler, de Mas, Ramsey, and Henderson in *Biotechnology Letters* (2013), was run to study the effect of $p = 6$ factors on a particular chemical reaction related to a candidate vaccine product. The experimenters were interested in the linear and quadratic main effects of the factors and the linear×linear interactions, so there were 28 contrasts of interest. They used the definitive screening design from the paper of Jones and Nachtsheim (2011), which is shown in Table 15.40, plus some extra repeats of the final treatment combination (which we are not using here).

The response is a measure of "extent of polymerization" and is shown in the last column of Table 15.40. The data were collected in a random order. An all-subsets regression run with all linear terms x_i (with values as in Table 15.40), quadratic terms x_i^2, and linear×linear terms $x_i x_j$, selects the linear trends in A, F, and D as being the most influential, with the possible addition of the linear$C$×linear$D$ trend. The regression model involving these four effects, together with the linear C main effect is

$$\hat{y} = 14.92 + 10.62x_1 + 1.16x_3 + 3.90x_4 + 5.46x_6 + 1.98x_3 x_4 \,,$$

and this accounts for 98% of the variability in the data.   □

15.9 Using SAS Software

15.9.1 Fractional Factorials

The analysis of a fractional factorial experiment by computer is identical to that of a single-replicate factorial experiment (Sect. 7.6, p. 225) except that only one effect should be entered into the model from each line of the aliasing scheme (and none from the defining relation). If two aliased effects are entered into the model, the Type I sum of squares and degrees of freedom will be zero for the second effect entered, and the Type III sum of squares and degrees of freedom will be zero for both effects.

In Table 15.41 we show a straightforward program for analyzing the sludge experiment of Example 15.2.1. The cell-means model in terms of the treatment combinations `TC` is used. (Variables `A`–`E` are created for later use.) Using `PROC GLM`, the analysis of variance is generated in the usual way by the `MODEL` statement, while the contrast estimates are obtained by the `ESTIMATE` statements, using the contrast coefficients listed in Table 15.7 (p. 503). The output is shown in Fig. 15.10. The main effects are each aliased with 2-factor (and 3-factor) interactions (see p. 503). The 2-factor interactions AC and BC are aliased with BE and AE, respectively. Since there is only one observation on each of the observed treatment combinations, the cell-means model leaves no degrees of freedom for error—this is why the p-values and values of test statistics and standard errors are either missing or meaningless. The inclusion of the `DIVISOR=4` options in the `ESTIMATE` statements ensures that all the contrasts listed in Table 15.7 will be divided by $v/2 = 4$ and give the same estimates as those in Table 15.8 (p. 503).

Table 15.41 SAS program for the sludge experiment—cell-means model

```
DATA SLUDGE;
 INPUT  A  B  C  D  E  Y;
 TC = 10000*A + 1000*B + 100*C + 10*D + E;
 LINES;
 0 0 0 1 0  195
 0 0 1 1 1  496
 0 1 0 0 1  87
 0 1 1 0 0  1371
 1 0 0 0 1  102
 1 0 1 0 0  1001
 1 1 0 1 0  354
 1 1 1 1 1  775
 ;
PROC GLM;
 CLASS TC;
 MODEL Y = TC;
 ESTIMATE 'A' TC -1 -1 -1 -1  1  1  1  1 / DIVISOR = 4;
 ESTIMATE 'B' TC -1 -1  1  1 -1 -1  1  1 / DIVISOR = 4;
 ESTIMATE 'C' TC -1  1 -1  1 -1  1 -1  1 / DIVISOR = 4;
 ESTIMATE 'D' TC  1  1 -1 -1 -1 -1  1  1 / DIVISOR = 4;
 ESTIMATE 'E' TC -1  1  1 -1  1 -1 -1  1 / DIVISOR = 4;
 ESTIMATE 'AC' TC 1 -1  1 -1 -1  1 -1  1 / DIVISOR = 4;
 ESTIMATE 'BC' TC 1 -1 -1  1  1 -1 -1  1 / DIVISOR = 4;
```

In Table 15.42, we show the SAS program for the equivalent model written in terms of main-effect and interaction parameters. The ESTIMATE statements for the main effects need no divisors, as they are automatically divided by 4 (the number of observations on each of the high and low levels). However, the ESTIMATE statements for the interaction contrasts include the option DIVISOR=2, to increase the actual divisor by a factor of 2. Without this option, the interaction estimates would be calculated with divisor 2 (the number of observations on each combination of levels of the two factors). The main-effect and interaction sums of squares are shown in Fig. 15.11. The output from the ESTIMATE statements is identical to that obtained from the cell-means model.

Again, there are no degrees of freedom for error, since a term has been included in the model from every row of the aliasing scheme. If all 2-factor interactions can be assumed to be negligible, then AC and BC would be omitted from the model, leaving 2 degrees of freedom for error.

Consider what would happen if two aliased terms were entered into the model. The defining relation for the $\frac{1}{4}$-fraction was stated in Example 15.2.1 to be $I = ABD = CDE = ABCE$. Consequently, A is aliased with BD. Adding BD into the model subsequent to A would give Type I sum of squares and degrees of freedom for BD equal to zero. This is because BD adds no more information if A is already in the model. The Type III sums of squares would be zero for both A and BD, since each would be added into the model assuming that the other is already in the model.

15.9.2 Design for the Control of Noise Variability

We now turn to the analysis of experiments involving design and noise factors, often known as Taguchi experiments. These were discussed in Sect. 15.7. There are two approaches to the analysis. The first approach involves the analysis of the mean and variance of the response observed for each design-treatment combination, calculated over the levels of the noise factors. The second approach involves

Fig. 15.10 Output from
SAS program for the
sludge
experiment—cell-means
model
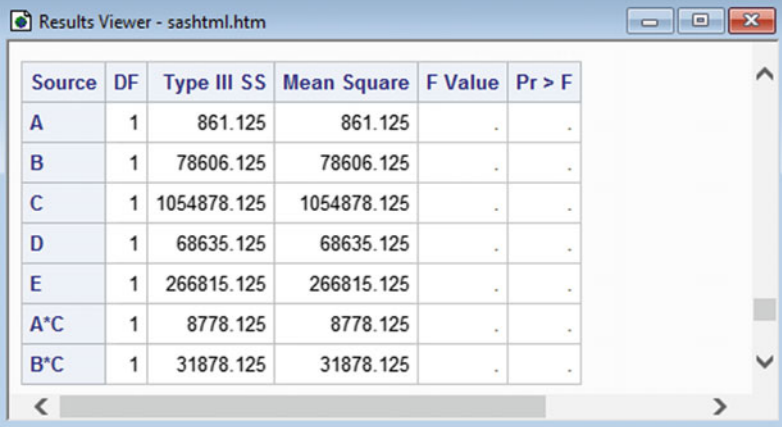
Table 15.42 SAS program for the sludge experiment—five-way model

```
PROC GLM;
 CLASSES A B C D E;
 MODEL Y= A B C D E A*C B*C;
 ESTIMATE 'A' A -1 1;
 ESTIMATE 'B' B -1 1;
 ESTIMATE 'C' C -1 1;
 ESTIMATE 'D' D -1 1;
 ESTIMATE 'E' E -1 1;
 ESTIMATE 'AC' A*C 1 -1 -1 1 / DIVISOR=2;
 ESTIMATE 'BC' B*C 1 -1 -1 1 / DIVISOR=2;
```

Fig. 15.11 Output from the SAS program for the sludge experiment—five-way model

the study of design-by-noise interactions. The first approach requires every noise combination to be observed with every design combination (that is, a product array), and the second approach requires randomization of all observed combinations of noise and design factors taken together.

Example 15.9.1 `Inclinometer experiment—product-array approach`

The inclinometer experiment was described in Sect. 15.7.1, and the data are shown in Table 15.29 (p. 526). Since this is a product array, the analysis will be done on the average and log variance of the responses over the levels of the noise factors, and these do not need to be identified in the SAS program input. Thus, the SAS program in Table 15.43 reads in the data corresponding to each combination of levels of the seven design factors (A–G) without identifying the levels of the noise factors. The average `AVY` and the log sample variance `LNVAR` of the observations for each design-treatment combination is computed and added to the data set. Since only 27 (i.e., 3^3) of the 3^7 design combinations are observed, we have a 3^{7-4} fractional factorial experiment with two possible response variables.

Two analyses are requested in Table 15.43. The first uses the response variable `LNVAR`. An analysis of variance table is requested via the `PROC GLM` statement for the model that includes main effects but no interactions. The least squares means for the levels of the design variables are requested via the `LSMEANS` statement, and these can be used to prepare plots such as those shown in Fig. 15.8 (p. 529). Linear and quadratic trends in each design factor can be tested via `ESTIMATE` or `CONTRAST` statements, only two of which are shown in the program. The final section of the program in Table 15.43 uses the response variable `AVY`. The output is similar to that shown in Tables 15.30, 15.31, 15.32 and 15.33, pp. 527–528.   □

For a mixed array, in which observed combinations of noise and design factors taken together are randomized, the analysis is done in a similar way to that of Chap. 7. In the SAS program, for the mixed-array approach, it is necessary to input the levels of the noise factors as well as those of the design factors. The model should include the main effects of the design and noise factors and at least the design-by-noise interactions. The design-by-noise interactions help in identifying those design factors whose levels give the most stable response as the noise factor levels change. Interaction plots for the significant design-by-noise interactions are made in the same way as those described in Sect. 6.8.3. If the noise factor levels are placed on the horizontal axis as in Fig. 15.7, p. 525, the levels of the design factor(s) that are most robust to the noise level fluctuations are those whose average responses result in lines closest to horizontal across the noise factor levels (cf. level 3 of factor A in Fig. 15.7(b), p. 525).

Table 15.43 SAS program for the product array analysis of the inclinometer experiment

```
DATA INCLP;
  INPUT A B C D E F G  Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8;
  AVY = (Y1 + Y2 + Y3 + Y4 + Y5 + Y6 + Y7 + Y8)/8;
  VAR = ((Y1*Y1 + Y2*Y2 + Y3*Y3 + Y4*Y4 + Y5*Y5 + Y6*Y6
                      + Y7*Y7 + Y8*Y8) - 8*AVY*AVY)/7;
LNVAR = LOG(VAR);
  LINES;
  0 0 0 0 0 0 0 0.62 3.54 3.56 0.62 3.09 0.71 0.73 3.20
  0 0 1 1 1 1 1 0.59 3.11 3.11 0.59 2.98 0.63 0.64 3.02
  : : : : : : :  :    :    :    :    :    :    :    :
  2 2 2 0 2 2 1 1.84 9.35 9.19 1.79 9.06 1.85 1.89 9.28
;
* Analysis of the log sample variance;
PROC GLM;
  CLASS A B C D E F G;
  MODEL LNVAR = A B C D E F G;
  ESTIMATE 'Lin A'  A -1  0  1;
  ESTIMATE 'Quad A' A  1 -2  1;
   :        :  :  :  :  :  :
  ESTIMATE 'Quad F' F  1 -2  1;
  CONTRAST 'Lin A'  A -1  0  1;
  CONTRAST 'Quad A' A  1 -2  1;
   :        :  :  :  :  :  :
  CONTRAST 'Quad F' F  1 -2  1;
  LSMEANS A B C D E F G;
;
* Analysis of the sample mean;
PROC GLM;
  CLASS A B C D E F G;
  MODEL AVY = A B C D E F G;
  ESTIMATE 'Lin A'  A -1  0  1;
  ESTIMATE 'Quad A' A  1 -2  1;
   :        :  :  :  :  :  :
  ESTIMATE 'Quad F' F  1 -2  1;
  CONTRAST 'Lin A'  A -1  0  1;
  CONTRAST 'Quad A' A  1 -2  1;
   :        :  :  :  :  :  :
  CONTRAST 'Quad F' F  1 -2  1;
  LSMEANS A B C D E F G;
```

15.9.3   Analysis of Small Screening Designs

For a small design with n_d observations, one method of searching for the k most influential factors, for any given k ($1 \le k \le n_d - 1$), is to select the factors that would result in a linear regression model with largest value of R^2. This is a simple method of analysis and is not guaranteed to select the best set of k factors, especially when the estimates are highly correlated. Nevertheless, we saw in Example 15.8.2 that, for simple problems, the method can work reasonably well and solutions for different values of k can be compared. In the SAS software, this method is easily achieved by adding an option to the model statement as follows:

Table 15.44 Results of the `RSQUARE` option

Number in model	R-Square	Variables in model	Number in model	R-Square	Variables in model
1	0.6597	D	1	0.7111	C
1	0.3808	E	1	0.5778	D
1	0.3452	H	1	0.2694	E
2	0.9774	B D	2	0.9693	C D
2	0.8730	E I	2	0.8670	C G
2	0.8672	C H	2	0.7970	C H
3	0.9985	B D G	3	0.9957	A C G
3	0.9871	B D I	3	0.9957	C D G
3	0.9871	B D E	3	0.9957	A C D
3	0.9871	D E I	3	0.9957	A D G

```
PROC REG;
MODEL Y = A B C D E F G H I J / SELECTION = RSQUARE BEST = 6;
```

The `SELECTION=RSQUARE` options asks SAS to check R^2 for linear models containing k variables, for all $k = 1, 2, 3, \ldots, n_d - 1$. The `BEST=6` options asks SAS to list the 6 sets of factors producing models with the 6 largest R^2.

For example, suppose we set `BEST=4`, for the supersaturated design obtained from Table 15.39 with rows corresponding to $+1$ in the branching column K, and use the data y_{BD}. Then, we would obtain the output in the left part of Table 15.44 (where we have deleted the fourth selection for the one- and two-factor models). From this table, we can see that the factors B and D, taken together, account for 97.7% of the variation in the data and little is gained from adding a third factor.

The right part of Table 15.44 shows the results of running the `RSQUARE` option using the data set y_{BDH} of Table 15.39. From this, one would most likely select factors C and D (incorrectly). Notice that the sets of $k = 3$ best factors indicate that selection of any three of A, C, D, G seems to be equivalent. This anomaly can be explained by the hidden aliasing which can be discovered by running `PROC REG` without any options, which says:

```
F =  -A + D - E
G =   A - C - D
H =  -B - C - D
I =  -B - D + E
J =  -A + B + C + D - E
```

and we see that the contrasts for A, C, D and G are linearly related. The correct selection of factors for these data is B, D and H. A linear model containing the main effects of three factors has an R^2 of only 0.971, which is still high but unlikely to be the set of three factors chosen. Thus, when there are many large effects as compared with the number of observations, their detection is difficult and, perhaps, impossible.

15.10   Using R Software

15.10.1   Fractional Factorials

The analysis of a fractional factorial experiment by computer is identical to that of a single-replicate factorial experiment (Sect. 7.7, p. 230) except that only one effect should be entered into the model from each line of the aliasing scheme (and none from the defining relation). If two aliased effects are

Table 15.45 R program and output for the sludge experiment—cell-means model

```
> sludge.data = read.table("data/sludge.txt", header = T) # Read A:E,y
> # Create variable TC within sludge.data:
> sludge.data = within(sludge.data, {TC = 10000*A+1000*B+100*C+10*D+E})
> # Create factor variables within sludge.data:
> sludge.data = within(sludge.data,
+ {fA = factor(A); fB = factor(B); fC = factor(C);
+  fD = factor(D);  fE = factor(E); fTC = factor(TC)})
> head(sludge.data, 3)

  A B C D E   y   TC  fTC fE fD fC fB fA
1 0 0 0 1 0 195   10   10  0  1  0  0  0
2 0 0 1 1 1 496  111  111  1  1  1  0  0
3 0 1 0 0 1  87 1001 1001  1  0  0  1  0

> # Analysis of variance: cell means model
> modelTC = lm(y ~ fTC, data = sludge.data)
> anova(modelTC)

Analysis of Variance Table
Response: y
          Df  Sum Sq Mean Sq F value Pr(>F)
fTC        7 1510452  215779
Residuals  0       0

> # Contrast estimates: cell means model
> lsmTC = lsmeans(modelTC, ~ fTC)
> contrast(lsmTC, list(A=c(-1,-1,-1,-1, 1, 1, 1, 1)/4,
+                      B=c(-1,-1, 1, 1,-1,-1, 1, 1)/4,
+                      C=c(-1, 1,-1, 1,-1, 1,-1, 1)/4,
+                      D=c( 1, 1,-1,-1,-1,-1, 1, 1)/4,
+                      E=c(-1, 1, 1,-1, 1,-1,-1, 1)/4,
+                      AC=c( 1,-1, 1,-1,-1, 1,-1, 1)/4,
+                      BC=c( 1,-1,-1, 1, 1,-1,-1, 1)/4))

contrast estimate  SE df t.ratio p.value
A           20.75 NaN  0     NaN     NaN
B          198.25 NaN  0     NaN     NaN
C          726.25 NaN  0     NaN     NaN
D         -185.25 NaN  0     NaN     NaN
E         -365.25 NaN  0     NaN     NaN
AC         -66.25 NaN  0     NaN     NaN
BC         126.25 NaN  0     NaN     NaN
```

entered into the model, the Type I sum of squares and degrees of freedom will be zero for the second effect entered, and the Type III sum of squares and degrees of freedom will be zero for both effects.

In Table 15.45 we show a straightforward program and output for analyzing the sludge experiment of Example 15.2.1. The cell-means model in terms of the factor variable fTC for the treatment combinations TC is used. (Factor variables fA–fE are created for later use.) The linear models function lm fits the model, and the anova function generates the analysis of variance table. Then the contrast estimates of interest are generated by a single contrast statement, using least squares means generated

Table 15.46   R program (continued) for the sludge experiment—five-way model

```
> # Analysis of variance: factorial effects model
> modelFE = lm(y ~ fA + fB + fC + fD + fE + fA:fC + fB:fC, data = sludge.data)
> anova(modelFE)

Analysis of Variance Table

Response: y
          Df   Sum Sq Mean Sq F value Pr(>F)
fA         1      861     861
fB         1    78606   78606
fC         1  1054878 1054878
fD         1    68635   68635
fE         1   266815  266815
fA:fC      1     8778    8778
fB:fC      1    31878   31878
Residuals  0        0

> # Contrast estimates: factorial effects model
> lsmA = lsmeans(modelFE, ~ fA); contrast(lsmA, list(A = c(-1, 1)))
> lsmB = lsmeans(modelFE, ~ fB); contrast(lsmB, list(B = c(-1, 1)))
> lsmC = lsmeans(modelFE, ~ fC); contrast(lsmC, list(C = c(-1, 1)))
> lsmD = lsmeans(modelFE, ~ fD); contrast(lsmD, list(D = c(-1, 1)))
> lsmE = lsmeans(modelFE, ~ fE); contrast(lsmE, list(E = c(-1, 1)))
> lsmAC = lsmeans(modelFE, ~ fA:fC); contrast(lsmAC, list(AC = c(1,-1,-1,1)/2))
> lsmBC = lsmeans(modelFE, ~ fB:fC); contrast(lsmBC, list(BC = c(1,-1,-1,1)/2))
```

by the lsmeans function and saved as lsmTC, and using the contrast coefficients listed in Table 15.7 (p. 503). Nicely formatted output is obtained by providing the contrasts as a list, including a name (i.e. A, B, etc.) for each.

The main effects are each aliased with 2-factor (and 3-factor) interactions (see p. 503). The 2-factor interactions *AC* and *BC* are aliased with *BE* and *AE*, respectively. Since there is only one observation on each of the observed treatment combinations, the cell-means model leaves no degrees of freedom for error—this is why the standard errors, test statistics and *p*-values generated by contrast are "not a number" (NaN). The inclusion of the divisor 4 each contrast ensures that all the contrasts listed in Table 15.7 will be divided by $v/2 = 4$ and give the same estimates as those in Table 15.8 (p. 503).

In Table 15.46, we show a continuation of the R program of Table 15.45, illustrating the analysis using a factorial effects model, providing selected output. Main effect and interaction contrast estimates are computed using the contrast statement of the least squares means function lsmeans. All of the contrasts as specified use coefficients $c_{ijk} = \pm 1/4$. For the main effects, the specified coefficients ± 1 are automatically divided by 4, averaging over the four combinations of the other two factors. For the two-factor interaction contrasts, the coefficients $\pm 1/2$ are specified and are automatically divided by 2, averaging over the two observations on each combination of these two factors. The main-effect and interaction sums of squares are shown in Table 15.46. The contrast estimates (not shown) are identical to those obtained from the cell-means model.

Again, there are no degrees of freedom for error, since a term has been included in the model from every row of the aliasing scheme. If all 2-factor interactions can be assumed to be negligible, then *AC* and *BC* would be omitted from the model, leaving 2 degrees of freedom for error.

Consider what would happen if two aliased terms were entered into the model. The defining relation for the $\frac{1}{4}$-fraction was stated in Example 15.2.1 to be $I = ABD = CDE = ABCE$. Consequently, A is aliased with BD. Adding BD into the model subsequent to A would give Type I sum of squares and degrees of freedom for BD equal to zero. This is because BD adds no more information if A is already in the model. The Type III sums of squares would be zero for both A and BD, since each would be added into the model assuming that the other is already in the model.

15.10.2   Design for the Control of Noise Variability

We now turn to the analysis of experiments involving design and noise factors, often known as Taguchi experiments. These were discussed in Sect. 15.7. There are two approaches to the analysis. The first approach involves the analysis of the mean and variance of the response observed for each design-treatment combination, calculated over the levels of the noise factors. The second approach involves the study of design-by-noise interactions. The first approach requires every noise combination to be observed with every design combination (that is, a product array), and the second approach requires randomization of all observed combinations of noise and design factors taken together.

Example 15.10.1 `Inclinometer experiment—product-array approach`

The inclinometer experiment was described in Sect. 15.7.1, and the data are shown in Table 15.29 (p. 526). Since this is a product array, the analysis will be done on the average and log variance of the responses over the levels of the noise factors, and these do not need to be identified in the R program input. Thus, the R program in Table 15.47 reads in the data corresponding to each combination of levels of the seven design factors (A–G) without identifying the levels of the noise factors. The average `Avy` and the log sample variance `LnVar` of the observations for each design-treatment combination is computed and added to the data set. Since only 27 (i.e., 3^3) of the 3^7 design combinations are observed, we have a 3^{7-4} fractional factorial experiment with two possible response variables.

Two analyses are requested in Table 15.47. The first uses the response variable `LnVar`. After fitting the linear model that includes main effects but no interactions via the `lm` function, an analysis of variance table is requested via the `anova` statement. The least squares means for the levels of each design factor are requested by the `lsmeans` function, and these can be used to prepare plots such as those shown in Fig. 15.8 (p. 529). Also, the `contrast` function uses the saved least squares means to estimate and test the linear and quadratic trends for each design factor. This use of `lsmeans` is only shown in the program for design factor A but can be used to obtain least squares means for the remaining design factors and the trend contrast estimates for B–F, (since they each have equally-spaced levels). A regression model is also fit, including linear and quadratic terms for each design factor, to generate the sum of squares associated with the linear and quadratic contrasts. This approach also provides the correct linear and quadratic trend contrast sums of squares for G, even though the levels of G are not equally spaced. The final section of the program in Table 15.47 uses the response variable `Avy`. The output is similar to that shown in Tables 15.30, 15.31, 15.32 and 15.33, pp. 527–528.   □

For a mixed array, in which observed combinations of noise and design factors taken together are randomized, the analysis is done in a similar way to that of Chap. 7. The model should include the main effects of the design and noise factors and at least the design-by-noise interactions. The design-by-noise interactions help in identifying those design factors whose levels give the most stable response as the noise factor levels change. Interaction plots for the significant design-by-noise interactions are made in the same way as those described in Sect. 6.9.3. If the noise factor levels are placed on the horizontal axis as in Fig. 15.7, the levels of the design factor(s) that are most robust to the noise level fluctuations are those whose average responses result in lines closest to horizontal across the noise factor levels (cf. level 3 of factor A in Fig. 15.7(b), p. 525).

Table 15.47 R program for the product array analysis of the inclinometer experiment

```
# Read data from file. Header: A B C D E F G y1 y2 y3 y4 y5 y6 y7 y8
ipa.data = read.table("data/inclinometer.product.txt", header = T)
# Create factor variables
ipa.data = within(ipa.data,
{fA = factor(A); fB = factor(B); fC = factor(C); fD = factor(D);
fE = factor(E); fF = factor(F); fG = factor(G)})
# Compute Avy and LnVar of data at each design treatment combo
ipa.data = within(ipa.data, {Avy = (y1 + y2 + y3 + y4 + y5 + y6 + y7 + y8)/8
    LnVar = log(((y1*y1 + y2*y2 + y3*y3 + y4*y4 + y5*y5 + y6*y6
                  + y7*y7 + y8*y8) - 8*Avy^2)/7)})
# Remove variables y1:y8
ipa.data = within(ipa.data, {remove(y1, y2, y3, y4, y5, y6, y7, y8)})
head(ipa.data,3)

# Analysis of log sample variance
model1 = lm(LnVar ~ fA + fB + fC + fD + fE + fF + fG, data = ipa.data)
anova(model1)
# Least square means and contrast estimates for A (similarly for B--F)
library(lsmeans)
lsmA1 = lsmeans(model1, ~ fA)
lsmA1
contrast(lsmA1, list(Alin = c(-1, 0, 1), Aquad = c(1, -2, 1)))
# Regression for log sample variance (to get contrast SS's)
model2 = lm(LnVar ~ A + I(A^2) + B + I(B^2) + C + I(C^2) + D + I(D^2)
            + E + I(E^2) + F + I(F^2) + G + I(G^2), data = ipa.data)
anova(model2)

# Analysis of sample mean
model3 = lm(Avy ~ fA + fB + fC + fD + fE + fF + fG, data = ipa.data)
anova(model3)
# Least square means and contrast estimates for A (similarly for B--F)
# library(lsmeans)
lsmA3 = lsmeans(model3, ~ fA)
lsmA3
contrast(lsmA3, list(Alin = c(-1, 0, 1), Aquad = c(1, -2, 1)))
# Regression for sample mean (to get contrast SS's, including for G)
model4 = lm(Avy ~ A + I(A^2) + B + I(B^2) + C + I(C^2) + D + I(D^2)
            + E + I(E^2) + F + I(F^2) + G + I(G^2), data = ipa.data)
anova(model4)
```

15.10.3   Analysis of Small Screening Designs

For a small design with n_d observations, one method of searching for the k most influential factors, for any given k ($1 \leq k \leq n_d - 1$), is to select the factors that would result in a linear regression model with largest value of R^2. This is a simple method of analysis and is not guaranteed to select the best set of k factors, especially when the estimates are highly correlated. Nevertheless, we saw in Example 15.8.2 that, for simple problems, the method can work reasonably well and solutions for different values of k can be compared. In the R software, the function regsubsets from the leaps package fits regression models containing all subsets of p of the factors up to nvmax and prints out the nbest models found for each value of p. Table 15.48 shows the output obtained from the R

Table 15.48 Results of the all subsets variable selection via `regsubsets`

```
> ssd <- data.frame(A = c(-1, -1,  1, -1,  1,  1),
+                   B = c(-1,  1, -1,  1, -1,  1),
+                   C = c(-1, -1,  1,  1,  1, -1),
+                   D = c( 1, -1, -1, -1,  1,  1),
+                   E = c( 1, -1, -1,  1, -1,  1),
+                   F = c( 1,  1, -1, -1,  1, -1),
+                   G = c(-1,  1,  1, -1, -1,  1),
+                   H = c( 1,  1,  1, -1, -1, -1),
+                   I = c( 1, -1,  1,  1, -1, -1),
+                   J = c(-1,  1, -1,  1,  1, -1),
+ y = c(40.83, 28.17, 13.98, 32.85, 39.78, 52.08))

> library(leaps)
> regsubsets.outssd <- regsubsets(y ~ A + B + C + D + E +
+   F + G + H + I + J, data = ssd, nbest = 3,
+   nvmax = 3, method = "exhaustive")
Warning message:
In leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
force.in = force.in,  :
  5  linear dependencies found

> summary.outssd <- summary(regsubsets.outssd)
> as.data.frame(summary.outssd$outmat)
          A B C D E F G H I J
1  ( 1 )          *
1  ( 2 )            *
1  ( 3 )                  *
2  ( 1 )    *    *
2  ( 2 )          *          *
2  ( 3 )      *              *
3  ( 1 )    *    *      *
3  ( 2 )    *      *          *
3  ( 3 )          * *          *

> round(summary.outssd$rsq, 4)
[1] 0.6597 0.3808 0.3452 0.9774 0.8730 0.8672 0.9985
[8] 0.9871 0.9871

> alias(aov(y ~ A + B + C + D + E + F + G + H + I + J, ssd))
Model :
y ~ A + B + C + D + E + F + G + H + I + J
Complete :
  (Intercept)  A  B  C  D  E
F  0          -1  0  0  1 -1
G  0           1  0 -1 -1  0
H  0           0 -1 -1 -1  0
I  0           0 -1  0 -1  1
J  0          -1  1  1  1 -1
```

software `regsubsets` command, for the supersaturated design obtained from Table 15.39 with rows corresponding to $+1$ in the branching column K, and the data y_{BD}.

The `regsubsets` call fits all possible models containing up to `nvmax = 3` factors and displays the `nbest = 3` of each size. The statement

   `as.data.frame(summary.outssd$outmat)`

produces a list of these best models with stars indicating which factors have been fitted in the model. The R^2 values for the listed models are obtained from the `round(summary.outssd$rsq, 4)` command, where the number 4 specifies the number of decimal places. The fourth entry in the list of R^2 values corresponds to the model containing B and D; we see that the $R^2 = 0.9774$, and little is gained by including other factors.

Notice the warning message in the output that says there were 5 linear dependencies found. These are the hidden aliases that can be identified by running

   `alias(aov(y   A + B + C + D + E + F + G + H + I + J, ssd))`

and observing that

```
F =   -A + D - E
G =    A - C - D
H =   -B - C - D
I =   -B - D + E
J =   -A + B + C + D - E
```

If we try to use this same design for the data set y_{BDH} from Table 15.39, we obtain

```
          A B C D E F G H I J
1  ( 1 )       *
1  ( 2 )         *
1  ( 3 )           *
2  ( 1 )       * *
2  ( 2 )       *       *
2  ( 3 )       *         *
3  ( 1 )       * *     *
3  ( 2 ) *         *     *
3  ( 3 ) *     *       *
```

with associated R^2 values of `0.7111 0.5778 0.2694 0.9693 0.8670 0.7970 0.9957 0.9957 0.9957`. From this, one would most likely select factors C and D (incorrectly). Notice that the sets of $k = 4$ best factors indicate that selection of any three of A, C, D, G seems to be equivalent. This anomaly can be explained by the hidden aliasing which can be discovered with the `alias` command as shown above and we see that the contrasts for A, C, D and G are linearly related. The correct selection of factors for these data is B, D and H. A linear model containing the main effects of three factors has an R^2 of only 0.971, which is still high but unlikely to be the set of three factors chosen. Thus, when there are many large effects as compared with the number of observations, their detection is difficult and, perhaps, impossible.

Exercises

1. **Decontamination experiment—beta particles, continued**

   Suppose that only the first block of the data (beta particles) had been obtained in the decontamination experiment described in Exercise 8 of Chap. 13 (p. 466). The design would then have been a $\frac{1}{2}$-fraction of a 2^4 experiment with defining relation $I = ABCD$. The half fraction is shown in Table 15.49. Analyze the data and compare your conclusions with those of the full experiment in

Table 15.49 Block I of the decontamination experiment

Treatment combinations (Response)							
1010	1111	0110	0000	1100	0101	0011	1001
(716)	(686)	(498)	(1437)	(527)	(579)	(1433)	(906)

Sources Barnett and Mead (1956). Copyright © 1956 Blackwell Publishers. Reprinted with permission

Table 15.50 Yields (in pounds) of mangold roots for Block III of the mangold experiment

Treatment combinations (Yield)							
00101	11001	01011	01110	10010	11100	00000	10111
(896)	(1284)	(996)	(860)	(1184)	(984)	(740)	(1468)

Source Design and Analysis of Experiments, by O. Kempthorne 1976, Reprinted by permission of Krieger Publishing Company Inc.

Exercise 8 of Chap. 13. Explain the circumstances under which a half fraction would be preferred to a single-replicate factorial experiment.

2. **Mangold experiment, continued**

The mangold experiment in Sect. 13.5, p. 447, was a single replicate confounded design for a 2^5 experiment in $b = 4$ blocks of size 8. The five factors were Sulphate of Ammonia (factor *A* at levels 0 or 0.6 cwt per acre), Superphosphate (factor *B* at levels 0 or 0.5 cwt per acre), Muriate of Potash (factor *C* at levels 0 or 1.0 cwt per acre), Agricultural Salt (factor *D* at levels 0 or 5 cwt per acre), and Dung (factor *E* at levels 0 or 10 tons per acre). All of the 3-, 4-, and 5-factor interactions were expected to be negligible. The two three-factor interactions *ABD*, *BCE* and their product *ACDE* were selected for confounding.

Suppose that the data from only the third block had been available, so that we have a $\frac{1}{4}$-fraction. The data are reproduced in Table 15.50.

(a) Write down the aliasing scheme for this fractional factorial experiment.
(b) Analyze the data. What conclusions can you draw?
(c) Comparing your conclusions with those of Sect. 13.5, what extra information do you gain by running the single-replicate design instead of the fraction?
(d) When would you recommend that an experimenter consider using a fractional factorial design rather than a single-replicate design?

3. **Dye experiment, continued**

The dye experiment was discussed in Sect. 14.2.4 (p. 478). There were three factors: the concentration of inorganic material *M* in the free water in the reaction mixture (factor *A* at three equally spaced levels), the volume of free water in the reaction mixture (factor *B* at three equally spaced levels), and the concentration of inorganic material *N* in the free water in the reaction mixture (factor *C* at three equally spaced levels). The data for the first replicate of the original experiment were given in Table 14.6 (p. 479) and the first block is reproduced in Table 15.51. The design for

Table 15.51 Volume of dyestuff for Block I of the dye experiment

Treatment combinations (Yield)								
000	021	012	110	101	122	220	211	202
(74)	(130)	(56)	(110)	(166)	(227)	(195)	(146)	(90)

Source Data adapted from *The Design and Analysis of Industrial Experiments*, Second edition, 1979, Ed: O.L. Davies. Published by Longman Group Limited

Table 15.52 Yields of sugar beet for Block III of the sugar-beet experiment

Treatment combinations (Yield)								
202	020	210	111	001	122	221	012	100
(2198)	(2093)	(2354)	(2268)	(1926)	(2152)	(2349)	(2025)	(2106)

Source Yates (1935). Copyright © 1935 Blackwell Publishers. Reprinted with permission

the first replicate was a single-replicate design that confounded (AB^2C^2; A^2BC). Analyze the data of Block I as though it had come from a $\frac{1}{3}$-fraction. State your conclusions.

4. **Sugar beet experiment, continued**

The sugar beet experiment described in Exercise 6 of Chap. 14 concerned the effects of three standard fertilizers, nitrogen, phosphate, and potassium (factors N, P, and K), each at three equally spaced levels, on sugar beet yield. The experiment was run as a single-replicate confounding the contrasts (NP^2K; N^2PK^2). Suppose the only data available were those of Block III, reproduced in Table 15.52.

(a) If the only data available were those from Block III, write out the aliasing scheme for the design.
(b) Analyze the data from Block III as though they came from a $\frac{1}{3}$-fraction. State your conclusions.

5. **Flour experiment, continued**

Suppose that the data from Block II of the 4×2^4 experiment in Table 15.16 (p. 514) had been lost, so that only Block I remained. This would then constitute a $\frac{1}{4}$-fraction.

(a) Write out the aliasing scheme for the design. What is the resolution number. Is this a good design?
(b) Bearing in mind the purpose of the experiment, can you find a better $\frac{1}{4}$-fraction? If so, write out the design and its aliasing scheme.
(c) Analyze the data from Block I of Table 15.16. What can you conclude?

6. **Handwheel experiment**

E.N. Corlett and G. Gregory describe an experiment in the 1960 issue of *Applied Statistics* that was concerned with finding the design of a machine tool handwheel that would maximize the accuracy on the part of the operator in the setting of the machine tool handwheel. The apparatus consisted of an optical dividing head with a dial mounted onto a mandrel to which was connected the handwheel spindle. The spindle was provided with an adjustable friction brake. The operator first offset the dial by 15° and then moved the handwheel so that a line on the dial was brought "into coincidence with a fixed line on the dividing head, making the final adjustment by means of a series of taps by hand on the handwheel rim."

Seven factors, each at two levels (coded 0 and 1) were investigated as follows.

A: Handwheel diameter (5.5 in., 10 in.)
B: Dial diameter (4 in., 8 in.)
C: Thickness of the dial line (0.008 in., 0.064 in.)
D: Friction of the spindle (7.5 lb.-in., 45 lb.-in.)

Table 15.53 Log variance of observations for the handwheel experiment

Block I		Block II	
ABCDEFG	$\ln(s^2)$	*ABCDEFG*	$\ln(s^2)$
0 0 0 0 0 0 0	0.7044	1 1 0 0 0 0 0	0.0561
1 0 1 0 0 0 0	0.5907	0 1 1 0 0 0 0	0.3615
0 0 1 0 0 0 1	−0.0297	1 1 1 0 0 0 1	−0.1158
1 0 0 0 0 0 1	0.3914	0 1 0 0 0 0 1	−0.1952
0 1 0 1 0 0 0	0.0792	1 0 0 1 0 0 0	0.4585
1 1 1 1 0 0 0	0.3228	0 0 1 1 0 0 0	0.2531
0 1 1 1 0 0 1	−0.1599	1 0 1 1 0 0 1	0.2727
1 1 0 1 0 0 1	−0.0996	0 0 0 1 0 0 1	0.6861
0 0 0 1 1 0 0	0.5878	1 1 0 1 1 0 0	−0.0074
1 0 1 1 1 0 0	0.3577	0 1 1 1 1 0 0	−0.2328
0 0 1 1 1 0 1	0.1847	1 1 1 1 1 0 1	−0.1046
1 0 0 1 1 0 1	0.5706	0 1 0 1 1 0 1	−0.2069
0 1 0 0 1 0 0	−0.1805	1 0 0 0 1 0 0	0.8051
1 1 1 0 1 0 0	−0.3224	0 0 1 0 1 0 0	0.4634
0 1 1 0 1 0 1	−0.1433	1 0 1 0 1 0 1	0.2904
1 1 0 0 1 0 1	0.1354	0 0 0 0 1 0 1	0.4692
Block III		Block IV	
ABCDEFG	$\ln(s^2)$	*ABCDEFG*	$\ln(s^2)$
0 1 0 0 0 1 0	−0.6760	1 0 0 0 0 1 0	0.5457
1 1 1 0 0 1 0	−0.3824	0 0 1 0 0 1 0	0.0846
0 1 1 0 0 1 1	−0.2996	1 0 1 0 0 1 1	0.4453
1 1 0 0 0 1 1	−0.4539	0 0 0 0 0 1 1	0.2361
0 0 0 1 0 1 0	0.2970	1 1 0 1 0 1 0	−0.5069
1 0 1 1 0 1 0	0.1646	0 1 1 1 0 1 0	−0.3299
0 0 1 1 0 1 1	0.3878	1 1 1 1 0 1 1	−0.3245
1 0 0 1 0 1 1	0.2168	0 1 0 1 0 1 1	−0.3233
0 1 0 1 1 1 0	0.0148	1 0 0 1 1 1 0	0.4199
1 1 1 1 1 1 0	−0.4898	0 0 1 1 1 1 0	0.2957
0 1 1 1 1 1 1	0.1308	1 0 1 1 1 1 1	0.2278
1 1 0 1 1 1 1	−0.1829	0 0 0 1 1 1 1	0.4269
0 0 0 0 1 1 0	0.0182	1 1 0 0 1 1 0	−0.4798
1 0 1 0 1 1 0	0.2070	0 1 1 0 1 1 0	−0.0669
0 0 1 0 1 1 1	0.1101	1 1 1 0 1 1 1	−0.0584
1 0 0 0 1 1 1	0.2642	0 1 0 0 1 1 1	−0.6856

Source Corlett and Gregory (1960). Copyright © 1960 Blackwell Publishers. Reprinted with permission

E: Level of operator's elbow relative to height of handwheel
(Level with center of spindle, 6 in. above spindle center)
F: Previous experience of operator (Practiced, Nonpracticed)
G: Knowledge of accuracy of previous setting (Feedback, No feedback)

The response variable was $\ln(s^2)$, where s^2 was the sample variance of 25 repeated observations for a particular treatment combination. It was estimated that each set of 25 repeated observations would take about 15 min to complete, including setup time. In a morning or afternoon session of four hours, therefore, sixteen observations could be taken. The experiment was to last over two

days, which meant that a 2^{7-1} fractional factorial experiment was required, divided into 4 blocks of 16.

The highest-order interaction was selected for the defining relation of the fraction, that is, $I = ABCDEFG$. Only two operators were used for the experiment, one for each level of practice. The difference between these operators was not of interest, only the interaction of the level of practice with the other factors. Rather unusually, then, the main effect of F was selected as one of the contrasts for confounding. All the 2-factor interactions and most of the 3-factor interactions were thought to be of interest. Unlikely 3-factor interactions included ACG and BDE, which were also chosen for confounding with blocks. The complete set of confounded contrasts was $F, ACG, ACFG$ together with its set of aliases $ABCDEG, BDEF, BDE$. All other main-effect, 2-factor, and 3-factor interaction contrasts could be estimated.

The data obtained from the experiment are shown in Table 15.53.

(a) Write out the aliasing scheme for the design.

(b) Using a computer package, estimate the (estimable) main-effect and interaction contrasts.

(c) Prepare a half-normal probability plot of the contrast estimates and identify the most important main effects and interactions.

(d) The authors of the article point out that if the responses are normally distributed and n is large (where n is the number of repeated observations, 25 in this experiment), then the response variable $\ln(s^2)$ has approximately constant variance equal to $2/(n-1)$. Calculate the standard error for each of the contrasts estimated in part (c). Using Bonferroni's method with an individual significance level of 0.001 for each test (giving an overall level of at most 0.06), which main effects and interactions are significantly different from zero? Do these results agree with the results from part (c)? Discuss why or why not.

(e) Draw interaction plots of the important interactions and discuss recommended settings for the six factors A, B, C, D, E, and G for the practiced and nonpracticed operators individually.

(f) Would you recommend further experimentation? If so, which factors and which settings would you recommend? Can you suggest a suitable design?

7. **Paint experiment**

(a) Suppose that you need to design an experiment involving 6 factors (A, B, C, D, E, F) at 2 levels each (64 treatment combinations) and that only 8 observations can be taken. You decide to sacrifice information on the $ABF, ACDF$, and $ABCE$ contrasts. Write out the defining relation and the two rows of the aliasing scheme showing the aliasing of A and the aliasing of AC.

(b) Explain what aliasing means.

(c) An experiment was run in Germany by S. Eibl, U. Kess, and F. Pukelsheim (*Journal of Quality Technology*, 1992) on the thickness of a paint coating. Prior to the experiment, the thickness achieved was around 2 mm, much higher than the target 0.8 mm. They selected the following six factors, each at two levels:

A: belt speed $\quad$ B: tube width $\quad$ C: pump pressure
D: paint viscosity $\quad$ E: tube height $\quad$ F: heating temperature

They used the $\frac{1}{8}$-fraction with the aliasing scheme in part (a), and they decided to ignore all interactions for this first experiment. Since they wanted to monitor the variation of the thickness, they took four observations on each of the 8 treatment combinations in the fraction.

Table 15.54 2^{10-6} resolution III fraction and pore diameter (nm) for the anatase experiment of Exercise 8

A	B	C	D	E	F	G	H	I	J	Pore diameter
0	0	0	0	1	0	0	1	0	0	6.4
1	0	0	0	0	0	1	1	1	1	5.7
0	1	0	0	0	1	0	1	1	1	6.1
1	1	0	0	1	1	1	1	0	0	7.6
0	0	1	0	0	1	1	0	0	1	3.5
1	0	1	0	1	1	0	0	1	0	3.5
0	1	1	0	1	0	1	0	1	0	3.7
1	1	1	0	0	0	0	0	0	1	6.5
0	0	0	1	0	1	1	0	1	0	10.1
1	0	0	1	1	1	0	0	0	1	3.6
0	1	0	1	1	0	1	0	0	1	15.6
1	1	0	1	0	0	0	0	1	0	12.3
0	0	1	1	1	0	0	1	1	1	12.1
1	0	1	1	0	0	1	1	0	0	17.3
0	1	1	1	0	1	0	1	0	0	14.9
1	1	1	1	1	1	1	1	1	1	15.3

Source Olsen et al. (2014), *Journal of Porous Materials*, 21, © 2014, Springer Science + Business Media New York. With permission of Springer

The data are shown in Table 16.1, p. 570, where the two levels of each factor are coded as -1 and 1 and shown for factors A–F in the columns labeled z_A–z_F.

Calculate the analysis of variance table and contrast estimates using response variable LNVAR (the log variance). What do you conclude?

(d) Assuming that the order of observations was completely randomized, calculate the analysis of variance table and also contrast estimates of interest, using the 32 observations separately (without combining them into an average). Remembering that the goal is to reduce the thickness, what conclusions would you draw from this particular experiment?

(e) The experimenters decided to run a followup experiment with at most 16 observations. You can use any of the original 6 factors and you can change the levels from their original settings. The ultimate goal is to achieve a coating of 0.8 mm. Suggest a followup experiment.

8. **Anatase experiment**

R.E. Olsen and coauthors described several experiments in the *Journal of Porous Materials* (2014). These concerned the study of anatase (a form of titanium dioxide) as a catalyst support. Specific catalyst support properties are required, such as certain surface area, pore volume, and pore diameter. Samples were prepared by mixing various chemicals with water in specified orders and speeds to produce a "slurry". For half the observations, the slurry was (i) dried, (ii) rinsed with distilled water and (iii) calcinated; this was called the DRC procedure. For the remaining observations, (ii) and (iii) were interchanged to give the DCR procedure.

In the first of their experiments, all factors had two levels, coded here as 0 or 1. These factors were A Mixing order, B speed of water addition (slow, fast), C amount of water (7 ml, 25 ml), D rinsing order (DRC, DCR), E drying time (3 h, 24 h), F drying temperature (25 °C, 100 °C), G calcination ramp rate (2 °C/min, 20 °C/min), H calcination temperature (400 °C, 700 °C), I calcination time (2 h, 20 h), J amount of an aluminium compound added (5, 22). A 2^{10-6} resolution III fraction was

run. Several responses were measured, including the pore diameters which are listed together with the fraction in Table 15.54. The observations were taken in a random order (not shown here).

(a) Part of the aliasing scheme for the design is:

$$E = ACD, \ F = BCD, \ G = ACD, \ H = CD, \ I = ABCD, \ J = ABC.$$

Write out the defining relation for the fraction (a total of 63 contrasts):

(b) Fifteen factorial effects can be included in the model. This can include the 10 main effects. Suggest five interaction contrasts to include (these cannot be aliased with main effects nor with each other.)

(c) Estimate the 15 contrasts in your model and plot a half normal probability plot.

(d) From part (c), which factors seem to affect the pore diameter the most. Suppose you were going to follow up these factors and their interactions in more detail in a later experiment. How would you design such an experiment if you could take 16 observations?

9. **Flour early experiment**

The flour experiment was introduced in Example 15.5.1, p. 513. In Table 15.55, we show part of the design for an early experiment (the first in a series of four experiments). Six ingredients, A, B, C, D, E, F, added to the flour were to be investigated in the experiment. In addition, there were three noise factors: Factor P (which was a combination of factors N and S in Example 15.5.1) had two levels ("high yeast with long proof time" or "low yeast with short proof time"), Factor Q, (as in Example 15.5.1, two levels "undermixing, little water, heavy pressure" or "overmixing, much water, little pressure"), and Factor R (two levels, underbake or overbake).

A crossed array was selected. The noise array was a $\frac{1}{2}$-fraction with defining relation $I = PQR$. Each of the four noise combinations was run on a single day, so that the experiment ran over four days. The design array was a $\frac{1}{4}$-fraction with defining relation $I = ABCD = BCEF = ADEF$, and this was run on each day. Thus the noise contrasts are confounded with days and cannot be analyzed. However, the object of the experiment was to examine the average yield (specific volume, ml/100 g) and the variance of the yield for the design factors across the noise factors.

(a) Calculate the average yield and the log variance of the yield for each design-treatment combination.

(b) Analyze the two sets of data separately. What recommendations would you make if the objective is to reduce the variability and increase the specific volume?

10. **Injection molding experiment**

S.R. Schmidt and R.G. Launsby in their book *Understanding Industrial Designed Experiments* describe an experiment on the effect of six factors on the shrinkage of a part produced by injection molding. The six factors were injection velocity (factor A), cooling time (factor B), barrel zone temperature (factor C), mold temperature (factor D), hold pressure (factor E), and back pressure (factor F). Each factor had two levels coded 0 and 1.

There were two responses of interest, the length and width of the part after shrinkage. The purpose of the experiment was to find settings of the six variables that would enable the parts to be "on target," that is, a post-shrinkage length of 14.5 units and width of 9.35 units.

Table 15.55 Specific volume for part of experiment 1 of the flour early experiment

Design combinations	Noise combinations			
	Day 1	Day 2	Day 3	Day 4
	(111)	(101)	(000)	(011)
000000	519	446	337	415
000011	503	468	343	418
001101	567	471	355	424
001110	552	489	361	425
010101	534	466	356	431
010110	549	461	354	427
011000	560	480	345	437
011011	535	477	363	418
100100	558	483	376	418
100111	551	472	349	426
101001	576	487	358	434
101010	569	494	357	444
110001	562	474	358	404
110010	569	494	348	400
111100	568	478	367	463
111111	551	500	373	462

Source Tuck et al. (1993). Copyright © 1993 Blackwell Publishers. Reprinted with permission

The orthogonal array in Table 15.22, p. 518, was selected with columns columns 1–6 labeled A, B, D, C, E, F, and columns 5 and 6 multiplied by -1. One degree of freedom (corresponding to column 7) is available to measure σ^2 or one of the two-factor interactions. Five parts were measured at each treatment combination, and the lengths and widths are recorded in Table 15.56.

(a) Write down the defining relation for the $\frac{1}{8}$-fraction and the aliasing scheme. The investigators assumed that all the interactions were negligible. If they had not done so, which interactions could have been measured?

(b) For the length data, calculate the average response and the standard deviation of the response for each treatment combination.

(c) Can you recommend which factors should be investigated more thoroughly in order to find a setting that would give the required length and also factors that could be set to reduce the variability?

(d) Repeat parts (a) and (b) for the width data.

(e) Can you make any overall recommendation?

(f) Write down the assumptions on the model that would need to be true in order to interpret the analysis of variance. Are these assumptions likely to be valid for this experiment?

11. **Spectrometer experiment, continued**

Read the details of the spectrometer experiment in Exercise 10 of Chap. 7. You will need to have access to your solutions to that exercise to answer this question.

Suppose that you are consultant for a different company and that they wish to run a similar experiment, with the same five factors, but with a total of 64 observations. To keep things simple, you might recommend that factors A and C be examined at 2 levels each rather than 3 levels in your first experiment (even though you may suspect that some of the factors have quadratic trends).

Table 15.56 Lengths and widths of parts after shrinkage in the injection molding experiment

| Treatment combinations | | | | | | Length (Deviation from 14.5) $\times 10^4$ | | | | | Width (Deviation from 9.35) $\times 10^4$ | | | | |
A	B	C	D	E	F										
0	0	0	0	0	0	0	5	0	0	5	75	60	70	85	90
0	0	0	1	1	1	75	90	70	65	65	50	40	40	40	45
0	1	1	0	0	1	45	50	45	45	45	45	45	45	50	40
0	1	1	1	1	0	100	105	105	110	105	130	130	125	135	135
1	0	1	0	1	0	105	110	105	120	100	55	60	60	55	60
1	0	1	1	0	1	45	55	65	50	50	80	65	50	40	45
1	1	0	0	1	1	150	140	155	50	145	100	80	85	90	85
1	1	0	1	0	0	55	65	55	55	60	65	60	65	65	60

Source Schmidt and Launsby (1992). Copyright © 1992 Air Academy Press. Reprinted with permission

Table 15.57 Analysis of variance for the industrial experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
A	1	262.205	262.205	54.57	0.0857
B	1	11.045	11.045	2.30	0.3712
C	1	981.245	981.245	204.21	0.0445
D	1	5.120	5.120	1.07	0.4899
E	1	1568.000	1568.000	326.33	0.0352
F	1	8.820	8.820	1.84	0.4048
Error	1	4.805	4.805		
Total	7	2841.240			

Thus, you have a 2^5 experiment. List 5 interactions that you are particularly interested in studying. You should use information from your answer to (a) and (b) of Exercise 10 of Chap. 7 in choosing the interactions. Design a factorial experiment in 4 blocks of size 8. State exactly how you chose your design. Write out at least three of the treatment combinations in two of the blocks and explain how you obtained them.

12. **Design of industrial experiment**

Suppose that you are asked to design an experiment for 6 treatment factors each having two levels. Only 64 observations can be taken in total, and these should be divided into 8 blocks of size 8. Suppose that you decide to confound the interaction contrasts *ABD*, *DEF*, and *ACDF*.

(a) Can all the other interaction contrasts be estimated?
(b) What does the statement "*ABD* is confounded" mean?
(c) How would you obtain the 8 blocks? Write out two blocks as an example.
(d) Suppose that the budget is cut before the experiment can take place, and only 8 observations can be taken in total. How would you decide which 8 observations to take? What can be estimated?
(e) Suppose that you were fairly sure that all interactions involving 4 factors or more were negligible and that neither *D* nor *F* interacts with any of the other factors. Suppose that the analysis of variance table obtained from the results of the experiment is as in Table 15.57. What would you investigate in a followup experiment? Give your reasons.

Table 15.58 Plackett–Burman design and data for Exercise 14

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	y
−1	−1	−1	−1	1	−1	1	−1	1	1	1	1	−1	−1	1	−1	−1	1	1	134.2
1	−1	−1	−1	−1	1	−1	1	−1	1	1	1	1	−1	−1	1	−1	−1	1	250.0
1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	1	1	−1	−1	1	−1	−1	59.3
−1	1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	1	1	−1	−1	1	−1	155.0
−1	−1	1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	1	1	−1	−1	1	293.4
1	−1	−1	1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	1	1	−1	−1	174.4
−1	1	−1	−1	1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	1	1	−1	182.3
−1	−1	1	−1	−1	1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	1	1	211.9
1	−1	−1	1	−1	−1	1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	1	13.6
1	1	−1	−1	1	−1	−1	1	1	−1	−1	−1	−1	1	−1	1	−1	1	1	95.5
1	1	1	−1	−1	1	−1	−1	1	1	−1	−1	−1	−1	1	−1	1	−1	1	158.0
1	1	1	1	−1	−1	1	−1	−1	1	1	−1	−1	−1	−1	1	−1	1	−1	366.4
−1	1	1	1	1	−1	−1	1	−1	−1	1	1	−1	−1	−1	−1	1	−1	1	164.8
1	−1	1	1	1	1	−1	−1	1	−1	−1	1	1	−1	−1	−1	−1	1	−1	98.9
−1	1	−1	1	1	1	1	−1	−1	1	−1	−1	1	1	−1	−1	−1	−1	1	45.8
1	−1	1	−1	1	1	1	1	−1	−1	1	−1	−1	1	1	−1	−1	−1	−1	170.1
−1	1	−1	1	−1	1	1	1	1	−1	−1	1	−1	−1	1	1	−1	−1	−1	107.3
−1	−1	1	−1	1	−1	1	1	1	1	−1	−1	1	−1	−1	1	1	−1	−1	283.6
−1	−1	−1	1	−1	1	−1	1	1	1	1	−1	−1	1	−1	−1	1	1	−1	130.2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	373.5

13. Suppose that you wish to run an experiment with four treatment factors (A, B, C, D) each having three levels. The only likely interactions are AB, AC, and ABC. The experiment needs to be run in blocks of size at most $k = 9$.

   (a) Design a 3^4 experiment in $b = 3^2$ blocks of size $k = 3^2$ confounding ABD and AB^2CD. What else is confounded? Are you happy with this design? Why or why not?
   (b) Show how you would obtain the nine blocks and show one of the blocks as an illustration.
   (c) Write out the degrees of freedom column for the analysis of variance table. (Read the question information again before you do this.)
   (d) Suppose that the blocks are randomly ordered. After the first block is run, the budget for the experiment is cut, so only nine observations are available. The design is now a 3^{4-2} fractional factorial design. Write down the defining relation for the design. Is this design going to be useful in examining the main effects and the interactions of interest? Why or why not?

14. **Plackett–Burman and supersaturated design**

   Table 15.58 shows a hypothetical set of data from a Plackett–Burman design with 20 observations and 19 factors. The data were simulated from a model in which factors C, J, K and P all have large main effects, all other main effects are drawn from a $N(0, 4.5)$ distribution and the random errors have a $N(0, 3)$ distribution. There are no interactions.

   (a) Using the data from Table 15.58, either fit a full main effects model or run an all-subsets regression to verify that the four important factors can be detected.
   (b) What are the estimates of the main effects of the four important factors?

(c) Using the *first* column of the design in Table 15.58 as a branching column, create a supersaturated design corresponding to +1 in the branching column.

(d) The supersaturated design in part (c) has 10 observations which may not be quite sufficient to be able to detect four large main effects. Run an all-subsets regression using the 10 observations. How many of the four important factors can be detected and what are their main effect estimates?

15. **Anatase experiment**

The anatase experiment was described in Exercise 8. After the initial experiment, several further experiments were run. Table 15.59 shows one of the definitive screening designs (where factor *D* of Exercise 8 has been set at level DCR, and factor *A* held constant). The response shown is pore diameter (nm), The randomized order of the observations is shown in the original paper. The design allows linear and quadratic trends in the main effects of some factors and some of the linear×linear interactions to be measured.

(a) Use an all subsets regression to find a model that explains much of the variability in the data (R^2 at least 0.90). Remember to include the main effects of any factors that are involved in large linear×linear interactions.

(b) If you were to design a follow-up experiment, which factors (and interactions) would you examine? Suggest a suitable design if you could take 16 more observations.

Table 15.59 Definitive screening design and pore diameter (nm) for the anatase experiment of Exercise 15

B	C	E	F	G	H	I	J	Pore diameter
0	−1	1	1	−1	1	1	1	9.0
0	1	−1	−1	1	−1	−1	−1	10.3
−1	0	−1	1	1	1	1	−1	14.6
1	0	1	−1	−1	−1	−1	1	16.9
−1	−1	0	1	1	−1	−1	0	6.6
1	1	0	−1	−1	1	1	−1	13.3
1	−1	1	0	1	1	−1	−1	15.3
−1	1	−1	0	−1	−1	1	1	7.8
−1	−1	1	−1	0	−1	1	−1	11.0
1	1	−1	1	0	1	−1	1	14.0
1	−1	−1	−1	1	0	1	1	11.2
−1	1	1	1	−1	0	−1	−1	9.3
−1	1	1	−1	1	1	0	1	8.5
1	−1	−1	1	−1	−1	0	−1	10.0
1	1	1	1	1	−1	1	0	18.6
−1	−1	−1	−1	−1	1	−1	0	11.8
0	0	0	0	0	0	0	0	12.5

Sources Olsen et al. (2014), *Journal of Porous Materials*, 21, © 2014, Springer Science + Business Media New York. With permission of Springer

Table 15.60 2^{p-s} fractions of 2^p experiments. For each defining relation, s independent generators are underlined, and s corresponding equations are given. To obtain the $v = 2^{p-s}$ treatment combinations in the fraction, list all v combinations of levels a_i of the $p - s$ factors not determined by the equations, then use the equations modulo 2 to complete each treatment combination. For two blocks, confound the effect in parentheses and its aliases

2^{p-s}	v	Defining relation	Equations
2^{5-2}_{III}	8	$I = \underline{ABCE} = \underline{ABD} = CDE$	$a_4 = a_1 + a_2$
		(AC)	$a_5 = a_1 + a_2 + a_3$
2^{6-2}_{IV}	16	$I = \underline{ABCD} = \underline{CDEF} = ABEF$	$a_4 = a_1 + a_2 + a_3$
		(ACE)	$a_6 = a_3 + a_4 + a_5$
2^{7-2}_{IV}	32	$I = \underline{ABCDE} = \underline{ABFG} = CDEFG$	$a_5 = a_1 + a_2 + a_3 + a_4$
		(AEF)	$a_7 = a_1 + a_2 + a_6$
2^{8-2}_{V}	64	$I = \underline{ABCDE} = \underline{DEFGH} = ABCFGH$	$a_5 = a_1 + a_2 + a_3 + a_4$
		(CEF)	$a_8 = a_4 + a_5 + a_6 + a_7$
2^{6-3}_{III}	8	$I = \underline{BCD} = \underline{ABE} = ACDE$	$a_4 = a_2 + a_3$
		$\quad = \underline{ABCF} = ADF = CEF$	$a_5 = a_1 + a_2$
		$\quad = BDEF \quad (AC)$	$a_6 = a_1 + a_2 + a_3$
2^{7-3}_{IV}	16	$I = \underline{ABCD} = \underline{CDEF} = ABEF$	$a_4 = a_1 + a_2 + a_3$
		$\quad = \underline{ACEG} = BDEG = ADFG$	$a_6 = a_3 + a_4 + a_5$
		$\quad = BCFG \quad (ACF)$	$a_7 = a_1 + a_3 + a_5$
2^{8-3}_{IV}	32	$I = \underline{ABCD} = \underline{CDEF} = ABEF$	$a_4 = a_1 + a_2 + a_3$
		$\quad = \underline{ACEGH} = BDEGH = ADFGH$	$a_6 = a_3 + a_4 + a_5$
		$\quad = BCFGH \quad (ABG)$	$a_8 = a_1 + a_3 + a_5 + a_7$
2^{9-3}_{IV}	64	$I = \underline{CDEF} = \underline{ACEGH} = ADFGH$	$a_4 = a_1 + a_2 + a_3$
		$\quad = \underline{ABCDJ} = ABEFJ = BDEGHJ$	$a_8 = a_1 + a_3 + a_5 + a_7$
		$\quad = BCFGHJ \quad (ACF)$	$a_9 = a_1 + a_2 + a_3 + a_4$
2^{7-4}_{III}	8	$I = \underline{ABCD} = \underline{BCE} = ADE$	$a_4 = a_1 + a_2 + a_3$
		$\quad = \underline{ACF} = BDF = ABEF$	$a_5 = a_2 + a_3$
		$\quad = CDEF = \underline{ABG} = CDG$	$a_6 = a_1 + a_3$
		$\quad = ACEG = BDEG = BCFG$	$a_7 = a_1 + a_2$
		$\quad = ADFG = EFG = ABCDEFG$	
2^{8-4}_{IV}	16	$I = \underline{ABCD} = \underline{CDEF} = ABEF$	$a_4 = a_1 + a_2 + a_3$
		$\quad = \underline{ADFG} = BCFG = ACEG$	$a_6 = a_3 + a_4 + a_5$
		$\quad = BDEG = \underline{ABGH} = CDGH$	$a_7 = a_1 + a_4 + a_6$
		$\quad = ABCDEFGH = EFGH = BDFH$	$a_8 = a_1 + a_2 + a_7$
		$\quad = ACFH = BCEH = ADEH$	
		(ADH)	
2^{9-4}_{IV}	32	$I = \underline{ABCD} = \underline{CDEF} = ABEF$	$a_4 = a_1 + a_2 + a_3$
		$\quad = \underline{ADFGH} = BCFGH = ACEGH$	$a_6 = a_3 + a_4 + a_5$
		$\quad = BDEGH = \underline{ABGJ} = CDGJ$	$a_7 = a_1 + a_4 + a_5 + a_6$
		$\quad = ABCDEFGH = EFGJ = BDFHJ$	$a_8 = a_1 + a_2 + a_7$
		$\quad = ACFHJ = BCEHJ = ADEHJ$	
		(ABH)	

Table 15.61 3^{p-s} fractions of 3^p experiments. For each defining relation, s independent generators are underlined, and s corresponding equations are given. To obtain the $v = 3^{p-s}$ treatment combinations in the fraction, list all v combinations of levels a_i of the $p - s$ factors not determined by the equations, then use the equations modulo 3 to complete each treatment combination

3^{p-s}	v	Defining relation	Equations
3^{4-2}_{III}	9	$I = \underline{AB^2C} = A^2BC^2$	$a_3 = 2a_1 + a_2$
		$\quad = \underline{ABD} = A^2CD = B^2C^2D$	$a_4 = 2a_2 + 2a_3$
		$\quad = A^2B^2D^2 = BCD^2 = AC^2D^2$	
3^{5-2}_{III}	27	$I = \underline{ABC^2D^2} = A^2B^2CD$	$a_4 = a_1 + a_2 + 2a_3$
		$\quad = \underline{ADE^2} = A^2BC^2E^2 = B^2CD^2E^2$	$a_5 = a_1 + a_4$
		$\quad = A^2D^2E = BC^2DE = AB^2CE$	
3^{6-2}_{IV}	81	$I = \underline{ABC^2D^2} = A^2B^2CD$	$a_4 = a_1 + a_2 + 2a_3$
		$\quad = \underline{ACEF} = A^2BD^2EF = B^2C^2DEF$	$a_6 = 2a_1 + 2a_3 + 2a_5$
		$\quad = A^2C^2E^2F^2 = BCD^2E^2F^2 = AB^2DE^2F^2$	
3^{6-3}_{III}	27	$I = \underline{ABC^2D^2} = A^2B^2CD$	$a_4 = a_1 + a_2 + 2a_3$
		$\quad = \underline{BDE^2} = AB^2C^2E^2 = A^2CD^2E^2$	
		$\quad = B^2D^2E = AC^2DE = A^2BCE$	$a_5 = a_2 + a_4$
		$\quad = \underline{CDF^2} = ABF^2 = A^2B^2C^2D^2F^2$	$a_6 = a_3 + a_4$
		$\quad = BCD^2E^2F^2 = AB^2DE^2F^2 = A^2C^2E^2F^2$	
		$\quad = B^2CEF^2 = AD^2EF^2 = A^2BC^2DEF^2$	
		$\quad = C^2D^2F = ABCDF = A^2B^2F$	
		$\quad = BC^2E^2F = AB^2CD^2E^2F = A^2DE^2F$	
		$\quad = B^2C^2DEF = ACEF = A^2BD^2EF$	
3^{7-3}_{IV}	81	$I = \underline{ABC^2D^2} = A^2B^2CD$	$a_4 = a_1 + a_2 + 2a_3$
		$\quad = \underline{ACEF} = A^2BD^2EF = B^2C^2DEF$	$a_6 = 2a_1 + 2a_3 + 2a_5$
		$\quad = A^2C^2E^2F^2 = BCD^2E^2F^2 = AB^2DE^2F^2$	
		$\quad = \underline{BC^2E^2FG} = AB^2CD^2E^2FG = A^2DE^2FG$	$a_7 = 2a_2 + a_3$
		$\quad = ABF^2G = A^2B^2C^2D^2F^2G = CDF^2G$	$\quad + a_5 + 2a_6$
		$\quad = A^2BCEG = B^2D^2EG = AC^2DEG$	
		$\quad = B^2CEF^2G^2 = AD^2EF^2G^2 = A^2BC^2DEF^2G^2$	
		$\quad = AB^2C^2E^2G^2 = A^2CD^2E^2G^2 = BDE^2G^2$	
		$\quad = AB^2FG^2 = C^2D^2FG^2 = ABCDFG^2$	

Table 15.62 Orthogonal arrays with 2^p observations and useful column labelings. Assign factors in alphabetical order

No. of factors   8 observations—Design of Table 15.22

Columns

No. of factors	1	2	3	12	13	23	123
3–6	A	B	C	E	F	G	D

No. of factors   16 observations—Design of Table 15.24

Columns

No. of factors	1	2	12	3	13	23	123	4	14	24	124	34	134	234	1234
4–5	A	B		C				D							E
6–15	A	B	K	C	L	M	D	E	N	P	F	Q	G	H	J

No. of factors   32 observations

Columns

No. of factors	1	2	12	3	13	23	123	4	14	24	124	34	134	234	1234
4–22	A	B	C			M	D				N		P	Q	G

Columns

No. of factors	5	15	25	125	35	135	235	1235	45	145	245	1245	345	1345	2345	12345
4–22	E			R		S	T	H		U	V	J	W	K	L	F

Table 15.63 Generators for cyclically generated orthogonal main-effect plans. These are saturated designs for factors each at two levels, for n observations with n divisible by 4 but not a power of 2. To generate a design, systematically cycle the generator to the right to obtain $n - 1$ rows; then include a final row of -1's

n	Generator
8	1 1 1 −1 1 −1 −1
12	1 1 1 −1 1 1 −1 1 −1 −1 −1
16	1 1 1 1 −1 1 −1 1 1 −1 −1 1 −1 −1 −1
20	1 1 1 1 −1 1 −1 1 −1 −1 −1 −1 1 1 −1 1 1 −1 −1
24	1 1 1 1 1 −1 1 −1 1 1 −1 −1 1 1 −1 −1 1 −1 1 −1 −1 −1 −1
28	Does not exist
32	−1 −1 −1 −1 −1 1 −1 −1 −1 1 1 1 −1 1 −1 1 −1 −1 1 −1 1 1 1 1 −1 −1 1 1 −1 1 1
36	−1 −1 −1 −1 −1 1 −1 −1 −1 1 1 −1 1 1 1 1 −1 1 −1 1 −1 −1 1 1 −1 1 1 −1 1 −1 −1 −1 1 1 1

Obtained via a computer program described by Dean and Draper (1999). Further generators are given in Plackett and Burman (1946)

Table 15.64 Orthogonal arrays for 18 observations: $L_{18}(3^6 \times 6)$ and $L_{18}(3^7 \times 2)$

							For $L_{18}(3^7 \times 2)$ replace 6-level column by the two columns below	
0	0	0	0	0	0	0	0	0
0	1	2	2	0	1	1	1	0
0	2	1	2	1	0	2	2	0
0	1	1	0	2	2	3	0	1
0	2	0	1	2	1	4	1	1
0	0	2	1	1	2	5	2	1
1	1	1	1	1	1	0	0	0
1	2	0	0	1	2	1	1	0
1	0	2	0	2	1	2	2	0
1	2	2	1	0	0	3	0	1
1	0	1	2	0	2	4	1	1
1	1	0	2	2	0	5	2	1
2	2	2	2	2	2	0	0	0
2	0	1	1	2	0	1	1	0
2	1	0	1	0	2	2	2	0
2	0	0	2	1	1	3	0	1
2	1	2	0	1	0	4	1	1
2	2	1	0	0	1	5	2	1

Sources $L_{18}(3^6 \times 6)$ from http://neilsloane.com/oadir/

Table 15.65 A 3^p orthogonal array for 27 observations: $L_{27}(3^{13})$

						Columns						
1	2	3	4	5	6	7	8	9	10	11	12	13
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	2	2	2	2	2	2	2	2	2
0	1	1	1	0	0	1	1	1	2	2	0	2
0	1	1	1	1	1	2	2	2	0	0	1	0
0	1	1	1	2	2	0	0	0	1	1	2	1
0	2	2	2	0	0	2	2	2	1	1	0	1
0	2	2	2	1	1	0	0	0	2	2	1	2
0	2	2	2	2	2	1	1	1	0	0	2	0
1	0	1	2	0	1	0	1	2	1	2	2	0
1	0	1	2	1	2	1	2	0	2	0	0	1
1	0	1	2	2	0	2	0	1	0	1	1	2
1	1	2	0	0	1	1	2	0	0	1	2	2
1	1	2	0	1	2	2	0	1	1	2	0	0
1	1	2	0	2	0	0	1	2	2	0	1	1
1	2	0	1	0	1	2	0	1	2	0	2	1
1	2	0	1	1	2	0	1	2	0	1	0	2
1	2	0	1	2	0	1	2	0	1	2	1	0
2	0	2	1	0	2	0	2	1	2	1	1	0
2	0	2	1	1	0	1	0	2	0	2	2	1
2	0	2	1	2	1	2	1	0	1	0	0	2
2	1	0	2	0	2	1	0	2	1	0	1	2
2	1	0	2	1	0	2	1	0	2	1	2	0
2	1	0	2	2	1	0	2	1	0	2	0	1
2	2	1	0	0	2	2	1	0	0	2	1	1
2	2	1	0	1	0	0	2	1	1	0	2	2
2	2	1	0	2	1	1	0	2	2	1	0	0
A	*B*			*C*			*D*	*E*	*F*	*G*		

Response Surface Methodology

16

16.1 Introduction

Response surface methodology was developed by Box and Wilson in 1951 to aid the improvement of manufacturing processes in the chemical industry. The purpose was to optimize chemical reactions to obtain, for example, high yield and purity at low cost. This was accomplished through the use of sequential experimentation involving factors such as temperature, pressure, duration of reaction, and proportion of reactants. The same methodology can be used to model or optimize any response that is affected by the levels of one or more quantitative factors. The models are generalizations of the polynomial regression models studied in Chap. 8.

The general scenario is as follows. The response is a quantitative continuous variable (e.g., yield, purity, cost), and the mean response is a smooth but unknown function of the levels of p factors (e.g., temperature, pressure), and the levels are real-valued and accurately controllable. The mean response, when plotted as a function of the treatment combinations, is a surface in $p + 1$ dimensions, called the *response surface*. For example, Fig. 16.1 shows a response surface for two factors A and B.

We will denote the levels of A by values of x_1 or x_A and the levels of B by values of x_2 or x_B. We will denote a treatment combination by $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ or by $\mathbf{x} = (x_A, x_B, \ldots, x_P)$ and the mean response at $\mathbf{x}$ by $\eta_{\mathbf{x}} = E[Y_{\mathbf{x}}]$. The general response surface model is of the form

$$Y_{\mathbf{x}} = \eta_{\mathbf{x}} + \epsilon_{\mathbf{x}} \,,$$

where $\epsilon_{\mathbf{x}}$ is a random error variable.

The objective of obtaining a response surface is twofold:

(i) to locate a feasible treatment combination $\mathbf{x}$ for which the mean response is maximized (or minimized, or equal to a specific target value); and

(ii) to estimate the response surface in the vicinity of this good location or region, in order to better understand the "local" effects of the factors on the mean response.

In general, throughout the chapter we will think about maximizing the response, but we show via an example that exactly the same techniques can be used for minimizing a response. The techniques can easily be adapted when the goal is to have the response close to a target value.

Fig. 16.1 Hypothetical response surface for two factors
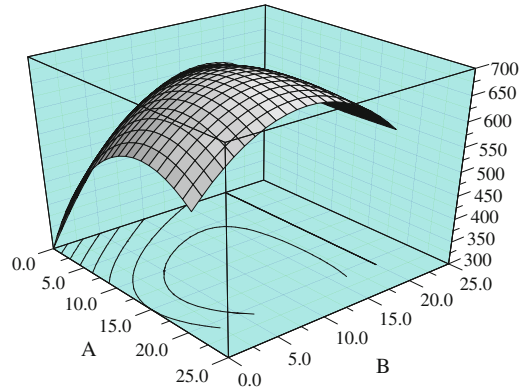
One possible approach to achieving the objective involves collecting observations at each location on a grid of treatment combinations spanning the entire experimental region of interest (as suggested by Fig. 16.1). However, the number of observations required by such a comprehensive approach can be very large, and it grows very quickly as the number of factors under study increases. Also, somewhat sophisticated modeling techniques would generally be needed to obtain an adequate fit of a model over the entire region. Instead, it is generally more efficient to conduct a sequence of small "local" experiments with which to search out the location of the peak mean response and then to study its vicinity.

Seeking out the peak is analogous to climbing an unfamiliar mountain under conditions of limited visibility—the mountain is the response surface, and your location on the mountainside is a treatment combination, say $\mathbf{x}_a$. Standing at position $\mathbf{x}_a$, you look around and can see enough to determine in which direction to go to continue a steep ascent. Then you climb in the determined direction as long as it continues to take you up, not looking about lest you lose footing. Then you stop and look around again to determine whether you are at the top of the mountain or in which direction you need to continue your ascent. Of course, when you reach a peak, due to the limited visibility, you may not be sure that you have actually reached the highest peak.

How does one do this experimentally? Looking around with limited visibility is equivalent to analyzing the data of a *local experiment*, consisting of observations on treatment combinations $\mathbf{x}$ close to your current position, $\mathbf{x}_a$. The local terrain is assessed by fitting a local model. Collecting observations in sufficiently close proximity to one another generally allows the local response surface to be well approximated by a rather simple polynomial regression model. When still far from the peak, a first-order model is often adequate. The fitted first-order model is a plane, from which the direction or *path of steepest ascent* is easily determined. Then observations are collected along this path as long as the response continues to increase. When the response stops increasing, another local experiment can be conducted to determine a new path of steepest ascent. This process can be iterated until the first-order model no longer adequately describes the local true surface. For example, close to the peak, the true surface generally exhibits greater curvature, and a first-order regression model becomes inadequate, exhibiting lack of fit. A larger number of observations is needed to fit a higher-order model with which to locate and study the vicinity of the peak. Typically, a second-order model is suitable.

A flow chart describing the steps in this process is shown in Fig. 16.2. While a surface is difficult to envisage in more than three dimensions, the process can work well for any number of factors. How well it works depends on several decisions requiring judgment on the part of the experimenter. The first part of this chapter (Sect. 16.2) looks at the left-hand portion of the flow chart and investigates first-order designs and first-order models, including lack of fit and the path of steepest ascent. Section 16.3

Fig. 16.2 Flow chart for response surface methods
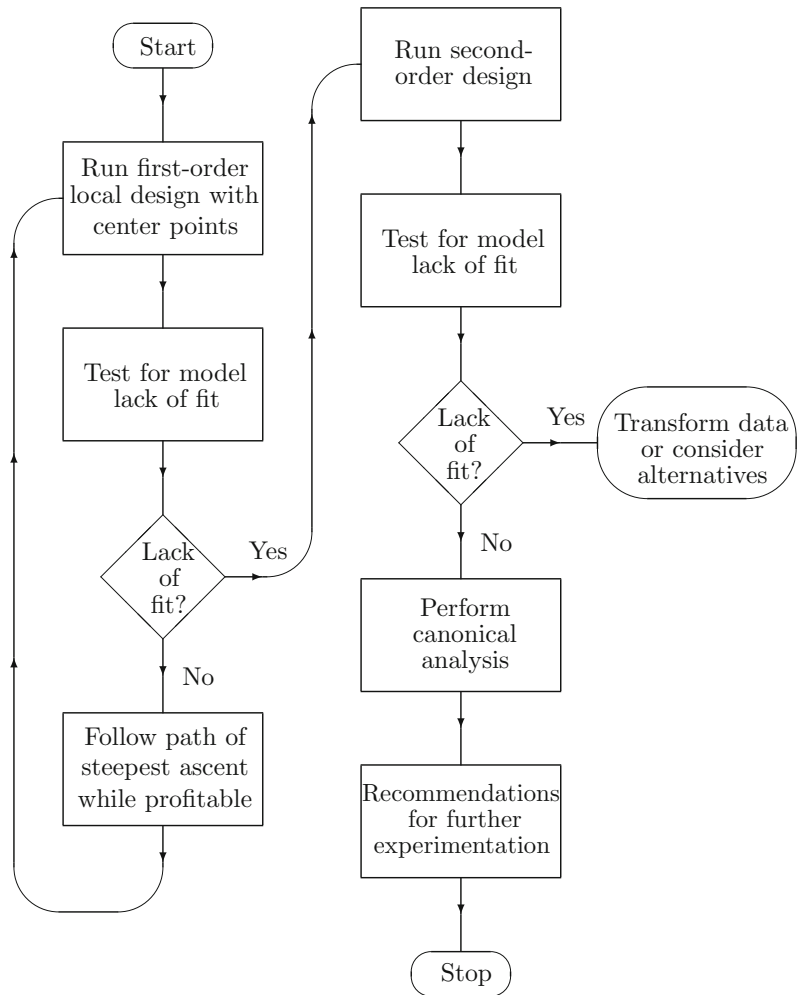
addresses the right-hand portion of the flow chart, which becomes relevant when the vicinity of the peak is reached. Second-order designs and models are described. More details about second-order designs are given in Sect. 16.4, and an experiment conducted in the flour milling industry is described in Sect. 16.5. The collection of observations as a block design is discussed in Sect. 16.6. Sections 16.7 and 16.8 describe the use of the SAS and R software, respectively.

16.2   First-Order Designs and Analysis

16.2.1   Models

Before the peak of the response surface is reached, a small local experiment is conducted to assess the local terrain. If the local experiment is not in the vicinity of the peak, then a first-order regression model often provides an adequate approximation to the local response surface. For p factors, the standard *first-order model* is a first-order polynomial regression model:

$$Y_{\mathbf{x},t} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon_{\mathbf{x},t} \,, \tag{16.2.1}$$

where $Y_{\mathbf{x},t}$ denotes the tth observation at treatment combination $\mathbf{x} = (x_1, \ldots, x_p)$, and the error variables $\epsilon_{\mathbf{x},t}$ are assumed to be independent with $N(0, \sigma^2)$ distributions. The parameter β_i is a measure of the local *linear effect* of the ith factor ($i = 1, \ldots, p$).

We often code the levels of each factor in each local experiment so that zero represents the *midrange* of the levels of the factor (the average of the highest and lowest levels included in the experiment) and $+1$ and -1 represent the highest and lowest levels of the factor, respectively. For the ith factor, such coded levels z_i are obtained as

$$z_i = (x_i - m_i)/h_i \,, \tag{16.2.2}$$

where m_i denotes the midrange of the values of x_i of the ith factor, and h_i denotes the *half-range*— half of the range. So, in terms of coded levels, the center of the design corresponds to the point $\mathbf{z}_0 = (0, 0, \ldots, 0)$.

The first-order model (16.2.1) can be rewritten in terms of the coded factor levels as follows:

$$Y_{\mathbf{z},t} = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_p z_p + \epsilon_{\mathbf{z},t} \,. \tag{16.2.3}$$

The parameters in models (16.2.1) and (16.2.3) are related, since

$$\beta_0 = \gamma_0 - \sum_i m_i \gamma_i / h_i \quad \text{and} \quad \beta_i = \gamma_i / h_i \ (i = 1, 2, \ldots, p) \,.$$

A design for estimating the parameters of a first-order model is called a *first-order design*. A first-order design should (i) allow for efficient estimation of each linear effect β_i or γ_i, (ii) allow a test for lack of fit of the first-order model, and (iii) be expandable to a good second-order design.

As long as there is no significant model lack of fit but there are significant linear effects, the fitted first-order model can be used to estimate the path of steepest ascent. If there is significant lack of fit of the first-order model, then additional observations may be collected to augment the first-order design so that a second-order polynomial regression model can be fitted to the data.

If there is no significant model lack of fit and also no significant linear effects, then more data may be needed to increase precision of the parameter estimators. Alternatively, the experimenters may need to change the factors under study or increase the range of levels.

16.2.2 Standard First-Order Designs

Throughout the rest of Sect. 16.2, we consider designs which we refer to as *standard first-order designs*. These designs consist of n_f "factorial points" and n_0 "center points." The *factorial points* consist of the treatment combinations of a 2^p factorial experiment run as a completely randomized design as in Chap. 7, or a 2^{p-s} fractional factorial design of Resolution III or higher. The *center points* are observations collected at the center of the local region under study; that is, at $z_0 = (0, 0, \ldots, 0)$. These are needed to provide error degrees of freedom and to provide adequate power for a test for model lack of fit.

Standard first-order designs are *orthogonal*, which means that

(i) for each factor, the sum of the coded levels used in the design is zero, ($\sum z_i = 0$, summing over observations), so half of the n_f factorial points in the design have each factor at its high level and the other half have each factor at its low level; and

(ii) for each pair of factors, the sum of cross products of the coded levels in the design is zero ($\sum z_i z_j = 0$, summing over observations).

The factorial portion of the design is chosen to be at least Resolution III so that the linear effects can be estimated. Higher resolution allows model lack of fit due to two-factor interaction effects to be tested. The 2^{p-s} orthogonal fractional factorial designs and the Plackett–Burman designs of Chap. 15 are the most efficient designs for estimation of the linear effects.

16.2.3 Least Squares Estimation

The method of least squares (as shown in optional Sect. 8.3) is used to fit a first-order model to the data. Denote the fitted model by

$$\hat{y}_{\mathbf{x}} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \tag{16.2.4}$$

or, in coded form,

$$\hat{y}_{\mathbf{z}} = \hat{\gamma}_0 + \hat{\gamma}_1 z_1 + \cdots + \hat{\gamma}_p z_p . \tag{16.2.5}$$

If a standard first-order design is used, with the extreme levels of each factor coded as $+1$ and -1, then the least squares estimator $\hat{\gamma}_i$ of the linear effect γ_i of the ith factor is

$$\hat{\gamma}_i = (\overline{Y}_{z_i(+1)} - \overline{Y}_{z_i(-1)})/2 , \tag{16.2.6}$$

where $\overline{Y}_{z_i(+1)}$ and $\overline{Y}_{z_i(-1)}$ denote the averages of the response variables at the high and the low level of the ith factor, respectively. The parameter $2\gamma_i$ denotes the change in the mean response between the high and low levels of the ith factor. This is the same as the main-effect contrast for the ith factor. The least squares estimator of β_i in the uncoded model is $\hat{\beta}_i = \hat{\gamma}_i / h_i$, where h_i is the half-range of the uncoded levels of the ith factor.

Example 16.2.1 `Paint experiment, continued`

Several experiments were run in Germany by Eibl et al. (1992) on the thickness of a paint coating. The first experiment in the series was examined in Exercise 7 of Chap. 15. To study how to decrease the mean thickness, the experimenters selected the following six factors, each at two levels:

$$A : \text{ belt speed} \qquad B : \text{ tube width } C : \text{ pump pressure}$$
$$D : \text{ paint viscosity } E : \text{ tube height } F : \text{ heating temperature}$$

They used a 2_{III}^{6-3} fractional factorial design with defining relation

$$I = BCD = ADE = ABCE = ABF = ACDF = BDEF = CEF .$$

Table 16.1 Paint thickness for the paint experiment

z_A	z_B	z_C	z_D	z_E	z_F	$y_{\mathbf{z}1}$	$y_{\mathbf{z}2}$	$y_{\mathbf{z}3}$	$y_{\mathbf{z}4}$	$\overline{y}_{\mathbf{z}.}$	$s_{\mathbf{z}}^2$
1	−1	1	−1	−1	−1	1.09	1.12	0.83	0.88	0.9800	0.021400
−1	−1	1	−1	1	1	1.62	1.49	1.48	1.59	1.5450	0.004967
1	1	−1	−1	−1	1	0.88	1.29	1.04	1.31	1.1300	0.042867
−1	1	−1	−1	1	−1	1.83	1.65	1.71	1.76	1.7375	0.005825
−1	−1	−1	1	−1	1	1.46	1.51	1.59	1.40	1.4900	0.006467
1	−1	−1	1	1	−1	0.74	0.98	0.79	0.83	0.8350	0.010700
−1	1	1	1	−1	−1	2.05	2.17	2.36	2.12	2.1750	0.017633
1	1	1	1	1	1	1.51	1.46	1.42	1.40	1.4475	0.002358

Source Eibl et al. (1992). Reprinted with Permission from Journal of Quality Technology © 1992 ASQ, www.asq.org

The experimenters decided to ignore all interactions for this first experiment. Since they wanted to monitor the variation of the paint thickness, they took four observations on each of the 8 treatment combinations in the fraction. The data are shown in Table 16.1, with factor levels coded as −1 and 1. If the data were collected in a completely random order, model (16.2.5) can be fitted using the 32 individual observations.

Using $z_A, \ldots, z_F$ rather than $z_1, \ldots, z_6$ to denote the factor levels, the fitted first-order model for the mean response is

$$\hat{y}_{\mathbf{z}} = \hat{\gamma}_0 + \hat{\gamma}_A z_A + \cdots + \hat{\gamma}_F z_F$$
$$= 1.42 - 0.32 z_A + 0.21 z_B + 0.12 z_C + 0.07 z_D - 0.03 z_E - 0.01 z_F \,,$$

where, for example, the parameter estimate $\hat{\gamma}_D$ is calculated as

$$\hat{\gamma}_D = (\overline{y}_{z_D(+1)} - \overline{y}_{z_D(-1)})/2 = (1.493125 - 1.348125)/2 = 0.0725 \approx 0.07 \,,$$

where $\overline{y}_{z_D(+1)}$ is the average of the observations with D at its high level and $\overline{y}_{z_D(-1)}$ is the average of the observations with D at its low level.   □

16.2.4 Checking Model Assumptions

Before progressing with the analysis of the fitted model, the model assumptions should be checked. We shall discuss tests for model lack of fit in Sect. 16.2.6.

If there is no model lack of fit, then the error assumptions may be checked using residual plots. If the observations were collected sequentially in a known run order, then the residuals are plotted against run order to check for independence of observations. Residuals are plotted against predicted values to assess equality of error variances. Normality is checked by plotting residuals versus normal scores.

16.2.5 Analysis of Variance

Suppose that a standard first-order design has been used, with the extreme levels of each factor coded as −1 and +1. Under the first-order model, it follows from Eq. (16.2.6) that

Table 16.2 Analysis of variance for a first order model for the paint experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value	Expected mean square
A	1	3.2640	3.2640	242.07	0.0001	$\sigma^2 + 32\gamma_A^2$
B	1	1.3448	1.3448	99.73	0.0001	$\sigma^2 + 32\gamma_B^2$
C	1	0.4560	0.4560	33.82	0.0001	$\sigma^2 + 32\gamma_C^2$
D	1	0.1540	0.1540	11.42	0.0024	$\sigma^2 + 32\gamma_D^2$
E	1	0.0221	0.0221	1.64	0.2127	$\sigma^2 + 32\gamma_E^2$
F	1	0.0066	0.0066	0.49	0.4902	$\sigma^2 + 32\gamma_F^2$
Error	25	0.3371	0.0135			σ^2
Total	31	5.5846				

<div align="center">Computational formulae</div>

$ssi = n_f \hat{\gamma}_i^2 = n_f (\overline{y}_{z_i(+1)} - \overline{y}_{z_i(-1)})^2 / 4, \quad$ for $i = A, B, C, \ldots$

ssE by subtraction $\qquad\qquad\qquad\qquad sstot = \sum_{\mathbf{z}} \sum_t y_{\mathbf{z}t}^2 - n\overline{y}_{..}^2$

$$\mathrm{Var}(\hat{\gamma}_i) = \left(\frac{\sigma^2}{n_f/2} + \frac{\sigma^2}{n_f/2} \right) / 4 = \sigma^2 / n_f \,,$$

for any $i = A, B, C, \ldots$. The sum of squares for testing that the main-effect contrast γ_A is zero (that is, $H_0^A : \gamma_A = 0$ against $H_A^A : \gamma_A \neq 0$) is

$$ssA = \hat{\gamma}_A^2 / (1/n_f) = n_f \hat{\gamma}_A^2 \,,$$

and since there is only one degree of freedom for the A contrast, $msA = ssA$. For the first-order model and a standard first-order design, we have expected mean square

$$E[MSA] = n_f E[\hat{\gamma}_A^2] = n_f \mathrm{Var}(\hat{\gamma}_A) + n_f (E[\hat{\gamma}_A])^2 = \sigma^2 + n_f \gamma_A^2 \,.$$

It can also be shown that $msE = ssE/(n - p - 1)$ is an unbiased estimate of σ^2, where ssE is obtained by subtraction in the analysis of variance table. Consequently, the decision rule for testing H_0^A against H_A^A is

$$\text{reject } H_0^A \quad \text{if} \quad msA/msE > F_{1,n-p-1,\alpha} \,.$$

Similar formulae hold for each main effect. The analysis of variance for the first-order model and a standard first-order design will be illustrated for the paint experiment.

Example 16.2.2 `Paint experiment, continued`

The paint experiment was discussed in Example 16.2.1, and the data were given in Table 16.1. The purpose of the experiment was to study the effects of six factors on paint thickness. The experimental design consisted of four observations on each of the treatment combinations of a 2_{III}^{6-3} design, which is an orthogonal factorial design with $n_f = 32$ factorial points and no center points. The analysis of variance for a first order model is shown in Table 16.2, together with the expected mean squares. The linear effect of each of factors A, B, C, and D is significantly different from zero, but factors E and F appear to have little effect on the response.   □

16.2.6 Tests for Lack of Fit

A first-order design allows the experimenter to determine when the first-order model is no longer adequate, provided that there are more design points than first-order model parameters, and the design includes replication at one or more points. There is said to be model *lack of fit* when the model does not adequately represent the mean response as a function of the factor levels. Lack of fit of the first-order model occurs when the local response surface is no longer a plane.

Generic Test

Let n_d denote the number of *distinct* coded treatment combinations $\mathbf{z}$. For each treatment combination for which there is replication, the sample variance $s_{\mathbf{z}}^2$ of the $n_{\mathbf{z}}$ observations at that treatment combination provides an unbiased estimate of the error variance σ^2, whether or not there is model lack of fit. These sample variances can be pooled together to obtain a *sum of squares for pure error*

$$ssPE = \sum_{\mathbf{z}} (n_{\mathbf{z}} - 1)s_{\mathbf{z}}^2 \qquad\qquad (16.2.7)$$

with $n - n_d$ degrees of freedom, giving a *mean square for pure error*

$$msPE = ssPE/(n - n_d)\,,$$

with $E[msPE] = \sigma^2$. The error sum of squares ssE with $n - p - 1$ degrees of freedom is obtained from fitting the first-order model (Table 16.2), and the difference

$$ssLOF = ssE - ssPE \qquad\qquad (16.2.8)$$

is called the *sum of squares for lack of fit*. The corresponding mean square is

$$msLOF = ssLOF/(n_d - p - 1)\,.$$

The expected mean square is $E[msLOF] = \sigma^2 + \theta^2$, where θ^2 is a quadratic function of any higher order parameters that are estimable due to the inclusion of more design points than needed to fit the first order model. Then the ratio

$$msLOF/msPE$$

is used to test the null hypothesis of no model lack of fit. The null hypothesis is rejected at level α if this ratio exceeds $F_{n_d-p-1,n-n_d,\alpha}$. This lack-of-fit test is summarized in Table 16.3.

Example 16.2.3 `Paint experiment, continued`

The paint experiment was described in Example 16.2.1. The analysis of variance for the first-order model was shown in Table 16.2, giving $ssE = 0.3371$ with 25 degrees of freedom. There were $n_{\mathbf{z}} = 4$ observations at each of eight factorial points, and the corresponding eight sample variances, each with three degrees of freedom, were given in Table 16.1, p. 570. These eight sample variances can be pooled together to obtain

$$ssPE = \sum_{\mathbf{z}} (4 - 1)s_{\mathbf{z}}^2 = 0.3367$$

Table 16.3 Generic lack-of-fit test for the first-order model

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	Expected mean square
Lack of fit	$n_d - p - 1$	$ssLOF$	$msLOF$	$msLOF/msPE$	$\sigma^2 + \theta^2$
Pure Error	$n - n_d$	$ssPE$	$msPE$		σ^2
Error	$n - p - 1$	ssE			

<div align="center">Computational formulae</div>

ssE from Table 16.2, $ssPE = \sum_{\mathbf{z}}(n_{\mathbf{z}} - 1)s_{\mathbf{z}}^2$, $ssLOF$ by subtraction,

n_d distinct design points, n observations total, p factors,

θ depends on the nature of estimable model lack of fit

Table 16.4 Generic lack-of-fit test for the paint experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Lack of fit	1	0.0004	0.0004	0.03	0.8594
Pure error	24	0.3367	0.0140		
Error	25	0.3371	0.0135		

based on $n - n_d = 32 - 8 = 24$ degrees of freedom. The sum of squares for lack of fit is

$$ssLOF = ssE - ssPE = 0.3371 - 0.3367 = 0.0004 \,,$$

and the test is summarized in Table 16.4. Since the p-value is large, there is no evidence of lack of fit of the first-order model. □

Test for Second-Order Lack of Fit

If the generic test indicates lack of fit of the first-order model, this provides no insight into why the model is not fitting well. To understand the nature of the lack of fit, it can be helpful to consider what the mean square for lack of fit measures in terms of higher-order models. If the first-order model is inadequate, the next possibility is that a second-order model would provide an adequate approximation to the local response surface. If so, then lack of fit of the first-order model is attributable to the presence of either two-factor interactions or to quadratic effects or to both.

If the only lack of fit is due to two-factor interaction effects, this corresponds to a twisting of the response surface. Such lack of fit can be tested if the first-order design allows estimation of two-factor interactions in addition to providing error degrees of freedom. In the paint experiment, for example, it is possible to estimate the AC interaction effect, in addition to the six main effects, provided that all other interaction effects are known to be negligible.

If the center of the experimental design is near the peak of the response surface, then one would expect quadratic effects, or curvature, to be present and a higher mean response near the design center than at the factorial points. Multiple center points $\mathbf{z}_0 = (0, \ldots, 0)$ are usually included in a first-order design, because comparison of the mean response at the center of the design region with the mean response at the factorial points provides an effective test for lack of fit due to quadratic effects.

So, to assess second-order lack of fit we fit a second-order polynomial regression model under the alternative hypothesis. With respect to the coded factor levels, the standard *second-order model* for p factors is

$$Y_{\mathbf{z},t} = \gamma_0 + \sum_i \gamma_i z_i + \sum_i \gamma_{ii} z_i^2 + \sum_{i<j} \gamma_{ij} z_i z_j + \epsilon_{\mathbf{z},t} \,,$$

where the parameter γ_i represents the linear effect of the ith factor, γ_{ii} represents the quadratic effect of the ith factor, and γ_{ij} represents the cross product effect between the ith and jth factors.

If the factorial portion of the standard first-order design is either a complete factorial design or a fraction of resolution V or higher, then all two-factor interaction parameters γ_{ij} in the second-order model are estimable (assuming higher-order interactions to be negligible). For testing for second-order lack of fit, we add the sums of squares for these two-factor interactions to obtain a pooled interaction sum of squares, *ssI*. If the factorial portion of the design is a fraction of resolution less than V, then not all two-factor interactions are estimable, and only the sums of squares of those two-factor interactions which are not aliased with main effects may be pooled—one sum of squares from each alias set.

The quadratic-effect parameters are not individually estimable from a standard first-order design. They are aliased with one another, and only their sum can be estimated. It can be shown that

$$E[\overline{Y}_f - \overline{Y}_0] = \sum_{i=1}^p \gamma_{ii} \,,$$

with

$$\mathrm{Var}(\overline{Y}_f - \overline{Y}_0) = \left(\frac{1}{n_0} + \frac{1}{n_f}\right)\sigma^2 = \frac{n}{n_f n_0}\sigma^2 \,,$$

where $\overline{Y}_f$ and $\overline{Y}_0$ denote the average of the n_f factorial points and the average of the n_0 center points, respectively. It follows that the corresponding sum of squares for testing whether or not the sum of the quadratic parameters is zero is

$$ssQ = \frac{n_f n_0}{n}(\overline{Y}_f - \overline{Y}_0)^2 \,,$$

with one degree of freedom. The expected mean square is

$$E[MSQ] = \sigma^2 + \frac{n_f n_0}{n}\left(\sum_{i=1}^p \gamma_{ii}\right)^2 .$$

In the generic test for lack of fit of the first-order model, *ssI* and *ssQ* are part of *ssLOF*. Thus, we can write

$$ssLOF = ssI + ssQ + ssH,$$

where *ssH* is the sum of squares for lack of fit due to a higher-order model. Then lack of fit due specifically to interaction terms and quadratic terms can be investigated separately. The tests are summarized in Table 16.5 for a standard first-order design.

For all tests for lack of fit, an adequate number of pure error degrees of freedom are needed for the test power to be reasonably high. Since $\mathrm{Var}(\overline{Y}_f - \overline{Y}_0) > \sigma^2/n_0$, the test for lack of fit due to quadratic effects will have low power if there are few center points. Typically, 3–6 center points would be used.

Example 16.2.4 Acid copper pattern plating experiment

Poon (1995) conducted a sequence of fractional factorial and response surface experiments each involving as many as seven factors to minimize the coating thickness variation of an acid copper-plating

Table 16.5 Lack-of-fit test for the first-order model, given the data of a standard first-order design, with p factors $A, B, \ldots$ and m alias sets for interaction effects clear of main effects

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	Expected mean square
Interaction	m	$ssI = ssAB + \cdots$	msI	$\dfrac{msI}{msPE}$	$\sigma^2 + \dfrac{n_f}{m}\theta_1$
Quadratic	1	ssQ	msQ	$\dfrac{msQ}{msPE}$	$\sigma^2 + \dfrac{n_0 n_f}{n}\theta_2^2$
Higher-order	$n_d - p - m - 2$	ssH			
Pure Error	$n - n_d$	$ssPE$	$msPE$		σ^2
Error	$n - p - 1$	ssE			

Computational formulae	
$ssAB = n_f \hat{\gamma}_{AB}^2 = n_f(\overline{y}_{z_A z_B(+1)} - \overline{y}_{z_A z_B(-1)})^2/4$	ssE from Table 16.2
$ssQ = (n_0 n_f/n)^2(\overline{y}_f - \overline{y}_0)^2$	$\theta_1 = \gamma_{AB}^2 + \cdots$
$ssPE = \sum_{\mathbf{z}}(n_{\mathbf{z}} - 1)s_{\mathbf{z}}^2$	$\theta_2 = \gamma_{AA} + \gamma_{BB} + \cdots$
$ssH = (ssE - ssPE) - ssI - ssQ$	

Table 16.6 Data for the acid copper pattern plating experiment

Anode–cathode separation (in.)		Current density (asf)		Standard deviation (μm)
Coded	Uncoded	Coded	Uncoded	
-1	9.5	-1	31	5.60
-1	9.5	1	41	6.45
1	11.5	-1	31	4.84
1	11.5	1	41	5.19
0	10.5	0	36	4.32
0	10.5	0	36	4.25

Source Poon (1995). Reprinted with permission

process. In the final experiments, conducted in the vicinity of minimum thickness variation, response surface methods were utilized to study the effects of anode-cathode separation (factor A) and cathodic current density (factor B) on the standard deviation of coating thickness. One experiment used the factorial points of a single replicate 2^2 design, augmented by two center points. The response was the standard deviation (in μm) of copper-plating thickness. The coded and uncoded factor levels, together with the resulting data, are given in Table 16.6.

The midrange of levels of factor A is $(11.5 + 9.5)/2 = 10.5$, and the half-range is $(11.5 - 9.5)/2.0 = 1.0$. So the coded levels are given by

$$z_A = x_A - 10.5.$$

The midrange and half-range of the factor B levels are $(41 + 31)/2 = 36$ and $(41 - 31)/2 = 5$, respectively, so the coded levels of factor B are

$$z_B = (x_B - 36)/5.$$

Table 16.7 shows the analysis of variance, including tests for lack of fit, due to a second-order model. The analyses are identical for coded and uncoded factor levels. There are significant quadratic effects—an indication that quadratic terms for either or both of factors A and B are needed to adequately model the response surface. The first-order design is inadequate, then, because not all parameters in

Table 16.7   Analysis of variance and lack-of-fit test for the acid copper pattern plating experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value	Expected mean square
A	1	1.0201	1.0201	1.46	0.3137	$\sigma^2 + n_f \gamma_A^2$
B	1	0.3600	0.3600	0.51	0.5250	$\sigma^2 + n_f \gamma_B^2$
Error	3	2.0986	0.6995			
Total	5	3.4787				
Interaction AB	1	0.0625	0.0625	25.51	0.1244	$\sigma^2 + n_f \gamma_{AB}^2$
Quadratic	1	2.0336	2.0336	830.05	0.0221	$\sigma^2 + \frac{n_0 n_f}{n} \theta^2$
Pure error	1	0.0025	0.0025			
Error	3	2.0986	0.6995			
where $\theta = \gamma_{AA} + \gamma_{BB}$						

the second-order model are estimable. The solution is to collect some additional observations, as will be illustrated in Example 16.3.1.   □

16.2.7 Path of Steepest Ascent

If there are significant linear effects and there is no significant lack of fit of the first-order model, then the *path of steepest ascent* may be followed to climb towards the maximum of the response surface.

Given the fitted first-order regression model (16.2.5), the path of steepest ascent from the current position $\mathbf{z}_a$ is determined as follows. If $\hat{\gamma}_i$ is positive, increase z_i to increase predicted mean response $\hat{y}_{\mathbf{z}}$. If $\hat{\gamma}_i$ is negative, decrease z_i to increase $\hat{y}_{\mathbf{z}}$. To follow the path of steepest ascent up the fitted response surface, change each z_i in proportion to the magnitude of $\hat{\gamma}_i$. So, if the value z_1 of the first factor is changed by $u\hat{\gamma}_1$ units for some real number u, then the level z_i of the ith factor should be changed by $u\hat{\gamma}_i$ for each other factor i.

The path of steepest ascent is defined above with respect to the coded variables. This presumes that the original variables have been coded in such a way to make the coded scales in some sense comparable. Since the original variable may be measured on scales that are not directly comparable, there is some art to the scaling of the coded variables.

Example 16.2.5 `Paint experiment, continued`

The paint experiment was described in Example 16.2.1, p. 569. The experimenters conducted an experiment to study how to decrease the thickness of a paint coating from about 2 mm to the target 0.8 mm. Four observations were taken at each treatment combination of a 2_{III}^{6-3} design and are shown in Table 16.1, p. 570.

The target thickness is approximately achieved at the experimental design point $\mathbf{z} = (+1, -1, -1, +1, +1, -1)$ so perhaps no further analysis or experimentation is needed. Nevertheless, we will use these data to illustrate how to move efficiently towards the lower target response surface value.

Since a lower mean response is required, we need to identify the path of steepest *descent*. The fitted first-order model is obtained from Example 16.2.1 as

$$\hat{y}_{\mathbf{z}} = \hat{\gamma}_0 + \hat{\gamma}_A z_A + \cdots + \hat{\gamma}_F z_F$$
$$= 1.42 - 0.32 z_A + 0.21 z_B + 0.12 z_C + 0.07 z_D - 0.03 z_E - 0.01 z_F .$$

The analysis of variance conducted in Example 16.2.2 suggests that only factors A, B, C, and D significantly affect the response. So, these four factors should be adjusted in an attempt to reduce paint thickness.

Based on the signs of the parameter estimates in the fitted model, we ought to be able to effect a reduction in mean response if we increase the level of factor A and decrease the level of any of factors B, C, and D. To follow the path of steepest descent, we change the levels of these factors each in proportion to the magnitude of its corresponding parameter estimate, $\hat{\gamma}_i$. So, if we increase z_A by $0.32u$ units for some real number u, then we decrease z_B by $0.21u$ units, decrease z_C by $0.12u$ units, and decrease z_D by $0.07u$ units.

Observations along the path of steepest descent moving away from the center of the current design, $\mathbf{z}_0 = (0, 0, 0, 0, 0, 0)$, consist of treatment combinations $(0.32u, -0.21u, -0.12u, -0.07u, 0, 0)$ corresponding to increasing values of u, such as $u = 3, 3.5, 4, \ldots$. The suggested values of u start at $u = 3$. This value is large enough to move the level of factor A near to the edge of the region of the current local experiment and corresponds to $\hat{y} = 0.9226$. For the value $u = 4$, the extrapolated prediction of the first order model is $\hat{y} = 0.7568$, already less than the target value of 0.8, making the step sizes reasonable or perhaps a bit too large. Certainly, other values of u could also have been chosen. Observations may then be collected along this path setting u equal to each value in turn until the target thickness is achieved, or until the response stops decreasing before reaching the target level. In the latter case, at the point of lowest response along the path another first-order design could be run to determine a new path of steepest descent.   □

In the previous example, the effects of factors E and F were not found to be significantly different from zero, so their levels were not changed in following the estimated path of steepest descent. There are a variety of reasons why the effect of a factor may be negligible. The obvious reason is that response is independent of the factor. However, it could also be that the levels used for the factor may be near the optimal value, so the response surface may be relatively flat with respect to small changes in the level of that factor. Alternatively, the levels of the factor may simply be too close together to give rise to a detectable change in the mean response. In subsequent experiments, the levels of such factors can be chosen farther apart to guard against the last scenario.

16.3   Second-Order Designs and Analysis

16.3.1 Models and Designs

Second-order designs and analysis are used when the test for lack of fit of the first-order model indicates that the vicinity of the maximum (or minimum) of the response surface has been reached and a second-order model should be fitted. For p factors, the standard second-order model is

$$Y_{\mathbf{x},t} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \sum_{i=1}^{p} \beta_{ii} x_i^2 + \sum_{i<j} \beta_{ij} x_i x_j + \epsilon_{\mathbf{x},t} , \qquad (16.3.9)$$

where $Y_{\mathbf{x},t}$ denotes the tth response observed for treatment combination $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. The random-error variables $\epsilon_{\mathbf{x},t}$ are assumed to be independent with $N(0, \sigma^2)$ distributions. The parameter β_i represents the linear effect of the ith factor. The parameter β_{ii} represents the quadratic effect of the ith factor, and β_{ij} represents the cross product effect, or interaction effect, between the ith and jth factors.

With respect to the coded factor levels $z_i = (x_i - m_i)/h_i$, the second-order model is

$$Y_{\mathbf{z},t} = \gamma_0 + \sum_{i=1}^{p} \gamma_i z_i + \sum_{i=1}^{p} \gamma_{ii} z_i^2 + \sum_{i<j} \gamma_{ij} z_i z_j + \epsilon_{\mathbf{z},t} \,. \tag{16.3.10}$$

Experimental designs used to fit a second-order model are referred to as *second-order designs*. A second-order design should (i) allow for efficient estimation of the response surface, in the sense of having $\text{Var}(\widehat{Y}_{\mathbf{z}})$ be small in the design region; (ii) allow a test for lack of fit of the second-order model; and (iii) allow for efficient estimation of all model parameters. Second-order designs must have at least $(p+1)(p+2)/2$ distinct design points; otherwise, not all of the $(p+1)(p+2)/2$ parameters in the second-order model can be estimated. We will consider only such designs in this chapter. Observations at even more points are needed, plus some replication, in order to be able to conduct a generic test for model lack of fit. Other properties of second-order designs that are sometimes desirable include rotatability, orthogonality, and orthogonal blocking—these will be discussed in Sects. 16.4.1–16.4.3.

The method of least squares is used to fit the second-order model to the data. This method is exactly as discussed in optional Sect. 8.3, with each second-order term z_i^2 or $z_i z_j$ being treated as a single regressor. In terms of the uncoded and coded factor levels, the fitted models are, respectively,

$$\hat{y}_{\mathbf{x}} = \hat{\beta}_0 + \sum_i \hat{\beta}_i x_i + \sum_i \hat{\beta}_{ii} x_i^2 + \sum_{i<j} \hat{\beta}_{ij} x_i x_j \tag{16.3.11}$$

and

$$\hat{y}_{\mathbf{z}} = \hat{\gamma}_0 + \sum_i \hat{\gamma}_i z_i + \sum_i \hat{\gamma}_{ii} z_i^2 + \sum_{i<j} \hat{\gamma}_{ij} z_i z_j \,, \tag{16.3.12}$$

where the parameters with hats denote the least squares estimates. Although it is possible to obtain explicit formulae for the least squares estimates for any specific design, the formulae for the quadratic parameter estimates $\hat{\gamma}_{ii}$ are complicated. Consequently, we rely on statistical computer software to obtain the least squares estimates (see Sects. 16.7 and 16.8 for the use of the SAS and R software, respectively).

As long as there is no significant lack of fit, the fitted second-order model can be used to study the local response surface. Generally, there will be a unique treatment combination $\mathbf{x}_s = (x_{s1}, x_{s2}, \ldots, x_{sp})$, called the *stationary point*, at which the fitted surface $\hat{y}_{\mathbf{x}}$ is neither increasing or decreasing—the tangent plane is level. At the stationary point, $\hat{y}_{\mathbf{x}}$ is maximized, minimized, or is at a saddle point. The surface near a *saddle point* is reminiscent of a horse saddle—rising up from front to back but sloping down from side to side. A saddle point yields neither a maximum nor a minimum for the fitted model. Instead, these will be found at the boundary of the design region.

If there is significant lack of fit of the second-order model, a higher-order model could be used, or a more local experiment could be run.

16.3.2 Central Composite Designs

Central composite designs were first described by Box and Wilson (1951), and they are nowadays the most popular second-order designs. Each design consists of a standard first-order design with n_f orthogonal factorial points and n_0 center points, augmented by n_a "axial points."

We follow the convention of coding the factor levels so the factorial points have coded levels ± 1 for each factor. However, it should be noted that some software packages will recode the levels in a central

Fig. 16.3 Central
composite designs for
$p = 2$ and $p = 3$ factors

(a) $p = 2$ **(b)** $p = 3$
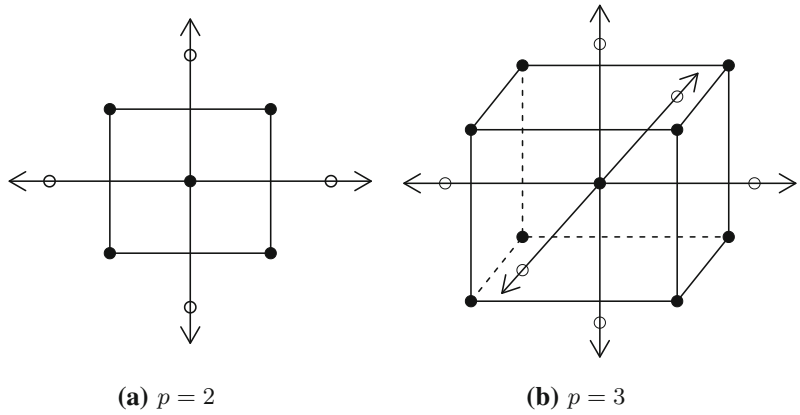
composite design before doing the analysis. In SAS, for example, the default is to code the extreme levels of each factor as ± 1, whereas R allows the user to specify the coding. Under our convention, *axial points* are points located at a specified distance α from the design center in each direction on each axis defined by the coded factor levels. On the z_i-axis, for example, two axial points are obtained by setting $z_i = \pm \alpha$, with $z_j = 0$ for all $j \neq i$. Thus, if there are p factors, there are $2p$ distinct axial points. Axial points are also commonly referred to as *star points*. Figure 16.3 shows central composite designs for $p = 2$ and $p = 3$ factors, with axial points represented by unfilled circles or balls.

A central composite design is easily built up from a standard first-order design by the addition of axial points, and possibly some extra factorial and center points. If the factorial portion of the design is a complete factorial or a fractional factorial of resolution V or more, all parameters of the second-order model are estimable. Otherwise, some aliasing will occur, and some terms will need to be omitted from the second-order model. A design should include enough replication, often at the center points, to allow for a test for model lack of fit. The axial points are located at a distance α from the center of the design, where the choice of α depends on the properties required of the design. A popular choice is $\alpha = (n_f)^{1/4}$ (see Sect. 16.4.1).

Example 16.3.1 `Acid copper pattern plating experiment, continued`

In Example 16.2.4, p. 574, a standard first-order design was used to study the effects of anode–cathode separation (factor A) and cathodic current density (factor B) on the standard deviation of a copper-plating thickness. The first-order design involved the $n_f = 4$ factorial points of a single-replicate 2^2 design, augmented by $n_0 = 2$ center points. There was significant lack of fit of the first-order model, so additional observations needed to be taken in order to fit a second-order model. The experimenters augmented the first-order design with four axial points, using $\alpha \approx (n_f)^{1/4} = \sqrt{2}$, say $\alpha = 1.4142$, giving the central composite design and data shown in Table 16.8.

The second-order model is fitted by a computer regression package. In terms of the uncoded factor levels, the fitted model is given by

$$\hat{y}_{\mathbf{x}} = \hat{\beta}_0 + \hat{\beta}_A x_A + \hat{\beta}_B x_B + \hat{\beta}_{AA} x_A^2 + \hat{\beta}_{BB} x_B^2 + \hat{\beta}_{AB} x_A x_B$$
$$= 84.1990 - 8.8689 x_A - 1.7526 x_B$$
$$+ 0.4419 x_A^2 + 0.0286 x_B^2 - 0.0250 x_A x_B ,$$

and, in terms of the coded factor levels, $z_A = (x_A - 10.5)$, $z_B = (x_B - 36)/5$, the fitted model is

Table 16.8 Data for the acid copper pattern plating experiment—central composite design

Anode–cathode separation (in.)		Current density (asf)		Standard deviation (μm)
Coded	Uncoded	Coded	Uncoded	
−1.0000	9.5000	−1.0000	31.0000	5.60
−1.0000	9.5000	1.0000	41.0000	6.45
1.0000	11.5000	−1.0000	31.0000	4.84
1.0000	11.5000	1.0000	41.0000	5.19
0.0000	10.5000	0.0000	36.0000	4.32
0.0000	10.5000	0.0000	36.0000	4.25
−1.4142	9.0858	0.0000	36.0000	5.76
1.4142	11.9142	0.0000	36.0000	4.42
0.0000	10.5000	−1.4142	28.9290	5.46
0.0000	10.5000	1.4142	43.0710	5.81

Source Poon (1995). Reprinted with permission

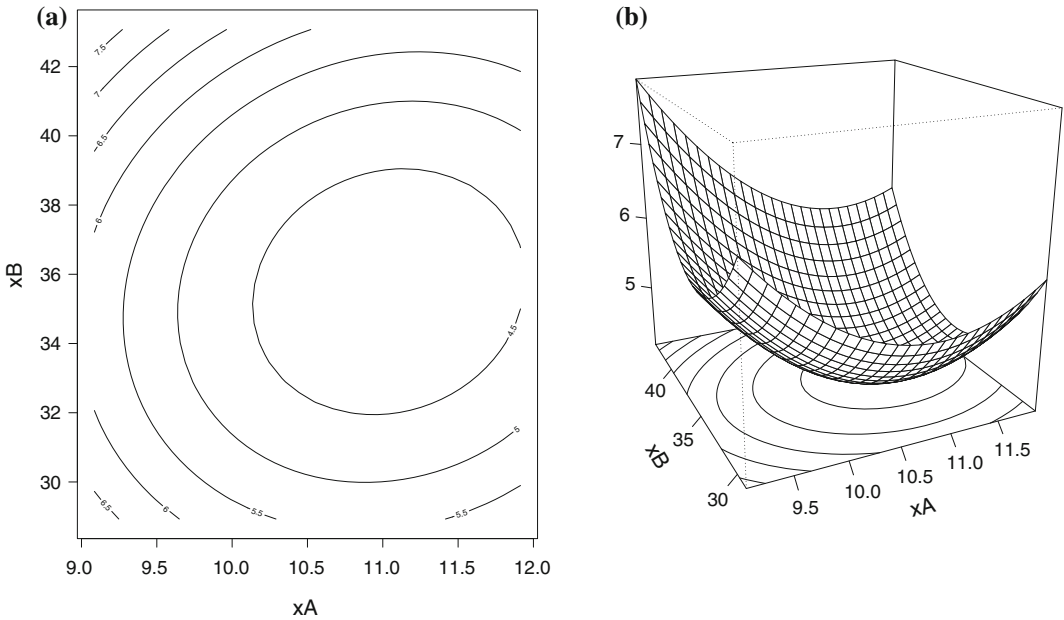**Fig. 16.4** Response surface contour plot and response surface plot of fitted second-order model for the acid copper pattern plating experiment

$$\hat{y}_{\mathbf{z}} = \hat{\gamma}_0 + \hat{\gamma}_A z_A + \hat{\gamma}_B z_B + \hat{\gamma}_{AA} z_A^2 + \hat{\gamma}_{BB} z_B^2 + \hat{\gamma}_{AB} z_A z_B$$
$$= 4.2850 - 0.4894 z_A + 0.2119 z_B$$
$$+ 0.4419 z_A^2 + 0.7144 z_B^2 - 0.1250 z_A z_B \ .$$

Figure 16.4 shows both a contour plot and a surface plot of the fitted model for uncoded factor levels. The stationary point is in the center of the ellipses. Clearly, the stationary point provides a minimum. The exact location of the stationary point will be determined in Sect. 16.3.5.   □

16.3.3 Generic Test for Lack of Fit of the Second-Order Model

If the second-order design includes n_d distinct treatment combinations, with n_d larger than the number of parameters $(p+2)(p+1)/2$, and replication at one or more of these, then a generic test for lack of fit of the second-order model can be conducted, just as for the first-order model (Sect. 16.2.6). The sum of squares for pure error, $ssPE$, and the sum of squares for lack of fit, $ssLOF$, are calculated as in (16.2.7) and (16.2.8). The error sum of squares ssE and the error degrees of freedom are obtained from the analysis of variance table of the second-order model. The test proceeds exactly as in Table 16.3 except that the error degrees of freedom are $n - [(p+2)(p+1)/2]$ and the degrees of freedom for lack of fit are then $n_d - [(p+2)(p+1)/2]$. The test will be illustrated for the acid copper-plating experiment in Example 16.3.2 in the next subsection.

16.3.4 Analysis of Variance for a Second-Order Model

Table 16.9 shows an outline analysis of variance table for a central composite design and second-order model, assuming that all parameters are estimable. The degrees of freedom associated with the linear effects have been added (pooled) together, as have those of the quadratic effects and those of the interaction (cross product) effects. Sequential, or Type I, sums of squares are listed for each of these pooled sources of variation. These include the sum of squares for all linear terms, $ss(L)$; the sum of squares for adding all quadratic terms to the model, given that all linear terms are already included, $ss(Q|L)$; and the sum of squares for adding all interaction terms to the model, given that all linear and quadratic terms are already in the model, $ss(I|L, Q)$. Using these sequential sums of squares, the analysis of variance is the same whether factor levels are coded or not. The coefficients a_i, a_{ii}, and a_{ij} listed in the expected mean squares are positive and depend on the design and the model. If coded factor levels are used, the expected mean squares would involve the parameters γ instead of the parameters β, but would have the same form.

If a central composite design is used and factor levels are coded in the usual way, the linear, quadratic and interaction sums of squares are independent of one another, so the corresponding sums of squares are the same, no matter in which order the terms are fitted. Also, the individual linear and interaction (cross product) parameters are estimated independently of one another and of the quadratic effects. The quadratic parameters are estimated independently of each other only if α and the number of center points n_0 are chosen to satisfy certain restrictions (see Sect. 16.4.2).

Table 16.9 Analysis of variance for a central composite design and second-order model

Source of variation	Degrees of freedom	Sum squares of (Type I)	Mean square (Type I)	Ratio	Expected mean square
L	p	ssL	msL	$\frac{msL}{msE}$	$\sigma^2 + \sum_i a_i \beta_i^2$
$Q|L$	p	$ss(Q|L)$	$ms(Q|L)$	$\frac{ms(Q|L)}{msE}$	$\sigma^2 + \sum_i a_{ii} \beta_{ii}^2$
$I|L, Q$	$\frac{1}{2}p(p-1)$	$ss(I|L, Q)$	$ms(I|L, Q)$	$\frac{ms(I|L,Q)}{msE}$	$\sigma^2 + \sum_{i<j} a_{ij} \beta_{ij}^2$
Error	df	ssE	msE		σ^2
Total	$n-1$	$sstot$			

Formula: $df = n - \frac{1}{2}(p+2)(p+1)$

Table 16.10 Analysis of variance for the acid copper pattern plating experiment

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
Linear	2	2.2751			
A_L	1	1.9159	1.9159	65.99	0.0012
B_L	1	0.3591	0.3591	12.37	0.0245
Quadratic	2	2.4361			
$A_Q\|B_Q$	1	0.8926	0.8926	30.74	0.0052
$B_Q\|A_Q$	1	2.3330	2.3330	80.36	0.0009
Interaction	1	0.0625	0.0625	2.15	0.2162
Error	4	0.1161	0.0290		
Total	9	4.8898			

Example 16.3.2 `Acid copper pattern plating experiment, continued`

The data for the central composite design of the acid copper pattern plating experiment were shown in Table 16.8, p. 580. The analysis of variance for the coded data is given in Table 16.10. The table shows the decomposition of the linear sum of squares with respect to the individual linear effects. Each of the two quadratic effects is shown adjusted for the other quadratic effect. If we test each hypothesis at individual level 0.01, the linear effect of factor A is significantly different from zero, as is the adjusted quadratic effect of each factor. Consequently, the model should include these three terms. We would also include the linear effect of B, since the higher-order (quadratic) term is included. The AB-interaction effect, or cross product effect, is not significantly different from zero.

Before settling on a final model, we should check the lack of fit of the second-order model. The only replication consisted of two center-point observations with values 4.32 and 4.25. The sample variance of these two observations is $s_0^2 = 0.00245$, so $ssPE = 0.00245$ with one degree of freedom. From the analysis of variance table, Table 16.10, we see that $ssE = 0.1161$ with 4 degrees of freedom. So,

$$ssLOF = ssE - ssPE = 0.1161 - 0.00245 = 0.11365$$

with $4 - 1 = 3$ degrees of freedom for lack of fit. There is significant lack of fit of the second-order model if

$$msLOF/msPE > F_{3,1,\alpha},$$

for appropriate significance level α. Here,

$$msLOF/msPE = (0.11365/3)/(0.00245/1) = 15.463,$$

which is less than $F_{3,1,0.10} = 53.6$, so there is no significant lack of fit of the second-order model, and the model fitted in Example 16.3.1, p. 579, should be a reasonable approximation to the true surface in the local region under study ($9.5 \le x_A \le 11.5$; $31 \le x_B \le 41$). □

16.3.5 Canonical Analysis of a Second-Order Model

After fitting a second-order model, we need to (i) determine the location of the stationary point and (ii) characterize the stationary point as providing a response surface minimum, maximum, or saddle point. The nature of the response surface at the stationary point may be evident from contour or surface plots, as is the case in Fig. 16.4, or its characterization may be done via canonical analysis. We provide an overview and illustration of canonical analysis in this section, leaving the computations to software (see Sects. 16.7.2 and 16.8.2 for the use of the SAS and R software, respectively).

In response surface methods, it is customary to perform the canonical analysis using the model fit to the coded data. We think of each coded treatment combination $\mathbf{z}$ as a point in p-dimensional space, $\mathbf{z} = (z_1, z_2, \ldots, z_p)$. Then the stationary point that we are trying to find is the point $\mathbf{z}_s = (z_{s1}, z_{s2}, \ldots, z_{sp})$ at which the fitted response surface $\hat{y}_\mathbf{z}$ is neither increasing nor decreasing—the tangent plane is level. The stationary point can be obtained via calculus as the critical point of the fitted surface $\hat{y}_\mathbf{z}$. The stationary point $\mathbf{x}_s$ for the model fit to the uncoded data can be obtained from $\mathbf{z}_s$ by simply uncoding each factor level z_{si}. In view of Eq. (16.2.2), this uncoding is accomplished by taking $x_{si} = h_i \times z_{is} + m_i$, for $i = 1, 2, \ldots, p$.

The second step—characterizing the response surface at the stationary point as a minimum, maximum, or saddle point—may be accomplished by putting the fitted second-order response surface into canonical form. To accomplish this, we change to a new coordinate system of points in two steps. First we set $\mathbf{v} = \mathbf{z} - \mathbf{z}_s$, so that $v_i = z_i - z_{si}$ for $i = 1, 2, \ldots, p$. This moves the coordinate system so that the stationary point is at the origin with respect to the v_i-axes, so the stationary point is now $\mathbf{v}_s = (0, 0, \ldots, 0)$. The other points $\mathbf{v} = \mathbf{z} - \mathbf{z}_s$ measure position relative to the stationary point $\mathbf{z}_s$. This eliminates all linear terms from the model. As the second step, the v_i-axes are rotated to obtain w_i-axes, with the rotation chosen to eliminate the cross product terms from the model.

In terms of each of these coordinate systems, the fitted model has the following equivalent representations:

$$\hat{y}_\mathbf{z} = \hat{\gamma}_0 + \sum_{i=1}^p \hat{\gamma}_i z_i + \sum_{i=1}^p \hat{\gamma}_{ii} z_i^2 + \sum_{i<j} \hat{\gamma}_{ij} z_i z_j \,,$$

$$\hat{y}_\mathbf{v} = \hat{y}_{\mathbf{v}_s} + \sum_{i=1}^p \hat{\gamma}_{ii} v_i^2 + \sum_{i<j} \hat{\gamma}_{ij} v_i v_j \,,$$

$$\hat{y}_\mathbf{w} = \hat{y}_{\mathbf{w}_s} + \sum_{i=1}^p \hat{\lambda}_{ii} w_i^2 \,,$$

where $\hat{y}_{\mathbf{v}_s}$ and $\hat{y}_{\mathbf{w}_s}$ are equal and each denotes the predicted response at the stationary point.

The last equation is said to be in *canonical form*, and in this form, we can immediately tell whether the stationary point is a maximum, a minimum, or a saddle point. If all of the $\hat{\lambda}_{ii}$'s are negative, then the fitted model is concave down and has a maximum at the stationary point. If all of the $\hat{\lambda}_{ii}$'s are positive, then the fitted model is concave up and has a minimum at the stationary point. If some of the $\hat{\lambda}_{ii}$'s are positive and some are negative, the stationary point is a saddle point. The $\hat{\lambda}_{ii}$ are called the *canonical coefficients*.

If a specific $\hat{\lambda}_{ii}$ is relatively large in magnitude, then $\hat{y}_\mathbf{w}$ will change rapidly for changes away from the stationary point $\mathbf{w}_s = (0, 0, \ldots, 0)$ in the w_i direction. Thus, if the stationary point is a saddle point and if $\hat{\lambda}_{\ell\ell}$ is the largest positive $\hat{\lambda}_{ii}$, movement in either direction away from the stationary point along the w_ℓ-axis provides a path of steepest ascent. On the other hand, if a specific $\hat{\lambda}_{ii}$ is relatively

small in magnitude, then $\hat{y}_{\mathbf{w}}$ is relatively unaffected by changes away from the stationary point along the w_i-axis.

Example 16.3.3 `Acid copper pattern plating experiment, continued`

In Example 16.3.1, p. 579, a second-order model was fitted to data collected from a central composite design. The experiment was run in order to study the effects of anode-cathode separation (factor A) and cathodic current density (factor B) on the standard deviation of copper-plating thickness. In terms of the coded factor levels $z_A = (x_A - 10.5)$, $z_B = (x_B - 36)/5$, the fitted model was

$$
\begin{aligned}
\hat{y}_{\mathbf{z}} &= \hat{\gamma}_0 + \hat{\gamma}_A z_A + \hat{\gamma}_B z_B + \hat{\gamma}_{AA} z_A^2 + \hat{\gamma}_{BB} z_B^2 + \hat{\gamma}_{AB} z_A z_B \\
&= 4.2850 - 0.4894 z_A + 0.2119 z_B \\
&\quad + 0.4419 z_A^2 + 0.7144 z_B^2 - 0.1250 z_A z_B .
\end{aligned}
$$

The following additional results are provided without computational details, since we are leaving those to software (see Sects. 16.7.2 and 16.8.2).

The stationary point is $\mathbf{z}_s = (z_{sA}, z_{sB}) = (0.5395, -0.1011)$. Using these values of z_A and z_B in the fitted model, we obtain the predicted response at the stationary point to be $\hat{y}_{\mathbf{z}_s} = 4.1423$.

The canonical coefficients are $\hat{\lambda}_{11} = 0.7280$ and $\hat{\lambda}_{22} = 0.4282$. Since both canonical coefficients are positive, the stationary point minimizes the estimated standard deviation of coating thickness. Now, $\hat{\lambda}_{11}$ is larger than $\hat{\lambda}_{22}$—nearly twice as large—so the surface will rise more rapidly as we move away from $\mathbf{z}_s$ in the w_1 direction than in the w_2 direction.

The w_1 canonical axis consists of all points (z_A, z_B) of the form

$$
(z_{sA}, z_{sB}) = (0.5395, -0.1011) + u(-0.2134, 0.9770) .
$$

The point $(-0.2134, 0.9770)$ has been scaled to be one unit from the origin (i.e. $(-0.2134, 0.9770)$ is a vector of length one), so a unit change in u corresponds to a step of size one along the w_1-axis. Since the second component of this point is nearly one, the w_1-axis is nearly parallel to the z_2-axis, or equivalently, to the B axis. This means that the coded level of B must be controlled more precisely than the coded level of A in order to maintain a minimum response. This conclusion is suggested by examining the fitted equation, since the coefficient of z_B^2 is somewhat larger than those of z_A^2 and $z_A z_B$.

Likewise, the w_2 canonical axis consists of all points (z_A, z_B) of the form

$$
(z_{sA}, z_{sB}) = (0.5395, -0.1011) + q(-0.9770, -0.2134) ,
$$

where a unit change in q corresponds to a step of size one along the w_2-axis. Since the first component of the point $(-0.9770, -0.2134)$ has magnitude nearly one, the w_2-axis is nearly parallel to the z_A-axis (or z_1-axis).   □

The canonical analysis has been described and illustrated here in terms of the coded factor levels. The SAS and R software likewise do the canonical analysis in terms of coded factor levels, though SAS software codes the levels somewhat differently, which impacts the canonical coefficients.

Canonical Analysis Formulas (Optional)

This subsection requires the knowledge of matrices and vectors. Consider the fitted second-order model

$$\hat{y}_{\mathbf{z}} = \hat{\gamma}_0 + \sum_i \hat{\gamma}_i z_i + \sum_i \hat{\gamma}_{ii} z_i^2 + \sum_{i<j} \hat{\gamma}_{ij} z_i z_j$$

for p factors. Let $\mathbf{b}$ denote the $p \times 1$ vector of linear parameter estimates, with ith entry $\hat{\gamma}_i$. Let $\mathbf{B}$ denote the $p \times p$ matrix with ith diagonal element $\hat{\gamma}_{ii}$ and with off-diagonal (ij)th entry $\hat{\gamma}_{ij}/2$. Then the least squares fitted model can be written in matrix terms as

$$\hat{y}_{\mathbf{z}} = \hat{\gamma}_0 + \mathbf{z}'\mathbf{b} + \mathbf{z}'\mathbf{Bz}.$$

Furthermore, the stationary point is

$$\mathbf{z}_s = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b},$$

with corresponding predicted mean response

$$\hat{y}_{\mathbf{z}_s} = \hat{\gamma}_0 - \mathbf{z}_s'\mathbf{Bz}_s = \hat{\gamma}_0 + \frac{1}{2}\mathbf{z}_s'\mathbf{b}.$$

The *canonical coefficients* $\hat{\lambda}_{ii}$ are the *eigenvalues* of the matrix $\mathbf{B}$. The *eigenvectors* of $\mathbf{B}$ determine the *canonical axes*, the canonical axis w_i being the normalized eigenvector of $\mathbf{B}$ corresponding to the eigenvalue $\hat{\lambda}_{ii}$. Obtaining the canonical coefficients and canonical axes using SAS and R software will be illustrated in Sects. 16.7.2 and 16.8.2, respectively.

16.4   Properties of Second-Order Designs: CCDs

In this section we discuss some desirable properties—rotatability, orthogonality, and orthogonal blocking—of second-order designs. The discussion here focuses on central composite designs (CCDs) because their properties can be controlled by judicious choice of the number of center points n_0 and the distance α of the axial points from the design center. In addition to rotatability, orthogonal blocking, and orthogonality, a design should include enough center points (say 3–6) to provide a reasonably sensitive test for lack of fit.

16.4.1 Rotatability

A design is *rotatable* if the variance $\text{Var}(\widehat{Y}_{\mathbf{z}})$ of the predicted response is the same for all coded points $\mathbf{z} = (z_1, z_2, \ldots, z_p)$ at any given distance $d = (\sum_i z_i^2)^{1/2}$ from the design center, $\mathbf{z}_0 = (0, 0, \ldots, 0)$. Thus, there is the same amount of information about the response surface at the same distance d in any direction from the design center. This is a reasonable requirement of a design, since data are generally collected without knowing in which direction from the design center the stationary point of the fitted surface will be located.

Rotatable Central Composite Designs

Suppose we take a central composite design for p factors, with one observation at each axial point located a distance α from the design center, and with one observation at each of the n_f factorial points.

It can be shown that such a central composite design is rotatable if

$$\alpha = (n_f)^{1/4}, \tag{16.4.13}$$

and if each axial point is observed r_a times, then the requirement for rotatability becomes

$$\alpha = (n_f/r_a)^{1/4}.$$

The details can be found in the articles by Box and Hunter (1957) and Draper (1982).

Example 16.4.1 `Acid copper pattern plating experiment, continued`

In Example 16.3.1, p. 579, a central composite design was used for $p = 2$ factors. The design involved one observation at each $n_f = 4$ factorial points and $n_a = 4$ axial points, plus two center points. If the model, in terms of coded factor levels, is fitted using $\alpha = (n_f)^{1/4} = \sqrt{2}$, the design is rotatable with respect to the coded factor levels. For example, it can be verified that the estimate of the variance is

$$\widehat{\text{Var}}(\widehat{Y}_{\mathbf{z}}) = 0.0182$$

at each point $\mathbf{z} = (z_1, z_2)$ at distance $\sqrt{2}$ from the design center. This includes each factorial point and each axial point. For comparison, $\widehat{\text{Var}}(\widehat{Y}) = 0.0145$ at the center point and $\widehat{\text{Var}}(\widehat{Y}) = 0.0100$ at the points $(-1, 0)$, $(1, 0)$, $(0, -1)$, and $(0, 1)$, which are each a distance 1.0 from the design center.   □

16.4.2 Orthogonality

The second-order model (16.3.10) includes $(p + 1)(p + 2)/2$ parameters, including the intercept γ_0. A second-order design is *orthogonal* if the sums of squares, $ss(\gamma_i|\gamma_0)$ $(i = 1, 2, \ldots, p)$, $ss(\gamma_{ii}|\gamma_0)$ $(i = 1, 2, \ldots, p)$, and $ss(\gamma_{ij}|\gamma_0)$ $(1 \le i < j \le p)$, each adjusted for the intercept γ_0, are independent. In the analysis of variance of an orthogonal design, the sums of squares associated with these $(p + 2)(p + 1)/2 - 1$ parameters are independent, and do not depend on the order in which the parameters are entered into the model. Orthogonality is advantageous if the experimenter is interested in evaluating which of the linear, quadratic, and cross product effects are significantly different from zero.

Orthogonal Central Composite Designs

Suppose we take a central composite design with one observation at each of the n_f factorial points and $2p$ axial points, and with n_0 observations at the center. As shown by Khuri and Cornell (1987), p. 119, the design is orthogonal if

$$(n_f + 2\alpha^2)^2 = n_f n,$$

where n is the total number of observations; that is, $n = n_f + 2p + n_0$. So, a central composite design with n_f factorial points and $2p$ axial points can be made orthogonal by appropriate choice of α or n_0. For example, if the number of center points is fixed at n_0, then n is fixed, and a central composite design is orthogonal if

$$\alpha = \left(\frac{\sqrt{n_f n} - n_f}{2} \right)^{1/2}. \tag{16.4.14}$$

If a central composite design is to be rotatable and n_0 is not fixed, then we would choose $\alpha = (n_f)^{1/4}$, and the design would also be orthogonal if the number of center points was chosen to be

$$n_0 = 4\sqrt{n_f} + 4 - 2p\,. \tag{16.4.15}$$

This may not be achievable, since n_0 must be an integer. Rounding (16.4.15) to the nearest integer gives a rotatable design that is nearly orthogonal.

Example 16.4.2 `Flour production experiment`

In Sect. 16.5, we will consider the last of four experiments described by Tuck et al. (1993) to develop robust bread flours. This experiment was run using a central composite design for three factors, with one observation at each of $n_f = 8$ factorial points and $2p = 6$ axial points. From Eq. (16.4.15), since $\sqrt{n_f} = \sqrt{8}$ is not an integer, the design with $n_f = 8$ cannot be both orthogonal and rotatable. The experimenters used only $n_0 = 2$ center points, giving $n = 16$ observations in total. From Eq. (16.4.14), the design is orthogonal if

$$\alpha = \left(\frac{\sqrt{(8)(16)} - 8}{2} \right)^{1/2} = 1.2872\,.$$

This value of α was used by the experimenters.   □

16.4.3 Orthogonal Blocking

If a second-order design is conducted as a block design, then the second-order model (16.3.10) is modified to include additive block effects. For p factors, the model is

$$Y_{h,\mathbf{z},t} = \gamma_0 + \theta_h + \sum_{i=1}^{p} \gamma_i z_i + \sum_{i=1}^{p} \gamma_{ii} z_i^2 + \sum_{i<j} \gamma_{ij} z_i z_j + \epsilon_{h,\mathbf{z},t}\,, \tag{16.4.16}$$

where $Y_{h,\mathbf{z},t}$ denotes the tth observation at coded treatment combination $\mathbf{z} = (z_1, z_2, \ldots, z_p)$ in block h, and the error variables $\epsilon_{h,\mathbf{z},t}$ are independent with $N(0, \sigma^2)$ distributions. The parameter θ_h denotes the effect of the hth block, and the other parameters are defined as in the second-order model (16.3.10).

A design is said to have *orthogonal blocking* if the least squares estimates of the linear, quadratic, and cross product effect parameters are the same under model (16.4.16), which includes block effects, as under the model (16.3.10) without block effects; that is, the linear, quadratic, and cross product effects are estimated independently of the block effects. The primary advantage of orthogonal blocking as compared with nonorthogonal blocking is that an orthogonally blocked design gives the smallest values of $\mathrm{Var}(\widehat{Y})$, $\mathrm{Var}(\hat{\gamma}_i)$, $\mathrm{Var}(\hat{\gamma}_{ii})$, and $\mathrm{Var}(\hat{\gamma}_{ij})$. A second advantage is that a rotatable design conducted with orthogonal blocking is still rotatable.

Given a design in b blocks with orthogonal blocking, the analysis under the block design model (16.4.16) is almost the same as it would be under model (16.3.10) for the design with no blocking. However, the sum of squares for blocks is extracted from the sum of squares for error, and there are $b - 1$ degrees of freedom for blocks giving $b - 1$ fewer degrees of freedom for error. The sum of squares for blocks is

$$ss\theta = \sum_{h=1}^{b} k_h (\overline{y}_{h..} - \overline{y}_{...})^2 = \sum_{h=1}^{b} y_{h..}^2 / k_h - y_{...}^2 / n\,,$$

where $y_{h..}$ is the sum of the observations in the hth block, k_h is the size of the hth block, and $y_{...}$ is the sum of all n observations in the design.

In their 1957 article, Box and Hunter developed the following general conditions under which a second-order design can be blocked orthogonally.

(1) Each block must be a first-order orthogonal design: that is, (i) for each block and each factor i, the sum of coded levels of the factor, $\sum z_i$, is zero; and (ii) for each block and each pair of factors i and j, the sum of cross products, $\sum z_i z_j$, is zero. (Each sum is over all the observations in the block.)

(2) For each block and each factor i, the sum of squares $\sum z_i^2$ of the coded levels of the ith factor in the block must be proportional to the number of observations in the block.

Orthogonal Blocking of Central Composite Designs

For a central composite design, we first divide the observations into two blocks: an *axial-points block* consisting of the n_a axial points plus n_{0a} center points, and a *factorial-points block* consisting of the n_f factorial points plus n_{0f} center points. This division into blocks is natural if, for example, a first-order design results in lack of fit, so that axial and additional center points are added at a later date to build up to a second-order design. Each of the blocks is a first-order orthogonal design, meeting condition (1) for orthogonal blocking. Concerning condition (2), the sum of squares $\sum z_i^2$ of the coded levels of each factor is $2\alpha^2$ in the axial block and n_f in the factorial block. So, condition (2) requires that

$$\frac{2\alpha^2}{n_f} = \frac{n_a + n_{0a}}{n_f + n_{0f}} .$$

Solving for α, a central composite design has orthogonal blocking if

$$\alpha = \left(\frac{n_f(n_a + n_{0a})}{2(n_f + n_{0f})} \right)^{1/2} . \tag{16.4.17}$$

The design is also rotatable if $\alpha = (n_f)^{1/4}$, in which case we require

$$n_{0f} = (\sqrt{n_f}/2)(n_a + n_{0a}) - n_f . \tag{16.4.18}$$

If the numbers of center points, n_{0a} and n_{0f}, in the blocks can be chosen to satisfy this equation, then the design will be rotatable and can be orthogonally blocked. When this is not possible, it is preferable to maintain orthogonal blocking but to relax rotatability. To accomplish this, the numbers n_{0a} and n_{0f} can be chosen such that Eq. (16.4.18) is approximately satisfied, and then α can be computed from Eq. (16.4.17).

It is sometimes possible to block a central composite design orthogonally in more than two blocks. The axial block cannot be further subdivided, but the factorial block can sometimes be divided into 2^m factorial blocks while maintaining orthogonal blocking if the number of factorial center points n_{0f} is divisible by 2^m so the factorial center points can be equally divided among the 2^m factorial blocks. This is done by confounding interaction effects between three or more factors. Box and Hunter (1957, p. 233) provide a table of blocking arrangements for rotatable and near-rotatable central composite designs. Notice that if center points are spread across b blocks, then they provide $b - 1$ fewer pure error degrees of freedom than they would in a design that is not blocked.

Example 16.4.3 `PAH recovery experiment`

Barnabas et al. (1995) used a central composite design to study the effects of four factors—pressure, temperature, extraction time, and methanol content—on the total recovery of polycyclic aromatic hydrocarbons (PAHs) when extracted from soil. The design was composed of $n_f = 2^4 = 16$ factorial points and $n_a = 2p = 8$ axial points. Taking $\alpha = 16^{1/4} = 2$ would give a rotatable design. From Eq. (16.4.18),

$$n_{0f} = (\sqrt{16}/2)(8 + n_{0a}) - 16 = 2n_{0a},$$

so use of twice as many factorial center points as axial center points would give a rotatable design that could be orthogonally blocked.

The experimenters chose to use $n_{0a} = 2$ axial center points and $n_{0f} = 4$ factorial center points. This gave an axial block of size 10 and a factorial block of size 20. They then subdivided the factorial block into two blocks each of size 10 by confounding the four-factor interaction and including two of the four factorial center points in each factorial block. The resulting design was rotatable with orthogonal blocking. Analysis of the design is discussed in Sects. 16.7.2 and 16.8.2 using the SAS and R software packages, respectively. The design itself is shown in Tables 16.16 (p. 596) and 16.19 (p. 604), where the first ten observations comprise the first factorial block, the second ten the second factorial block, and the final ten the axial block.   □

16.5   A Real Experiment: Flour Production Experiment, Continued

Tuck et al. (1993) described a series of four related experiments, involving quality improvement in the milling industry. The collective purpose of the experiments was to develop a bread flour that would give high loaf volume despite fluctuations in the bread-making process. We consider here their fourth experiment.

Bread flour consists of wheat plus a small number of minor ingredients. Their fourth experiment was concerned with the effects of three such ingredients (labeled design factors B, C, and D) on loaf volume. An orthogonal central composite design, involving eight factorial points, six axial points, and two center points, was used. For the axial points, the value $\alpha = 1.2872$ was used to make the design orthogonal (see Example 16.4.2).

When a product consists of a mixture of ingredients, and the total volume of the mixture is held constant, the fractions associated with the ingredients in the mixture necessarily sum to one. This has implications for the model and data analysis. However, in this experiment, the minor ingredients constituted such a small portion of the mixture that the total volume did not need to remain fixed, and standard response surface methods could be used to study the design factors.

There were a number of sources of variation in the production process that constituted noise factors. The production factors were paired in order to keep the experiment small. So, noise factor G represented oven bake and proof time, noise factor J represented yeast and water level, and noise factor K represented degree of mixing and moulding pressure. Each of these composite factors had two levels, "high" and "low." A 2_{III}^{3-1} fraction in the composite noise factors was used, with defining relation

$$I = GJK.$$

The experimental design used was a product array. It included $16 \times 4 = 64$ observations—each of the 16 design factor combinations of the central composite design was observed with each of the four noise factor combinations of the noise array. Also, the noise factors were difficult to change, so each

Table 16.11 Flour production experiment: average specific volume $y_{h\mathbf{z}}$ of loaves on half-day h; $\alpha = 1.2872$

z_B	z_C	z_D	$y_{1\mathbf{z}}$	$y_{2\mathbf{z}}$	$y_{3\mathbf{z}}$	$y_{4\mathbf{z}}$	$\overline{y}_{.\mathbf{z}}$	$100\log_{10}(s_{\mathbf{z}})$
-1	-1	-1	586	399	418	404	451.75	195.36
-1	-1	1	615	411	435	421	470.50	198.60
-1	1	-1	611	422	431	439	475.75	195.63
-1	1	1	639	436	444	454	493.25	198.88
1	-1	-1	603	422	400	430	463.75	197.17
1	-1	1	622	411	425	436	473.50	199.79
1	1	-1	634	471	436	425	491.50	198.68
1	1	1	673	433	423	462	497.75	207.19
α	0	0	618	414	419	477	482.00	197.80
$-\alpha$	0	0	586	421	420	455	470.50	189.60
0	α	0	621	426	427	458	483.00	196.94
0	$-\alpha$	0	629	412	412	426	469.75	202.68
0	0	α	631	411	433	453	482.00	200.35
0	0	$-\alpha$	587	413	419	430	462.25	192.15
0	0	0	604	432	416	438	472.50	194.53
0	0	0	602	425	407	439	468.25	195.48

Source Tuck et al. (1993). Copyright © 1993 Blackwell Publishers. Reprinted with permission

noise factor combination constituted a different block, and in each block the design factor treatment combinations ($z_B z_C z_D$) were randomly ordered. Observations were collected over two days using half-days as blocks, with the four blocks collected in the order ($z_G z_J z_K$) = 111, 100, 010, 001. As a result, noise factor effects are also confounded with changes in conditions from half-day to half-day. For each observation, three loaves were baked from a single dough, then the average specific volume of the three loaves recorded. The resulting data $y_{h\mathbf{z}}$ are shown in Table 16.11.

For each of the 16 treatment combinations $\mathbf{z}$ of the central composite design in turn, the sample mean $\overline{y}_{.\mathbf{z}}$ and the log sample variance ($\times 100$) were computed from the observations $y_{h\mathbf{z}}$ in the four blocks ($h = 1, 2, 3, 4$). The effects of the design factors on these two response variables were studied separately by fitting second-order response surface regression models to each set of 16 responses.

The analysis of variance for fitting the second-order model to the response $\overline{y}_{.\mathbf{z}}$ is shown in Table 16.12. Because the design is orthogonal, the effects can be assessed for significance independently of their order of entry into the model. The only effects that are significantly different from zero at an individual significance level of 0.01 are the main effects of factors C and D. The overall significance level for the nine tests is at most 0.09. The experimenters decided also to retain the main effect of factor B, for which $p = 0.0204$. If the corresponding first-order model is fitted to $\overline{y}_{.\mathbf{z}}$, we obtain

$$\hat{\overline{y}}_{.\mathbf{z}} = 475.50 + 4.42z_B + 10.24z_C + 6.87z_D .$$

The coefficients of z_B, z_C, and z_D are all positive. Thus, increasing the level of design factors B, C, and D has a positive effect on the mean loaf specific volume.

The analysis of variance for the response $100\log_{10}(s_{\mathbf{z}})$ is shown in Table 16.13. No effects can be regarded as significantly different from zero at an individual 0.01 significance level. However, in this setting it would not be particularly bad to make a Type I error, and if we raise the individual significance level we would select the linear effects of factors B and D and the quadratic effect of C as being the important effects. If the corresponding model is fitted and the linear effect of C is also included, we obtain

Table 16.12 Flour production experiment: analysis of variance for $\overline{y}_{.\mathbf{z}}$

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio	p-value
z_B	1	221.4366	221.4366	9.77	0.0204
z_C	1	1185.3603	1185.3603	52.30	0.0004
z_D	1	533.2415	533.2415	23.53	0.0029
z_B^2	1	48.0081	48.0081	2.12	0.1958
z_C^2	1	50.4906	50.4906	2.23	0.1862
z_D^2	1	1.1997	1.1997	0.05	0.8257
$z_B z_C$	1	3.4453	3.4453	0.15	0.7101
$z_B z_D$	1	51.2578	51.2578	2.26	0.1833
$z_C z_D$	1	2.8203	2.8203	0.12	0.7363
Error	6	135.9897	22.6650		
Total	15	2233.2500			

Table 16.13 Flour production experiment: analysis of variance for $100 \log_{10}(s_{\mathbf{z}})$

Source of variation	Degrees of freedom	Sum of squares	Mean squares	Ratio	p-value
z_B	1	54.9174	54.9174	8.24	0.0284
z_C	1	0.3730	0.3730	0.06	0.8208
z_D	1	70.1514	70.1514	10.53	0.0176
z_B^2	1	0.6409	0.6409	0.10	0.7669
z_C^2	1	61.4625	61.4625	9.23	0.0229
z_D^2	1	7.8587	7.8587	1.18	0.3191
$z_B z_C$	1	8.7175	8.7175	1.31	0.2963
$z_B z_D$	1	2.6931	2.6931	0.40	0.5484
$z_C z_D$	1	4.3269	4.3269	0.65	0.4511
Error	6	39.9752	6.6625		
Total	15	251.1164			

$$100 \widehat{\log_{10}}(s_{\mathbf{z}}) = 195.19 + 0.18 z_C + 3.35 z_C^2 + 2.20 z_B + 2.49 z_D$$
$$\approx 195.19 + 3.35(z_C + 0.027)^2 + 2.20 z_B + 2.49 z_D \, .$$

Taking the two fitted models, we see that not only does the mean response increase as the levels of factors B, C, and D are increased, but so does the variability. The minimum variability with respect to factor C is achieved at $z_C = -0.027$. However, the amount of factor C in the loaf cannot be negative, and so the minimum variability is achieved when the amount of factor C, as well as factors B and D, is zero.

   The end result was that the experimenters set $z_C = 0$ to achieve low variability and adjusted the level of factor B (which has the slightly smaller effect on the variance, and may have been less costly than factor D) to raise mean response to the desired level.

16.6   Box–Behnken Designs

A central composite design has five levels for each factor, ± 1, $\pm \alpha$, 0. For a given experiment, circumstances may dictate the use of fewer levels, but at least three levels per factor are needed for quadratic terms to be estimable in the second-order model. Use of 3^p factorial designs or regular 3^{p-s} fractional factorial designs might be considered. These tend to be large, however, and the smaller ones tend to be of resolution III or IV so that two-factor interactions are confounded with main effects or other two-factor interactions. For fitting a second-order response model a different type of design, called a *Box–Behnken design*, is often preferred, since interaction parameter estimates are not completely confounded, and in many cases, these designs are considerably smaller than 3^{p-s} fractional factorial designs.

A Box–Behnken design for p factors is constructed by a composition of an incomplete block design for p treatments in b blocks of size k and a 2^k factorial design having factor levels coded $+1$ and -1. The method of composition is illustrated in Example 16.6.1. In addition to the points generated by the composition, center points must be added to the design for all model parameters to be estimable.

A list of Box–Behnken designs can be found in the article of Box and Behnken (1960). The designs have p factors with each factor observed at 3 levels, for $p = 3$–7, 9–12, and 16. The designs for $p = 4$ and 7 are rotatable, and the others are nearly rotatable. The designs for $p = 4$–7, 9, 10, 12, and 16 allow orthogonal blocking. All of the designs possess a high degree of orthogonality, the only correlation being between the estimators of the intercept and the quadratic terms.

Example 16.6.1 Construction of a Box-Behnken Design: $p = 4$

Suppose we require a second-order design for $p = 4$ factors, each observed at three levels, and with a total of 27 observations. As illustrated by Box and Behnken (1960), a Box–Behnken design can be constructed from a composition of a balanced incomplete block design in $b = 6$ blocks of size $k = 2$ and a 2^2 factorial design as follows. The balanced incomplete block design, shown below left, consists of all possible combinations of four treatment labels taken two at a time. Shown to its right are the $v = 4$ treatment combinations of a 2^2 design, with factor levels coded $+1$ and -1. These two designs are composed as follows. In each of the six blocks of the incomplete block design, the treatment labels are replaced by the symbol ± 1 and the blank "$-$" by 0 to give the Box–Behnken design represented in condensed form (and without center points) below right.

$$\begin{bmatrix} 1 & 2 & - & - \\ - & - & 3 & 4 \\ 1 & - & 3 & - \\ - & 2 & - & 4 \\ 1 & - & - & 4 \\ - & 2 & 3 & - \end{bmatrix} \text{ with } \begin{bmatrix} -1 & -1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix} \text{ gives } \begin{bmatrix} \pm 1 & \pm 1 & 0 & 0 \\ 0 & 0 & \pm 1 & \pm 1 \\ \pm 1 & 0 & \pm 1 & 0 \\ 0 & \pm 1 & 0 & \pm 1 \\ \pm 1 & 0 & 0 & \pm 1 \\ 0 & \pm 1 & \pm 1 & 0 \end{bmatrix}$$

The same design, but expanded out and augmented with three center points, is shown in Table 16.14. The first ± 1 in each row of the condensed design is replaced by the first column of levels of the 2^2 design, the second ± 1 is replaced by the second column of levels, and each 0 is replaced by a column of $v = 4$ zeros. With the addition of three center points, this gives the design with 27 treatment

Table 16.14 Box–Behnken design: $p = 4$ factors, $n = 27$ treatment combinations

$$
\begin{bmatrix}
-1 & -1 & 0 & 0 \\
-1 & 1 & 0 & 0 \\
1 & -1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & -1 & -1 \\
0 & 0 & -1 & 1 \\
0 & 0 & 1 & -1 \\
0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0
\end{bmatrix}
\quad
\begin{bmatrix}
-1 & 0 & -1 & 0 \\
-1 & 0 & 1 & 0 \\
1 & 0 & -1 & 0 \\
1 & 0 & 1 & 0 \\
0 & -1 & 0 & -1 \\
0 & -1 & 0 & 1 \\
0 & 1 & 0 & -1 \\
0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0
\end{bmatrix}
\quad
\begin{bmatrix}
-1 & 0 & 0 & -1 \\
-1 & 0 & 0 & 1 \\
1 & 0 & 0 & -1 \\
1 & 0 & 0 & 1 \\
0 & -1 & -1 & 0 \\
0 & -1 & 1 & 0 \\
0 & 1 & -1 & 0 \\
0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
$$

combinations shown as the 27 rows of Table 16.14. Although this design has the same number of treatment combinations as a 3_{IV}^{4-1} design, it does not have complete confounding of the two-factor interactions in pairs. □

In general, the composition of an incomplete block design for p treatments in b blocks of size k with a factorial design with $v = 2^k$ treatment combinations yields a Box–Behnken design for p factors with bv treatment combinations. The ith of the k treatment labels in each block is replaced by the ith of the k columns of the factorial design, and each "$-$" is replaced by a column of v zeros.

In general, if the incomplete block design is a balanced incomplete block design with $r = 3\lambda$, as in Example 16.6.1, then the resulting Box–Behnken design is rotatable—otherwise not. If there does not exist a balanced incomplete block design with $r = 3\lambda$, then one can either use a balanced incomplete block design with $r \neq 3\lambda$ or use a partially balanced incomplete block design. If a partially balanced incomplete block design is used, each pair of treatment labels must occur together in at least one block for all second-order model parameters to be estimable.

Orthogonal Blocking

Many Box–Behnken designs can be blocked orthogonally. The requirements for orthogonal blocking of a second-order design were given in Sect. 16.4.3, and these imply that a Box–Behnken design can be blocked orthogonally under either of two circumstances.

First, if the blocks of the incomplete block design in the composition can be partitioned into equirepli-cate sets, then the same partition of observations in the Box–Behnken design provides orthogonal blocking as long as the same number of center points is included in each block. Such is the case for the design of Example 16.6.1, since each pair of blocks in the balanced incomplete block design includes every treatment label exactly once. For the resulting Box–Behnken design in Table 16.14, each bracketed set of nine treatment combinations is a corresponding block with one center point included.

The second situation that allows orthogonal blocking occurs when interactions involving three or more factors can be confounded in the generating factorial design. An example follows:

Example 16.6.2 Example of orthogonal blocking

For $p = 4$ factors, the balanced incomplete block design with blocks consisting of the four combinations of three treatment labels can be combined with the 2^3 factorial design as follows.

$$\begin{bmatrix} 1 & 2 & 3 & - \\ 1 & 2 & - & 4 \\ 1 & - & 3 & 4 \\ - & 2 & 3 & 4 \end{bmatrix} \text{ with } \begin{bmatrix} -1 & -1 & -1 \\ -1 & -1 & 1 \\ -1 & 1 & -1 \\ -1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix} \text{ gives } \begin{bmatrix} \pm1 & \pm1 & \pm1 & 0 \\ \pm1 & \pm1 & 0 & \pm1 \\ \pm1 & 0 & \pm1 & \pm1 \\ 0 & \pm1 & \pm1 & \pm1 \end{bmatrix},$$

where the ith occurrence of ±1 in any row of the combined design is replaced by the ith column of the factorial design, and each 0 in the combined design is replaced by a column of eight 0's. The resulting 32-run Box–Behnken design can be partitioned into two blocks of size 16 by confounding the three-factor interaction in the generating factorial design. Thus, treatment combinations in the combined design are divided into two blocks, the division depending on whether they include an even or odd number of factors at level "-1." An equal number of center points must be added to each block. $\square$

16.7   Using SAS Software

In this section we illustrate the analysis of a standard first-order design and a central composite design using the SAS procedures GLM and RSREG, respectively.

16.7.1 Analysis of a Standard First-Order Design

The acid copper pattern plating experiment of Poon (1995) was introduced in Example 16.2.4 (p. 574). This small experiment involved four factorial points and two center points. A SAS program using the GLM procedure for the analysis of this standard first-order design is shown in Table 16.15. After reading

Table 16.15 SAS program for first-order response surface regression

```
* Enter data of the first-order design and code levels;
DATA COPPER;
  INPUT XA XB S;
  ZA = (XA - 10.5);
  ZB = (XB - 36)/5;
  LINES;
   9.5  31  5.60
   9.5  41  6.45
  11.5  31  4.84
  11.5  41  5.19
  10.5  36  4.32
  10.5  36  4.25
;
* Analysis of the first-order design;
PROC GLM;
  MODEL S = ZA ZB;
* Add model terms to test for lack of fit;
PROC GLM;
  MODEL S = ZA ZB ZA*ZB ZA*ZA;
```

the data and coding the factor levels, there are two calls of PROC GLM. Neither of these calls includes a CLASS statement, since the goal is to fit a regression model to the levels of the quantitative factors and not to compare the effects of their levels.

In the first call, the first-order model (16.2.3) is fitted, generating the output shown in Fig. 16.5. Neither main effect is significantly different from zero, indicating either that the experimental region is in the vicinity of the peak, or that neither factor affects the response.

In the second call of PROC GLM, the interaction term and one quadratic term are added to the model to test for lack of fit of the first-order model—the model would contain too many parameters if both quadratic terms were added. Some of the resulting output is shown in Fig. 16.6. At an overall level of

Fig. 16.5 SAS output from the first call of PROC GLM: analysis of variance and parameter estimates for a first-order design

The GLM Procedure
Dependent Variable: S

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.38010000	0.69005000	0.99	0.4686
Error	3	2.09858333	0.69952778		
Corrected Total	5	3.47868333			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ZA	1	1.02010000	1.02010000	1.46	0.3137
ZB	1	0.36000000	0.36000000	0.51	0.5250

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	5.108333333	0.34144980	14.96	0.0006
ZA	-0.505000000	0.41818889	-1.21	0.3137
ZB	0.300000000	0.41818889	0.72	0.5250

Fig. 16.6 SAS output from the second call of PROC GLM: test for lack of fit of the first-order model

Dependent Variable: S

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3.47623333	0.86905833	354.72	0.0398
Error	1	0.00245000	0.00245000		
Corrected Total	5	3.47868333			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ZA	1	1.02010000	1.02010000	416.37	0.0312
ZB	1	0.36000000	0.36000000	146.94	0.0524
ZA*ZB	1	0.06250000	0.06250000	25.51	0.1244
ZA*ZA	1	2.03363333	2.03363333	830.05	0.0221

Table 16.16 SAS program for response surface regression (PAH recovery experiment)

```
DATA PAH;
  INPUT RUN B1  B2 PRES TEMP   ET   MC     Y;
  LINES;
        1  1   0  250   55  47.5  15  391.8
        2  1   0  150   85  47.5  15  413.6
        3  1   0  250   55  22.5   5   68.7
        4  1   0  250   85  47.5   5  143.0
        5  1   0  150   85  22.5   5  104.0
        6  1   0  150   55  22.5  15  309.1
        7  1   0  200   70  35.0  10  400.6
        8  1   0  250   85  22.5  15  402.5
        9  1   0  150   55  47.5   5   77.7
       10  1   0  200   70  35.0  10  426.5
       11  0   1  250   85  47.5  15  457.5
       12  0   1  150   55  22.5   5   56.9
       13  0   1  250   85  22.5   5   94.1
       14  0   1  250   55  22.5  15  409.7
       15  0   1  150   55  47.5  15  410.9
       16  0   1  150   85  22.5  15  375.8
       17  0   1  150   85  47.5   5  110.5
       18  0   1  200   70  35.0  10  387.8
       19  0   1  250   55  47.5   5  103.0
       20  0   1  200   70  35.0  10  399.1
       21 -1  -1  200   70  35.0  10  416.9
       22 -1  -1  200   40  35.0  10  359.8
       23 -1  -1  200   70  10.0  10  276.1
       24 -1  -1  200   70  60.0  10  462.3
       25 -1  -1  100   70  35.0  10  311.5
       26 -1  -1  200   70  35.0  10  346.5
       27 -1  -1  200   70  35.0   0   46.8
       28 -1  -1  200   70  35.0  20  418.7
       29 -1  -1  200  100  35.0  10  413.9
       30 -1  -1  300   70  35.0  10  429.4
  ;
  * Sort by independent variables for lack of fit test;
  PROC SORT; BY B1 B2 PRES TEMP ET MC;
  * Response surface regression, including contour plots;
  ODS GRAPHICS ON; * Needed for contour plots;
  PROC RSREG PLOTS = SURFACE;
    MODEL Y = B1 B2 PRES TEMP ET MC / COVAR = 2 LACKFIT;
    RUN;
  ODS GRAPHICS OFF;
```

0.10 for the four tests (each being done at individual level $\alpha^* = 0.025$), the quadratic term ZA*ZA is significantly different from zero, indicating the presence of significant curvature. This fact caused the experimenters to add axial points to the first-order design to obtain a central composite design (see Example 16.3.1).

16.7.2 Analysis of a Second-Order Design

The SAS procedure RSREG is used to fit a second-order response surface regression model. This is illustrated in Table 16.16 in the context of the PAH recovery experiment that was introduced in Example 16.4.3 (p. 589). A rotatable central composite design with orthogonal blocking was used to study the effects of four factors—pressure (PRES), temperature (TEMP), extraction time (ET), and methanol content (MC)—on the total recovery of polycyclic aromatic hydrocarbons (Y) when extracted from soil.

The SAS program shown in Table 16.16 reads the run number, the levels of the block indicator variables, the uncoded levels of the four factors, and the data into data set ONE. Until now, we have always declared a block variable to be a classification variable via the CLASS statement and listed its levels as $1, 2, \ldots, b$. However, PROC RSREG does not recognize classification variables, and if a single block factor were included in the model, it would be interpreted as a single covariate—a quantitative variable possessing one degree of freedom. We have included in the model the pair of covariates (B1, B2), for which we have selected the three coded pairs of levels (1, 0), (0, 1) and $(-1, -1)$. The three pairs of levels distinguish the three blocks and provide two block degrees of freedom.

Only the factor *names* need be listed in the MODEL statement in RSREG, as all quadratic and cross product terms in the factors are automatically included in the model. To avoid treatment–block interactions from being included, B1 and B2 are declared to be covariates. This is done via the option COVAR = 2, which indicates that the first two listed independent variables are covariates and should not be included in any interactions.

A generic test for model lack of fit can optionally be requested if the SAS data set has been sorted by the independent variables in the model to cluster replicated observations. PROC SORT is used to sort the data, and a test for lack of fit is requested via the option LACKFIT in the model statement of PROC RSREG.

PROC RSREG codes the levels of each factor so that $+1$ and -1 represent the extreme levels of each factor. For example, the axial points of a central composite design would typically be coded ± 1 by SAS software rather than the conventional $\pm \alpha$. Figure 16.7 shows how SAS codes the factor levels, as well as the resulting analysis of variance table. The analysis of variance table includes Type I sums of squares for covariates, linear terms, quadratic terms, and cross product terms, adding the terms to the model in that order. These Type I sums of squares are the same, whether coded or uncoded factor levels are specified in the model statement. Observe that the cross product terms are not significantly different from zero, and there is no significant lack of fit of the model.

Type III sums of squares are also provided for each factor, pooling together the sums of squares for all terms—linear, quadratic and interaction—involving the factor. This information can be used for assessing whether any single factor can be removed from the model. These Type III sums of squares are also the same using either the coded or uncoded factor levels. The Type III sums of squares indicate that the factor methanol content (MC) is needed in the model, but perhaps not the other factors. Further analysis could explore what additional terms are needed, if any.

Figure 16.8 contains the parameter estimates and corresponding t-tests. Clearly the linear and quadratic methanol content terms are needed in the model, providing some clarification to the analysis of variance results.

In Fig. 16.9, the canonical analysis is shown, including the stationary point (Critical Value) in terms of both coded and uncoded factor levels; the predicted value at the stationary point; the canonical coefficients (Eigenvalues); classification of the stationary point as a maximum, minimum, or saddle point; and the direction of each canonical axis (Eigenvectors). The canonical coefficients and axes are with respect to the coded factor levels.
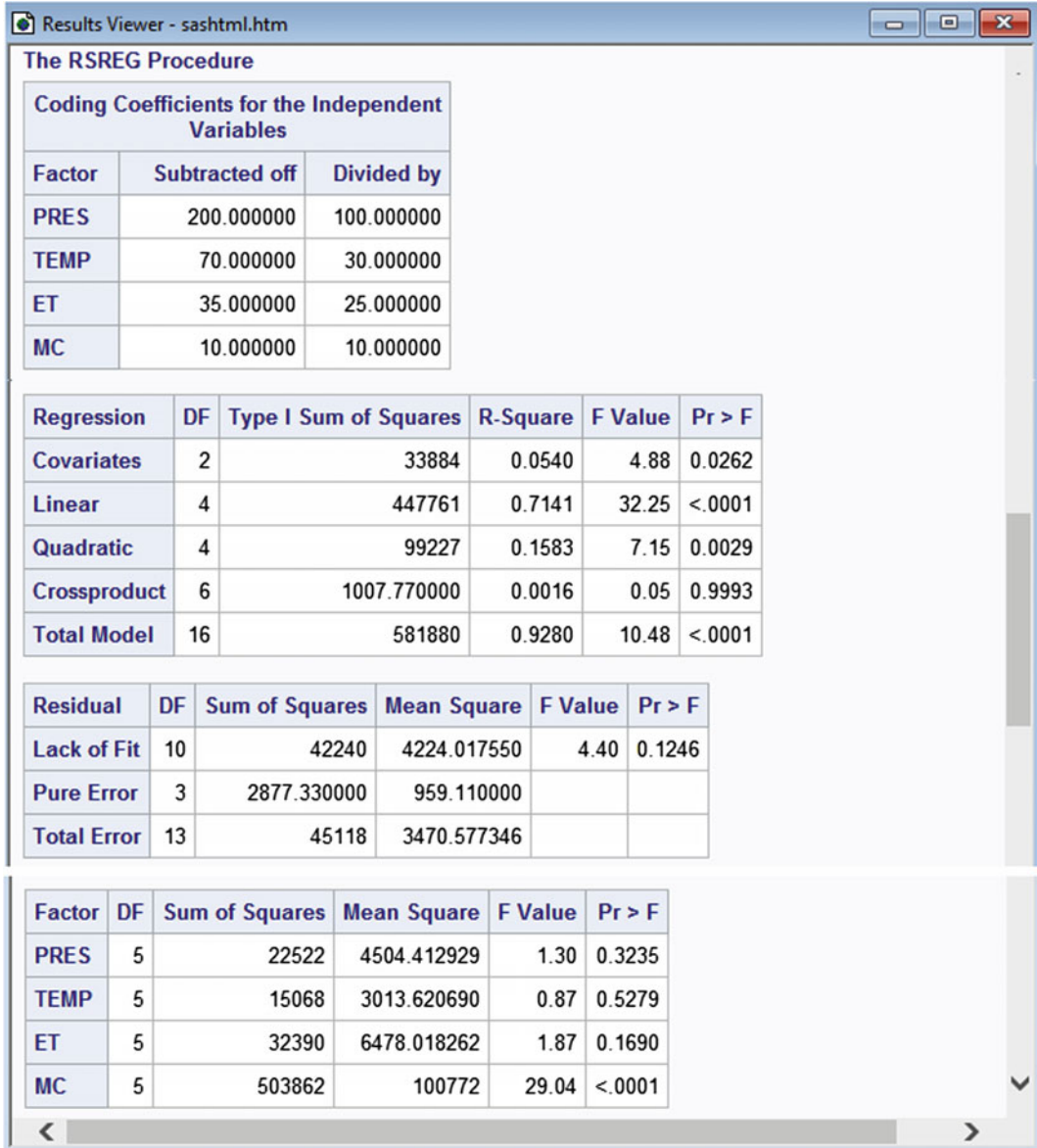
Fig. 16.7 SAS output from `PROC RSREG`: coding of factor levels, analysis of variance, and lack-of-fit test

For this experiment, all eigenvalues (canonical coefficients) are negative, so the stationary point is a maximum. The eigenvectors are each scaled to be of length one. The last eigenvalue, with value -227.865046, is the largest in magnitude. For the corresponding eigenvector, the primary component is that of MC with value 0.994301. So, the fitted model has greatest curvature at the stationary point when moving in either direction determined by this fourth eigenvector, which is nearly parallel to the MC-axis. This is evident from the contour plots in Fig. 16.10, where MC is the x-axis variable of plots (b), (e) and (f). Such a panel of contour plots is generated by inclusion of `PLOTS = SURFACE` as an option of `PROC RSREG` in Table 16.16, whereas changing the option to `PLOTS = 3D` would

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	-1238.272778	552.190337	-2.24	0.0430	396.233333
PRES	1	3.998267	2.492765	1.60	0.1327	37.300000
TEMP	1	12.755444	8.744702	1.46	0.1684	31.783333
ET	1	12.320000	9.182487	1.34	0.2027	54.966667
MC	1	65.649667	21.967351	2.99	0.0105	263.066667
PRES*PRES	1	-0.008863	0.004499	-1.97	0.0706	-88.625000
TEMP*PRES	1	-0.002117	0.019637	-0.11	0.9158	-6.350000
TEMP*TEMP	1	-0.080250	0.049994	-1.61	0.1325	-72.225000
ET*PRES	1	-0.004660	0.023565	-0.20	0.8463	-11.650000
ET*TEMP	1	0.003067	0.078549	0.04	0.9695	2.300000
ET*ET	1	-0.143800	0.071991	-2.00	0.0671	-89.875000
MC*PRES	1	0.023100	0.058912	0.39	0.7013	23.100000
MC*TEMP	1	-0.014500	0.196372	-0.07	0.9423	-4.350000
MC*ET	1	0.066200	0.235646	0.28	0.7832	16.550000
MC*MC	1	-2.263250	0.449945	-5.03	0.0002	-226.325000
B1	1	-27.073333	15.210911	-1.78	0.0985	-27.073333
B2	1	-20.293333	15.210911	-1.33	0.2051	-20.293333

Fig. 16.8 SAS output from `PROC RSREG`: parameter estimates for uncoded and coded factor levels

generate response surface plots. Note that such graphics require `ODS GRAPHICS ON`, and `PROC RSREG` must run before turning `ODS GRAPHICS OFF`.

16.8   Using R Software

In this section we illustrate the analysis of a standard first-order design and a central composite design using the response surface methods function `rsm` of the `rsm` library. We then illustrate generation of central composite and Box–Behnken designs using functions of the `rsm` package.

16.8.1 Analysis of a Standard First-Order Design

The acid copper pattern plating experiment of Poon (1995) was introduced in Example 16.2.4 (p. 574). This small experiment involved four factorial points and two center points. An R program and output for the analysis of this standard first-order design is shown in Tables 16.17 and 16.18. In Table 16.17,

Fig. 16.9 SAS output
from PROC RSREG:
canonical analysis

the data are read from file, coded using the `coded.data` function of the `rsm` package, saved as
`copper1`, and displayed.

In the R program continuation in Table 16.18, the first-order analysis is generated using the `rsm`
function. In the statement

```
model1 = rsm(s ~ FO(zA, zB), data = copper1)
```

the syntax `FO(zA, zB)` fits the first order model in both coded response variables, using the coded
data `copper1`, saving the results as `model1`. Then the `summary(model1)` command generates
pertinent information, including: parameter estimates, standard errors, and corresponding t-tests; the
analysis of variance table, including a lack-of-fit test; and the direction of steepest ascent with respect
to coded and uncoded variables. Finally, the command

```
steepest(model1, dist=seq(0, 5, by = 1), descent = F)
```

provides predicted response at steps along the path of steepest ascent, stepping from the origin (which
is the design center point for coded data) at distances from zero to five in unit increments, showing the
location of each step in terms of the coded and uncoded predictor variables.

Based on the t tests, neither main effect is significantly different from zero, indicating either that
the experimental region is in the vicinity of the peak, or that neither factor affects the response. The
lack-of-fit test yields a p-value of 0.034, indicating significant lack-of-fit of the first order model,
though the test does not distinguish whether this is due to interaction or quadratic effects. The reader
may verify that the significant lack-of-fit is due to a quadratic effect. This fact caused the experimenters
to add axial points to the first-order design to obtain a central composite design (see Example 16.3.1).
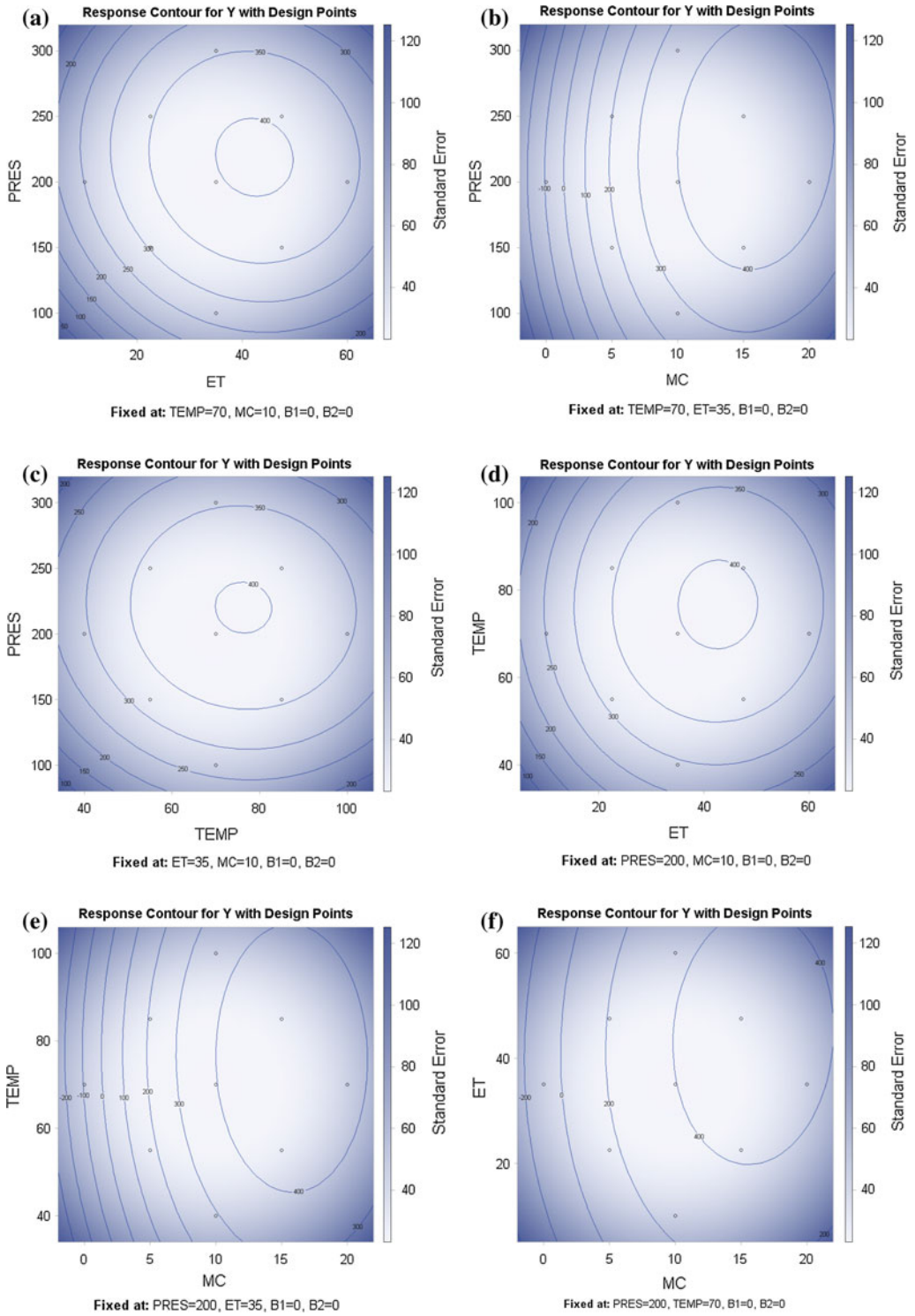
Fig. 16.10 Response surface contour plots for the PAH recovery experiment

Table 16.17 R program and output for first-order response surface regression: data entry and coding

```
> # Read first 6 observations from file
> copper.data = head(read.table("data/copper.txt", header = T), 6)
> # Code data
> # install.packages("rsm")
> library(rsm)
> copper1 = coded.data(copper.data, zA ~ xA - 10.5, zB ~ (xB - 36)/5)
> copper1

    xA xB    s
1  9.5 31 5.60
2  9.5 41 6.45
3 11.5 31 4.84
4 11.5 41 5.19
5 10.5 36 4.32
6 10.5 36 4.25

Data are stored in coded form using these coding formulas ...
zA ~ xA - 10.5
zB ~ (xB - 36)/5
```

16.8.2 Analysis of a Second-Order Design

In the previous section, the syntax FO was used with the rsm function to fit a first order response surface regression model. Analogously, the syntax SO is used to fit a second order model. This is illustrated in the R program beginning in Table 16.19, in the context of the PAH recovery experiment that was introduced in Example 16.4.3 (p. 589). A rotatable central composite design with orthogonal blocking was used to study the effects of four factors—pressure (Pres), temperature (Temp), extraction time (ET), and methanol content (MC)—on the total recovery of polycyclic aromatic hydrocarbons (y) when extracted from soil.

The R program beginning in Table 16.19 reads the data from a file pah.txt which contains the run number, the block level, the uncoded levels of the four factors, and the response variable. A factor variable for blocks is created and all the information is saved in the data set pah.data. The function coded.data of the response surface methods package rsm then codes the levels of each of the factors, saving the coded data as pah.ccd.data. The data are then displayed (see Table 16.19), including the coding formulas. The coding formulas used here code the extreme levels of each factor as ± 1, so the results presented here are consistent with those in the prior SAS software section, though one could certainly choose instead to code the levels of the factorial points as ± 1.

The R program is continued in Table 16.20, where the output shown is all generated by the following two program lines.

```
model2 = rsm(y ~ fBlock + SO(zP, zT, zET, zMC), data = pah.ccd.data)
summary(model2)
```

The first line fits the response surface regression model, saving the results as model2. Block effects are included in the model additively, whereas the syntax SO(zP, zT, zET, zMC) causes inclusion of all terms up to second order in the four factors, including linear, interaction, and quadratic effects. The summary command then generates the output shown in Table 16.20, as well as the canonical analysis shown in Table 16.21.

Table 16.18 R program and output for first-order response surface regression: parameter estimates and analysis of variance

```
> # First-order model analysis
> model1 = rsm(s ~ FO(zA, zB), data = copper1)
> summary(model1)

Call:
rsm(formula = s ~ FO(zA, zB), data = copper1)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.108      0.341   14.96  0.00065
zA            -0.505      0.418   -1.21  0.31373
zB             0.300      0.418    0.72  0.52495

Multiple R-squared:  0.397, Adjusted R-squared:  -0.00545
F-statistic: 0.986 on 2 and 3 DF,  p-value: 0.469

Analysis of Variance Table

Response: s
            Df Sum Sq Mean Sq F value Pr(>F)
FO(zA, zB)   2  1.380   0.690    0.99  0.469
Residuals    3  2.099   0.700
Lack of fit  2  2.096   1.048  427.78  0.034
Pure error   1  0.002   0.002

Direction of steepest ascent (at radius 1):
      zA        zB
-0.85974  0.51074

Corresponding increment in original units:
      xA        xB
-0.85974  2.55368

> steepest(model1, dist = seq(0, 5, by = 1), descent = F)

Path of steepest ascent from ridge analysis:
  dist     zA     zB |      xA      xB |  yhat
1    0  0.000  0.000 | 10.500 36.000 | 5.108
2    1 -0.860  0.511 |  9.640 38.555 | 5.696
3    2 -1.720  1.021 |  8.780 41.105 | 6.283
4    3 -2.579  1.532 |  7.921 43.660 | 6.870
5    4 -3.439  2.043 |  7.061 46.215 | 7.458
6    5 -4.299  2.554 |  6.201 48.770 | 8.046
```

In Table 16.20, the analysis of variance table includes Type I sums of squares for blocks, linear or first order terms, two way interaction terms, and pure quadratic terms, adding the terms to the model in that order. These Type I sums of squares would be the same modeling either coded or uncoded factor levels. A lack-of-fit test is also provided. Observe that the cross product terms are not significantly different from zero, and there is no significant lack of fit of the model. The linear and quadratic components are clearly significant. Looking at the parameter estimates and corresponding t-tests in Table 16.20, clearly the linear and quadratic methanol content terms are needed in the model, providing some clarification to the analysis of variance results.

Table 16.19 R program and output for second-order response surface regression: data entry and coding

```
> pah.data = read.table("data/pah.txt", header = T)
> pah.data$fBlock = factor(pah.data$Block)
> library(rsm)
> pah.ccd.data = coded.data(pah.data, zP ~ (Pres - 200)/100,
+   zT ~ (Temp - 70)/30, zET ~ (ET - 35)/25, zMC ~ (MC - 10)/10)
> pah.ccd.data
   Run Block Pres Temp   ET MC      y fBlock
1    1     1  250   55 47.5 15 391.8      1
2    2     1  150   85 47.5 15 413.6      1
3    3     1  250   55 22.5  5  68.7      1
4    4     1  250   85 47.5  5 143.0      1
5    5     1  150   85 22.5  5 104.0      1
6    6     1  150   55 22.5 15 309.1      1
7    7     1  200   70 35.0 10 400.6      1
8    8     1  250   85 22.5 15 402.5      1
9    9     1  150   55 47.5  5  77.7      1
10  10     1  200   70 35.0 10 426.5      1
11  11     2  250   85 47.5 15 457.5      2
12  12     2  150   55 22.5  5  56.9      2
13  13     2  250   85 22.5  5  94.1      2
14  14     2  250   55 22.5 15 409.7      2
15  15     2  150   55 47.5 15 410.9      2
16  16     2  150   85 22.5 15 375.8      2
17  17     2  150   85 47.5  5 110.5      2
18  18     2  200   70 35.0 10 387.8      2
19  19     2  250   55 47.5  5 103.0      2
20  20     2  200   70 35.0 10 399.1      2
21  21     3  200   70 35.0 10 416.9      3
22  22     3  200   40 35.0 10 359.8      3
23  23     3  200   70 10.0 10 276.1      3
24  24     3  200   70 60.0 10 462.3      3
25  25     3  100   70 35.0 10 311.5      3
26  26     3  200   70 35.0 10 346.5      3
27  27     3  200   70 35.0  0  46.8      3
28  28     3  200   70 35.0 20 418.7      3
29  29     3  200  100 35.0 10 413.9      3
30  30     3  300   70 35.0 10 429.4      3

Data are stored in coded form using these coding formulas ...
zP ~ (Pres - 200)/100
zT ~ (Temp - 70)/30
zET ~ (ET - 35)/25
zMC ~ (MC - 10)/10
```

In Table 16.21, the canonical analysis is shown, including: the stationary point expressed in terms of both coded and uncoded units; the eigenvalues, or canonical coefficients; and the eigenvectors, giving the direction of each canonical axis. The canonical coefficients and axes are with respect to the coded factor levels.

For this experiment, all eigenvalues (canonical coefficients) are negative, so the stationary point is a maximum. The eigenvectors are each scaled to be of length one. The last eigenvalue, with value -227.865, is the largest in magnitude. For the corresponding eigenvector, the primary component is

Table 16.20 R program and output for second-order response surface regression: parameter estimates and analysis of variance

```
> # Second-order model and analysis
> model2 = rsm(y ~ fBlock + SO(zP, zT, zET, zMC), data = pah.ccd.data)
> summary(model2)

Call:
rsm(formula = y ~ fBlock + SO(zP, zT, zET, zMC), data = pah.ccd.data)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   369.16      28.46   12.97  8.2e-09
fBlock2         6.78      26.35    0.26  0.80094
fBlock3        74.44      26.35    2.83  0.01431
zP             37.30      24.05    1.55  0.14492
zT             31.78      24.05    1.32  0.20911
zET            54.97      24.05    2.29  0.03972
zMC           263.07      24.05   10.94  6.3e-08
zP:zT          -6.35      58.91   -0.11  0.91581
zP:zET        -11.65      58.91   -0.20  0.84630
zP:zMC         23.10      58.91    0.39  0.70133
zT:zET          2.30      58.91    0.04  0.96945
zT:zMC         -4.35      58.91   -0.07  0.94226
zET:zMC        16.55      58.91    0.28  0.78319
zP^2          -88.63      44.99   -1.97  0.07056
zT^2          -72.23      44.99   -1.61  0.13246
zET^2         -89.88      44.99   -2.00  0.06714
zMC^2        -226.33      44.99   -5.03  0.00023

Multiple R-squared:  0.928, Adjusted R-squared:  0.839
F-statistic: 10.5 on 16 and 13 DF,  p-value: 0.0000582

Analysis of Variance Table

Response: y
                         Df Sum Sq Mean Sq F value     Pr(>F)
fBlock                    2  33884   16942    4.88     0.0262
FO(zP, zT, zET, zMC)      4 447761  111940   32.25  0.0000012
TWI(zP, zT, zET, zMC)     6   1008     168    0.05     0.9993
PQ(zP, zT, zET, zMC)      4  99227   24807    7.15     0.0029
Residuals                13  45118    3471
Lack of fit              10  42240    4224    4.40     0.1246
Pure error                3   2877     959
```

that of MC with value 0.994301. So, the fitted model has greatest curvature at the stationary point when moving in either direction determined by this fourth eigenvector, which is nearly parallel to the MC-axis, as is evident from the SAS contour plots in Fig. 16.10 (p. 601), where MC is the x-axis variable of plots (b), (e) and (f). To generate similar plots using R, add the following code to the end of the R program in Tables 16.19 and 16.20.

```
par(mfrow = c(3, 2))
contour(model2, ~ zP + zT + zET + zMC,
              at = round(xs(model2), 3), las = 1)
```

Table 16.21 R program output for second-order response surface regression: canonical analysis

```
Stationary point of response surface:
     zP       zT      zET      zMC
0.25947 0.19593 0.34721 0.60522

Stationary point in original units:
   Pres    Temp       ET      MC
225.947  75.878   43.680  16.052

Eigenanalysis:
$values
[1] -71.257  -84.168  -93.761 -227.865

$vectors
          [,1]       [,2]       [,3]       [,4]
zP    0.233360   0.712899  -0.655835  -0.084838
zT   -0.964432   0.255316  -0.067269   0.012632
zET  -0.121729  -0.652941  -0.744877  -0.063313
zMC   0.024413   0.016008  -0.102535   0.994301
```

The `contour` statement above generates six black-and-white contour plots—one for each pair of the four factors listed. The `par` statement causes the six contour plots to be arranged in a 3×2 layout; otherwise, one would obtain six separate plots. The option `at = xs(model2)` would use the stationary point rather than the center point to fix the levels of the unplotted factors, whereas the option `at = round(xs(model2),3)` does the same but rounds each component to 3 digits, providing cleaner subheadings. For color plots, include the option `image = T`.

16.8.3 Generating Designs

In this section, we illustrate some capabilities of the R software to generate response surface designs, using functions of the `rsm` package. Table 16.22 contains sample R code generating most of the designs used in examples in this chapter.

The function `cube` generates first order designs consisting of factorial (or cube) points and center points. For example, the statement

```
cube(basis = ~ A+B+C, generators = c(D ~ B*C, E ~ A*D, F ~ A*B),
     reps = 4, n0 = 0)
```

generates the design for the paint experiment introduced in Example 16.2.1. The syntax `basis = ~ A+B+C` causes inclusion of all eight combinations of *A*, *B* and *C* at two levels each, then the `generators` option defines the levels of *D*, *E* and *F* in terms of prior variables using the generators *BCD*, *ADE* and *ABF*. Each of the resulting eight treatment combinations is replicated four times, since `reps = 4`, and the design includes `n0 = 0` center points.

Central composite designs can be generated either in one or two steps, as illustrated in the code in Table 16.22 for the acid copper plating experiment of Example 16.2.4. The `ccd` function generates a complete central composite design. Alternatively, the `cube` function generates the first order design, the `star` function generates the axial or star points and any additional center points, and the `djoin` function joins these factorial and axial parts together. The default is for the design to have two blocks—one for each of the factorial and axial parts—unless `oneblock = T` is specified. The design generated

Table 16.22 Sample R code for generating response surface designs used in the chapter

```
library(rsm)

# Paint experiment
cube(basis = ~ A+B+C, generators = c(D ~ B*C, E ~ A*D, F ~ A*B),
     n0 = 0, reps = 4)

# Paint experiment: same design, different notation for factors
cube(3, generators = c(x4 ~ x2*x3, x5 ~ x1*x4, x6 ~ x1*x2),
     n0 = 0, reps = 4)

# Acid copper pattern plating experiment: rotatable CCD
# Create design in 2 parts, then join the parts
dsgn1 = cube(2, n0 = 2, reps = 1,
             coding = list(x1 ~ (A - 10.5)/1, x2 ~ (B - 36)/5))
dsgn2 = star(dsgn1, n0 = 0, alpha = "rotatable", reps = 1)
dsgn12 = djoin(dsgn1, dsgn2)
dsgn12 # Show design in randomized order
stdorder(dsgn12) # Show design in standard order
# Create the design all at once: defaults is factorial and axial blocks
dsgn = ccd(2, n0 = c(2, 0), alpha = "rotatable", oneblock = T,
  randomize = F, coding = list(x1 ~ (A - 10.5)/1, x2 ~ (B - 36)/5))
varfcn(dsgn, ~ SO(x1, x2), contour = T) # Contour of scaled variances
# Add data to the coded data set containing the design
dsgn$s = c(5.60, 4.84, 6.45, 5.19, 4.32, 4.25, 5.76, 4.42, 5.46, 5.81)
# Data analysis
model2 = rsm(s ~ SO(x1, x2), data = dsgn)
summary(model2)
contour(model2, ~ x1 + x2, at = round(xs(model2), 3), image = T,
                las = 1)

# Flour experiment: orthogonal CCD
ccd(basis = ~ B+C+D, n0 = 2, alpha = 1.2872, randomize = F)

# PAH experiment: rotatable CCD, with orthogonal blocking
ccd(4, alpha = "rotatable", n0 = 2, blocks = ~ x1*x2*x3*x4,
        randomize = F)

# Flour experiment: noise array only
cube(basis = ~ G+J, generators = c(K ~ G*J), n0 = 0, randomize = F)

# Box-Behnken example design: 4 factors, 1 ctr pt per block
bbd(4, n0 = 3, block = F, randomize = F)
# Same design, except 3 blocks
bbd(4, n0 = 1, block = T, randomize = F)
```

is in terms of coded variables, though the `coding` option allows the user to specify the coding formula relating each variable to its coded levels. By default, the design is randomized separately within each block, in which case the function `stdorder` can be used to display the design in standard, or unrandomized, order. One can specify a specific value of α (e.g. `alpha = 1.2872` for the flour experiment), or request that the value be set corresponding to a desired design property, such as: `alpha = "rotatable"` for a rotatable design; `alpha = "orthogonal"` for orthogonal blocking (not orthogonality); `alpha = "spherical"` for the axial and factorial points to be the same distance from the center points, and `alpha = "faces"` for the axial points to be on the faces of the cube (same as `alpha = 1`). Given a design, the variance function `varfcn` can generate a contour plot of the scaled variance for a given design and model, so one can see rotatability or non-rotatability, for example.

Using the `bbd` function, one can generate Box–Behnken designs for 3–7 factors, including designs with orthogonal blocking for either four factors and three blocks or five factors and two blocks.

For sake of completeness, the code for the acid copper pattern plating experiment illustrates one way to add data to the coded design data set, as needed for data analysis.

Exercises

1. **Paint experiment, continued**

   The paint experiment of Eibl et al. (1992) was discussed in Example 16.2.1 (p. 569), where the first-order model was fitted to the data. For the fitted first-order model, do the following.

   (a) Plot the residuals versus run order, and use the plot to check the independence assumption. (The order of the observations was not randomized in this experiment. Rather, the observations were collected in the order they are shown row by row in Table 16.1, p. 570.)
   (b) Plot the residuals versus the predicted values, and use the plot to check the assumption of equal variance.
   (c) Plot the residuals versus their normal scores, and use the plot to check the normality assumption.
   (d) Verify that the design is orthogonal.

2. **Paint followup experiment**

   The data of the second paint experiment described by Eibl et al. (1992) are given in Table 16.23. This experiment involves factors A–D, as these had significant effects in the first experiment (Example 16.2.1). The factors are

   $$A: \text{belt speed} \qquad B: \text{tube width}$$
   $$C: \text{pump pressure} \quad D: \text{paint viscosity}$$

   All four factors are at lower levels than in the first experiment. Lowering the levels of factors B–D was indicated by the analysis of the first experiment. Lowering the level of factor A was based on a conjecture of the experimenters.

   (a) The experiment consists of two replicates of a half-fraction. Find the defining relation for the half-fraction.
   (b) Fit the first-order model, recoding the factor levels as ± 1.
   (c) Test for lack of fit of the first-order model.

Table 16.23 Paint thickness y_{zt} for the paint followup experiment

z_A	z_B	z_C	z_D	y_{z1}	y_{z2}
−1.5	0	−2	0	1.71	1.61
0.5	0	−2	0	0.91	1.30
−1.5	−2	0	0	1.71	1.60
0.5	−2	0	0	1.15	1.29
−1.5	0	0	−2	1.33	1.06
0.5	0	0	−2	1.74	1.98
−1.5	−2	−2	−2	0.64	0.78
0.5	−2	−2	−2	1.51	1.18

Source Eibl et al. (1992). Reprinted with Permission from Journal of Quality Technology © 1992 ASQ, www.asq.org

(d) What would you recommend the experimenters do next?

3. **Fractionation experiment**

Sosada (1993) studied the effects of extraction time, solvent volume, ethanol concentration, and temperature on the yield and phosphatidylcholine enrichment (PCE) of deoiled rapeseed lecithin when fractionated with ethanol. Initially, a single-replicate 2^4 experiment was conducted, augmented by three center points.

(a) The results for the 16 factorial points are shown as the first 16 runs in Table 16.24. Fit the first-order model for the response variable "PCE" and conduct the corresponding analysis of variance.
(b) The design also included $n_0 = 3$ center-point observations of PCE. The sample variance of these three observations was $s_0^2 = 1.120$. Test the first-order model for lack of fit, using a 5% level of significance. (Hint: Since the factorial points include no replication, $msPE = s_0^2$, and ssE based on all 19 runs is equal to ssE from the factorial portion of the design plus $(n_0 - 1)s_0^2$.)
(c) Based on the results of parts (a) and (b), what subsequent experimentation would you recommend?

4. **Fractionation experiment, continued**

The fractionation experiment was described in Exercise 3, where the response PCE was used. Consider here, instead, the analysis of "Yield".

(a) Fit the first-order model for the response variable "Yield" based on the initial first-order 2^4 factorial design, shown as the first 16 runs in Table 16.24. Conduct the corresponding analysis of variance.
(b) At the design center point, three additional observations were collected, for which the sample variance was $s_0^2 = 0.090$. Test the first-order model for lack of fit, using a 5% level of significance. (Hint: Since the factorial points include no replication, $msPE = s_0^2$, and ssE based on all 19 runs is equal to ssE from the factorial portion of the design plus $(n_0 - 1)s_0^2$.)
(c) Based on the results of parts (a) and (b), what subsequent experimentation would you recommend?

Table 16.24 Purified lecithin yield and phosphatidylcholine enrichment (PCE), given extraction time (z_1), solvent volume (z_2), ethanol concentration (z_3), and temperature (z_4); fractionation experiment

Run	z_1	z_2	z_3	z_4	Yield	PCE
1	1	1	1	1	27.6	43.8
2	−1	−1	1	1	16.6	27.2
3	1	−1	−1	1	15.4	23.6
4	−1	1	−1	1	17.4	26.2
5	1	−1	1	−1	17.0	27.8
6	−1	1	1	−1	19.0	30.2
7	1	1	−1	−1	17.4	25.2
8	−1	−1	−1	−1	12.6	18.8
9	1	−1	1	1	18.6	28.8
10	−1	1	1	1	22.4	36.8
11	1	1	−1	1	21.4	33.4
12	−1	−1	−1	1	14.0	21.0
13	1	1	1	−1	24.0	38.0
14	−1	−1	1	−1	15.6	23.6
15	1	−1	−1	−1	13.0	20.2
16	−1	1	−1	−1	14.4	22.6
17	0	0	0	0	22.6	
18	$\sqrt{2}$	0	0	0	23.4	
19	$-\sqrt{2}$	0	0	0	20.6	
20	0	$\sqrt{2}$	0	0	22.6	
21	0	$-\sqrt{2}$	0	0	13.4	
22	0	0	$\sqrt{2}$	0	20.6	
23	0	0	$-\sqrt{2}$	0	15.6	
24	0	0	0	$\sqrt{2}$	21.0	
25	0	0	0	$-\sqrt{2}$	17.6	

Source Sosada (1993). Copyright © 1993 American Oil Chemists Society. Reprinted with permission

5. **Fractionation experiment, continued**

   The fractionation experiment was described in Exercise 3, and analysis of the first-order model for "Yield" was considered in Exercise 4. Based on the analysis of the first-order design, the experimenter chose to augment the 16 factorial points of the first-order design into a 25-run central composite design, the yields from which are shown in Table 16.24.

   (a) Determine whether the central composite design used is rotatable or orthogonal.
   (b) Fit the second-order response surface model and determine which effects are significantly different from zero.
   (c) Conduct a canonical analysis and discuss the results with respect to the following items. What is the nature of the critical point? Noting that the objective is to increase yield, in what direction should one move in subsequent experimentation?

6. **Film viscosity experiment**

   Cuq et al. (1995, *Journal of Food Science*) used a central composite design to study the effects of protein concentration (g/100 g solution), pH, and temperature (°C), denoted by P, H, and T,

respectively, on the apparent viscosity Y (mPa) of film-forming solution, in the development of edible packaging films based on fish myofibrillar proteins. The data are shown in Table 16.25.

(a) Is this central composite design rotatable or orthogonal?
(b) Fit the second-order model to the data using the coded factor levels, and check the model assumptions. Would you recommend that a transformation of the data be taken?
(c) Fit the second-order model to the natural log of the data, $\ln(y)$, using the coded factor levels.
(d) Conduct the test for lack of fit of the second-order model for $\ln(y)$.
(e) Check the model assumptions for $\ln(y)$.
(f) Conduct the canonical analysis for $\ln(y)$.
(g) Conduct the analysis of variance for $\ln(y)$.
(h) Compute the coefficient of multiple determination R^2 for the second-order model for $\ln(y)$.
(i) Assess the results of the experiment, based on the model for $\ln(y)$.

7. **Flour production experiment, continued**

Consider again the flour production experiment of Sect. 16.5. The data were given in Table 16.11 (p. 590), along with the statistics $\bar{y}_{.\mathbf{z}}$ and $100 \log_{10}(s_{\mathbf{z}})$ computed for the observations at each design-factor combination $\mathbf{z}$.

(a) Plot $\log_{10}(s_{\mathbf{z}})$ versus $\log_{10}(\bar{y}_{.\mathbf{z}})$, and use the methods of Sect. 5.6.2 to determine an appropriate variance-stabilizing transformation for these data. (Use of $\log_{10}$ is equivalent to use of $\ln$ for choosing a transformation.)

Table 16.25 Apparent viscosity y_{zt} of film-forming solution, for combinations of levels of protein concentration (g/100 g solution), pH, and temperature (°C)

Design point	P		H		T		
	z_P	x_P	z_H	x_H	z_T	x_T	y
1	−1	1.25	−1	2.75	−1	20	50
2	1	2.75	−1	2.75	−1	20	48
3	−1	1.25	1	3.25	−1	20	16700
4	1	2.75	1	3.25	−1	20	560
5	−1	1.25	−1	2.75	1	40	320
6	1	2.75	−1	2.75	1	40	18
7	−1	1.25	1	3.25	1	40	19000
8	1	2.75	1	3.25	1	40	5000
9	−2	0.50	0	3.00	0	30	12700
10	2	3.50	0	3.00	0	30	182
11	0	2.00	−2	2.50	0	30	14
12	0	2.00	2	3.50	0	30	27800
13	0	2.00	0	3.00	−2	10	133
14	0	2.00	0	3.00	2	50	4300
15	0	2.00	0	3.00	0	30	57
16	0	2.00	0	3.00	0	30	70
17	0	2.00	0	3.00	0	30	58
18	0	2.00	0	3.00	0	30	56

Source Cuq et al. (1995). Copyright © 1995 Inst. of Food Technologists. Reprinted with permission

Table 16.26 Resin impurity content y_{zt} (ppm)

Design point	Time	Temp.	$y_{x,t}$
1	7.0	232.4	18.5
2	3.0	220.0	22.5
3	11.0	220.0	17.2
4	1.3	190.0	42.2
5	7.0	190.0	28.6
6	7.0	190.0	19.8
7	7.0	190.0	23.6
8	7.0	190.0	24.1
9	7.0	190.0	24.2
10	12.7	190.0	19.1
11	3.0	160.0	54.1
12	11.0	160.0	33.8
13	7.0	147.6	55.4

(b) Repeat the first analysis of variance of Sect. 16.5, for which the response variable was $\overline{y}_{.z}$, after applying the transformation determined in part (a) to the observations y_{hz}. Compare your conclusions with those reached in Sect. 16.5.

(c) Repeat the second analysis of variance of Sect. 16.5, for which the response variable was $100 \log_{10}(s_z)$, after applying the transformation determined in part (a) to the observations y_{hz}. Compare your conclusions to those reached in Sect. 16.5.

8. **Central composite design**

Consider using a central composite design for three factors, to include eight factorial points and six axial points.

(a) Determine the value of α to make the design rotatable.

(b) Investigate how α and the number of center points should be chosen to make the design both rotatable and orthogonal, if possible. If this is not possible, how can the design be made rotatable and nearly orthogonal?

(c) Investigate whether the design can be rotatable with orthogonal blocking. If not, then investigate whether orthogonal blocking is possible. If so, how many blocks could be used? Investigate whether orthogonal blocking and near rotatability is possible.

9. **Central composite design**

Repeat Exercise 8 for a central composite design for four factors, to include 16 factorial points and eight axial points.

10. **Resin impurity experiment**

An experiment was conducted using a design close to a central composite design to study the effects of drying time (hours) and temperature (°C) on the content y (ppm) of undesirable compounds in a resin. The data are shown in Table 16.26.

Table 16.27 Resin degradation (ppm) for the resin moisture experiment

Design point	T		H		P		
	z_T	x_T	z_H	x_H	z_P	x_P	y
1	-1	150	-1	0	0	4	83
2	-1	150	0	50	-1	1	103
3	-1	150	0	50	1	9	94
4	-1	150	1	100	0	4	98
5	0	185	-1	0	-1	1	51
6	0	185	-1	0	1	9	48
7	0	185	1	100	-1	1	106
8	0	185	1	100	1	9	108
9	1	220	-1	0	0	4	36
10	1	220	0	50	-1	1	153
11	1	220	0	50	1	9	107
12	1	220	1	100	0	4	87
13	0	185	0	50	0	4	80
14	0	185	0	50	0	4	81
15	0	185	0	50	0	4	77
16	0	185	0	50	0	4	80
17	0	185	0	50	0	4	82

(a) Determine the coded levels of time and temperature, as well as the values of n_f, n_a, n_0. What values of α for each factor were selected by the experimenters for the axial points? Why is the design not quite a central composite design?

(b) Fit the second-order model, using coded factor levels.

(c) Test for model lack of fit.

(d) Check the equal variance and normality assumptions of the model using residual plots.

(e) Conduct the canonical analysis.

(f) Conduct the analysis of variance.

(g) Summarize the results.

11. **Resin moisture experiment**

A Box–Behnken design was used to determine whether specific drying conditions for a process could yield a resin that is sufficiently devoid of moisture and low-molecular-weight components. The three factors T, H, and P under study were temperature (150, 185, 220°C), relative humidity (0, 50, 100%), and air pressure (1, 5, 9 torr). The response variable y was a measure of product degradation (ppm). The design and data are shown in Table 16.27.

(a) Fit the second-order model, using coded factor levels.

(b) Test for model lack of fit.

(c) Check the equal-variance and normality model assumptions using residual plots.

(d) Conduct the canonical analysis.

(e) Conduct the analysis of variance.

(f) Summarize the results.

12. **Box–Behnken design**

   (a) Construct a Box–Behnken design for three factors based on the balanced incomplete block design for three treatments in three blocks of size two and the 2^2 factorial design.
   (b) Determine whether the design constructed in part (a) is rotatable.
   (c) For the design constructed in part (a), determine whether orthogonal blocking is possible.

13. **Box–Behnken design**

   (a) Construct a Box–Behnken design for five factors based on the balanced incomplete block design for five treatments in 10 blocks of size two.
   (b) Determine whether the design constructed in part (a) is rotatable.
   (c) For the design constructed in part (a), determine whether orthogonal blocking is possible.

Random Effects and Variance Components

17

17.1 Introduction

Until now, we have looked only at treatment factors whose levels have been specifically chosen. We have tested hypotheses about, and calculated confidence intervals for, comparisons in the effects of these particular treatment factor levels. These treatment effects are known as *fixed effects*, since we represent them in the model as unknown constants (parameters). Models that contain only fixed effects are called *fixed-effects models*.

As mentioned in step (f) of the checklist in Sect. 2.2, p. 7, there are occasions when we are interested in a large population of possible levels of a treatment factor, and the levels that are actually used in the experiment are a random sample from this population. The effects of the levels used in the experiment are then represented as random variables whose distributions are the distributions of values in the population. Such treatment-factor effects are called *random effects*, and the corresponding models are called *random-effects models*. We are not interested in just the levels that happen to be in the experiment. Rather, we are concerned with the variability of the effects of all the levels in the population. Consequently, random effects are handled somewhat differently from fixed effects. Some examples of experiments involving random effects are given in Sect. 17.2.

In Sect. 17.3 we look at experiments with a single random effect. The selection of sample sizes and model assumption checking are discussed in Sects. 17.4 and 17.5. These ideas are extended to experiments with two or more random effects in Sect. 17.6.

An experiment may involve both random and fixed effects, and the corresponding model is then known as a *mixed model*. Such experiments are discussed in Sect. 17.7. Block effects may also be random effects, and these are discussed in Sect. 17.9. Rules for obtaining confidence intervals and testing hypotheses for random effects are given in Sect. 17.8. The use of SAS and R software is considered in Sects. 17.10 and 17.11, respectively.

17.2 Some Examples

Before running an experiment, the checklist (Sect. 2.2, p. 7) should be completed as usual. In the case of a random effect, the treatment factor will have an extremely large number of levels, only a very small proportion of which can be observed in the experiment. Throughout this chapter, we will assume that the total possible number of levels of each treatment factor will be at least 100 times larger than the numbers of levels that can be observed. Typically, the population of possible levels *will* meet this

requirement, and for the purposes of writing down a model, we may regard the population as infinite. Otherwise, one needs to make a correction for a "finite population" in all of the formulae, and this is beyond the scope of this book. Some examples of "infinite" populations are given in Example 17.2.1.

Example 17.2.1 `Infinite populations`

Suppose that a manufacturer of canned tomato soup wishes to reduce the variability in the thickness of the soup. Suppose that the most likely causes of the variability are the quality of the cornflour (cornstarch) received from the supplier and the actions of the machine operators. Let us consider two different scenarios:

Scenario 1: The machine operators are highly skilled and have been with the company for a long time. Thus, the most likely cause of variability is the quality of the cornflour delivered to the company. The treatment factor is "cornflour," and its possible levels are all the possible batches of cornflour that the supplier could deliver. Theoretically, this is an infinite population of batches. We are interested not only in the batches of cornflour that have currently been delivered, but also in all those that might be delivered in the future. If we assume that the batches delivered to the company are a random sample from all batches that could be delivered, and if we take a random sample of delivered batches to be observed in the experiment, then the effect of the cornflour on the thickness is a random effect and can be modeled by a random variable.

Scenario 2: It is known that the quality of the cornflour is extremely consistent, so the most likely cause of variability is due to the different actions of the machine operators. The company is large and machine operators change quite frequently. Consequently, those that are available to take part in the experiment are only a small sample of all operators employed by the company at present or that might be employed in the future. If we can assume that the operators available for the experiment are representative of the population, then we can assume that they are similar to a random sample from a very large population of possible operators, present and future. Since we would like to know about the variability of the entire population, we model the effect of the operators as random variables, and call them random effects.   □

In the absence of any blocking factors, a completely randomized design would be used. The levels of the random-effects treatment factor are first selected at random from the population of all possible levels, and then the experimental units are randomly assigned to these selected levels as usual. At step (h) of the checklist, we need to calculate the number of levels v of the treatment factor to be observed in the experiment in addition to r, the number of observations on each level. Since this calculation uses the formulas for confidence intervals and hypothesis tests, we will postpone the discussion to Sect. 17.4. As a general rule, if the variability of the treatment effects is much greater than the error (measurement) variability, then v should be large and r small; and vice versa.

Example 17.2.2 `Clean wool experiment`

The clean wool experiment was reported by J.M. Cameron, of the National Bureau of Standards, in the 1951 volume of *Biometrics*. The following checklist has been compiled from the information given in the article.

(a) **Define the objectives of the experiment**.

> Raw wool contains varying amounts of grease, dirt, and foreign material which must be removed before manufacturing begins. The purchase price and customs levy of a shipment are based on the actual amount

of wool present, i.e., on the amount of wool present after thorough cleaning—the "clean content." The clean content is expressed as the percentage the weight of the clean wool is of the original weight of the raw wool.

The experiment was run in order to estimate the variability in "clean content" of bales of wool in a shipment.

(b) **Identify all sources of variation**.

(i) Treatment factors and their levels.
The treatment factor was "wool bale" and its levels were the entire population of bales in a particular shipment. Seven bales were observed in the experiment, and these were selected at random from the shipment. The shipment was large enough to allow the bales used in the experiment to be regarded as a random sample from an infinite population of bales. The treatment factor "wool bale" was therefore regarded as a random effect.

(ii) Experimental units.
The experimental units were time slots, so that allocation of these to the levels of the treatment factor determined the order in which the wool bales were observed.

(iii) Blocking factors, noise factors, and covariates.
No nuisance factors were identified as major sources of variation.

(c) **Choose a rule by which to assign the experimental units to the levels of the treatment factors**.

A completely randomized design was selected.

(d) **Specify the measurements to be made, the experimental procedure, and the anticipated difficulties**.

A machine was used to bore through a bale of wool and extract a core of wool. Several cores were taken from each of the seven selected bales so that several observations on clean content could be made on each bale. Each core of wool was weighed and then cleaned by scouring, removing burrs, etc. After cleaning, the wool was reweighed and the clean content calculated as the ratio of the clean wool to the initial weight, times 100%.
An anticipated difficulty was that the scouring process, which works well with large amounts of wool, proves difficult with a small core of wool, so that the experimental error observed in the experiment may be larger than would normally be observed in routine production.

The observations on the clean content of the seven bales are shown in Table 17.1. Model selection for this experiment and its analysis via SAS and R software are discussed in Sects. 17.10 and 17.11, respectively.   □

Table 17.1   Data for the clean wool experiment

	Bale						
	1	2	3	4	5	6	7
Clean content	52.33	56.99	54.64	54.90	59.89	57.76	60.27
	56.26	58.69	57.48	60.08	57.76	59.68	60.30
	62.86	58.20	59.29	58.72	60.26	59.58	61.09
	50.46	57.35	57.51	55.61	57.53	58.08	61.45

Source Cameron (1951). Copyright © 1951 International Biometric Society. Reprinted with permission

17.3   One Random Effect

17.3.1   The Random-Effects One-Way Model

For a completely randomized design, with v randomly selected levels of a treatment factor T, the random-effects one-way model is

$$Y_{it} = \mu + T_i + \epsilon_{it}\,, \tag{17.3.1}$$

$$\epsilon_{it} \sim N(0, \sigma^2)\,, \quad T_i \sim N(0, \sigma_T^2)\,,$$

$$\epsilon_{it}\text{'s and } T_i\text{'s are all mutually independent}\,,$$

$$t = 1, \ldots, r_i, \quad i = 1, \ldots, v\,.$$

Compare this with the fixed-effects one-way analysis of variance model (3.3.1), p. 33. The form of the model and the error assumptions are exactly the same. The only difference is in the modeling of the treatment effect. Since the ith level of the treatment factor T observed in the experiment has been randomly selected from the "infinite" population, its observed effect is an observation of a random variable T_i. The distribution of T_i is the distribution of treatment effects in the whole population. We have assumed in (17.3.1) that the population of effects follows a normal distribution with variance σ_T^2, and this assumption will need to be checked along with the error assumptions. The mean of the treatment-effect population has been absorbed into the constant μ, so the distribution of T_i is listed as $N(0, \sigma_T^2)$. The variance σ_T^2 is the parameter of interest, since if the effects of all of the treatment-factor levels are the same, then σ_T^2 is zero. If the effects of the levels are quite different, then σ_T^2 is quite large.

Our final assumption is one of independence. If the treatment-factor levels are selected at random, then the assumption of independence of $T_1, T_2, \ldots, T_v$ is reasonable. However, if, as in Example 17.2.1 Scenario 2, the levels are a "convenient sample," then this assumption should be investigated carefully. Independence of the T_i and ϵ_{it} requires that the treatment factor not affect any source of variation that has been absorbed into the error variable.

In a random-effects model, the expected value of the response is μ, since

$$E[Y_{it}] = E[\mu] + E[T_i] + E[\epsilon_{it}] = \mu\,.$$

The variance of Y_{it} is

$$\mathrm{Var}(Y_{it}) = \mathrm{Var}(\mu + T_i + \epsilon_{it}) = \mathrm{Var}(T_i) + \mathrm{Var}(\epsilon_{it}) + 2\mathrm{Cov}(T_i, \epsilon_{it}) = \sigma_T^2 + \sigma^2\,,$$

since T_i and ϵ_{it} are mutually independent and so have zero covariance. Therefore, the distribution of Y_{it} is

$$Y_{it} \sim N(\mu,\ \sigma_T^2 + \sigma^2). \tag{17.3.2}$$

The two components σ_T^2 and σ^2 of the variance of Y_{it} are known as *variance components*. Observations on the same treatment are correlated, with

$$\mathrm{Cov}(Y_{it}, Y_{is}) = \mathrm{Cov}(\mu + T_i + \epsilon_{it}, \mu + T_i + \epsilon_{is}) = \mathrm{Var}(T_i) = \sigma_T^2\,.$$

17.3.2 Estimation of σ^2

In order to be able to test hypotheses about σ_T^2 or to calculate confidence intervals, we need an unbiased estimate of σ^2. The random-effects one-way model (17.3.1) is very similar to the fixed effects one-way analysis of variance model (3.3.1), p. 33, so a natural question is whether the fixed-effects mean square for error MSE provides an unbiased estimator for σ^2 in the random-effects model also. The answer, happily, is "yes," and we can check it by calculating $E[MSE]$ for the random-effects model, as shown below.

From (3.4.6), p. 39, the fixed-effects sum of squares for error is

$$SSE = \sum_{i=1}^{v} \sum_{t=1}^{r_i} Y_{it}^2 - \sum_{i=1}^{v} r_i \overline{Y}_{i.}^2 .$$

Remember that the variance of a random variable X is calculated as $\mathrm{Var}(X) = E[X^2] - (E[X])^2$. So, we have

$$E[Y_{it}^2] = \mathrm{Var}(Y_{it}) + (E[Y_{it}])^2 = (\sigma_T^2 + \sigma^2) + \mu^2 .$$

Now,

$$\overline{Y}_{i.} = \mu + T_i + \frac{1}{r_i} \sum_{t=1}^{r_i} \epsilon_{it} ,$$

so

$$\mathrm{Var}(\overline{Y}_{i.}) = \sigma_T^2 + \frac{\sigma^2}{r_i} \quad \text{and} \quad E[\overline{Y}_{i.}] = \mu . \tag{17.3.3}$$

Consequently,

$$E[\overline{Y}_{i.}^2] = \left(\sigma_T^2 + \frac{\sigma^2}{r_i}\right) + \mu^2.$$

Thus,

$$E[SSE] = \sum_{i=1}^{v} \sum_{t=1}^{r_i} (\sigma_T^2 + \sigma^2 + \mu^2) - \sum_{i=1}^{v} r_i \left(\sigma_T^2 + \frac{\sigma^2}{r_i} + \mu^2\right)$$

$$= n\sigma^2 - v\sigma^2 \quad \left(\text{where } n = \sum_{i=1}^{v} r_i\right)$$

$$= (n - v)\sigma^2 ,$$

giving

$$E[MSE] = E[SSE/(n - v)] = \sigma^2.$$

So MSE is an unbiased estimator for σ^2, and the observed value of the mean square for error, msE, is an unbiased estimate for σ^2 in the random-effects one-way model, as well as in the fixed-effects one-way model.

Confidence bounds for σ^2 can be computed as under fixed-effects models (Sect. 3.4.6), that is,

$$\sigma^2 \le \frac{ssE}{\chi^2_{n-v,1-\alpha}} \,, \tag{17.3.4}$$

where $\chi^2_{n-v,1-\alpha}$ is the percentile of the chi-squared distribution with $n - v$ degrees of freedom and with probability of $1 - \alpha$ in the right-hand tail.

17.3.3   Estimation of σ^2_T

Since the fixed-effects mean square for error msE provides an unbiased estimate of σ^2, the next question that is natural to ask is whether the fixed-effects mean square for treatments msT provides an unbiased estimate for σ^2_T. The answer is "not quite," but we can certainly use it to find an estimate. Now $msT = ssT/(v - 1)$, and ssT is given in (3.5.11), p. 43, as

$$ssT = \sum_{i=1}^{v} r_i \bar{y}_{i.}^2 - n\bar{y}_{..}^2$$

with corresponding random variable

$$SST = \sum_{i=1}^{v} r_i \overline{Y}_{i.}^2 - n\overline{Y}_{..}^2 \,.$$

Using the same type of calculation as in Sect. 17.3.2 above, we have

$$\overline{Y}_{..} = \mu \,+\, \frac{1}{n}\sum_i r_i T_i \,+\, \frac{1}{n}\sum_{i=1}^{v}\sum_{t=1}^{r_i} \epsilon_{it} \,.$$

So

$$E[\overline{Y}_{..}] = \mu \quad\text{and}\quad \mathrm{Var}(\overline{Y}_{..}) = \frac{\sum r_i^2}{n^2}\sigma^2_T + \frac{n}{n^2}\sigma^2 \,.$$

Also, from (17.3.3),

$$E[\overline{Y}_{i.}] = \mu \quad\text{and}\quad \mathrm{Var}(\overline{Y}_{i.}) = \sigma^2_T + \frac{\sigma^2}{r_i}.$$

Therefore,

$$E[SST] = \sum_{i=1}^{v} r_i \left(\sigma^2_T + \frac{\sigma^2}{r_i} + \mu^2 \right) - n\left(\frac{\sum r_i^2}{n^2}\sigma^2_T + \frac{\sigma^2}{n} + \mu^2 \right)$$

$$= \left(n - \frac{\sum r_i^2}{n} \right)\sigma^2_T + (v - 1)\sigma^2 \,.$$

Since $MST = SST/(v-1)$, we have

$$E[MST] = c\sigma_T^2 + \sigma^2, \ \text{where} \ c = \frac{n^2 - \sum r_i^2}{n(v-1)}.$$

Notice that if all r_i are equal to r, then $n = vr$ and $c = r$.

We see that MST is an unbiased estimator of $c\sigma_T^2 + \sigma^2$, not σ_T^2. Nevertheless, we can easily find an unbiased estimator of σ_T^2, since

$$E\left[\frac{MST - MSE}{c}\right] = \sigma_T^2. \tag{17.3.5}$$

It is, unfortunately, possible for the observed value of this estimator to be negative even though σ_T^2 cannot be negative. This will occur when msE happens to be greater than msT, and this is most likely when σ_T^2 is close to zero. If msE is considerably greater than msT, then the model should be questioned, as it is unlikely to be a good description of the data.

Example 17.3.1 `Ice cream experiment`

The following experiment was run by Sue Hubbard in 1986 to determine whether or not different flavors of ice cream melt at different speeds. A random sample of three flavors was selected from a large population of flavors offered to the customer by a single manufacturer in May 1986. It is not obvious that the selected flavors are representative of all possible ice cream flavors, since some may include an ingredient that inhibits melting. The theoretical population is therefore the population of all flavors that could be made with ingredients similar to those flavors available.

The three flavors of ice cream were stored in the same freezer in similar-sized containers. For each observation, one teaspoonful of ice cream was taken from the freezer, transferred to a plate, and the melting time at room temperature was observed to the nearest second. Eleven observations were taken on each flavor. These are shown, together with their order of observation, in Table 17.2 and plotted in Fig. 17.1.

Now,

$$\begin{aligned} ssE &= \sum\sum y_{it}^2 - 11\sum \bar{y}_{i.}^2 \\ &= 30,206,485 - 30,003,028.8181 \\ &= 203,456.1819. \end{aligned}$$
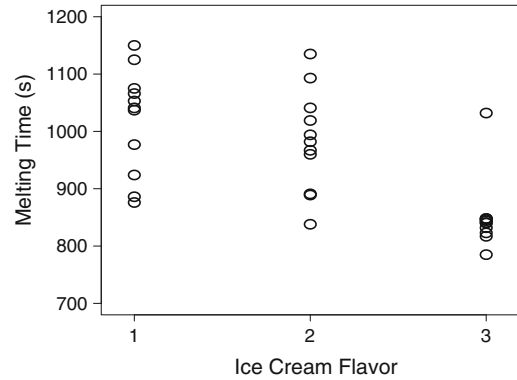
So an unbiased estimate of σ^2 is

$$msE = ssE/(33-3) = 6781.8727 \ \text{seconds}^2.$$

Table 17.2 Melting times for three randomly selected flavors of ice cream. Order of observation in parentheses

Flavor	Time in seconds (order of observation)					
1	924 (1)	876 (2)	1150 (5)	1053 (7)	1041 (10)	1037 (12)
	1125 (15)	1075 (16)	1066 (20)	977 (22)	886 (25)	
2	891 (3)	982 (4)	1041 (8)	1135 (13)	1019 (14)	1093 (18)
	994 (27)	960 (30)	889 (31)	967 (32)	838 (33)	
3	817 (6)	1032 (9)	844 (11)	841 (17)	785 (19)	823 (21)
	846 (23)	840 (24)	848 (26)	848 (28)	832 (29)	

Fig. 17.1 Plot of data for
the ice cream experiment

Similarly,

$$ssT = 11 \sum \bar{y}_{i.}^2 - 33\bar{y}_{..}^2$$
$$= 30{,}003{,}028.8181 - 29{,}830{,}018.9393$$
$$= 173{,}009.8787 \,.$$

So $msT = ssT/(3 - 1) = 86{,}504.9394$ seconds^2, and an unbiased estimate of σ_T^2 is given by

$$\frac{msT - msE}{c} = \frac{86{,}504.9394 - 6781.8727}{11}$$
$$= 7247.5515 \text{ seconds}^2.$$

$\square$

17.3.4 Testing Equality of Treatment Effects

When the treatment factor has random effects, we are interested in the variability of the treatment effects in the entire population of levels, not just those in the experiment. Since the variance of the effects in the population is σ_T^2, the null hypothesis of interest is of the form

$$H_0^T : \sigma_T^2 = 0 \,,$$

and the alternative hypothesis is

$$H_A^T : \sigma_T^2 > 0.$$

It would be very convenient if we could use the same hypothesis-testing rule as we used for testing equality of treatment effects in the fixed-effects model. The fixed-effects decision rule was to reject the hypothesis of no difference in the treatments if $msT/msE > F_{v-1,n-v,\alpha}$, (see (3.5.15), p. 43). Let us examine the ratio msT/msE in the random-effects one-way model (17.3.1). In Sect. 17.3.2 we showed that

$$E[MSE] = \sigma^2 \,,$$

and in Sect. 17.3.3 we showed that

$$E[MST] = c\sigma_T^2 + \sigma^2 \,,$$

where $c = (n^2 - \sum r_i^2)/n(v - 1)$, and if all r_i are equal to r, then $c = r$.

So, if $H_0^T : \sigma_T^2 = 0$ is true, then the expected value of the numerator of the ratio MST/MSE is equal to σ^2, the same as the expected value of the denominator. Then, if H_0^T is true, the ratio should be in the region of 1.0. But if σ_T^2 is large, the expected value of the numerator is larger than the denominator, and the ratio should be large and positive. This situation is similar to that for the fixed-effects case. The only remaining question is whether MST/MSE has an F distribution with $v - 1$ and $n - v$ degrees of freedom when H_0^T is true.

It can be shown that

$$SST/(c\sigma_T^2 + \sigma^2) \sim \chi_{v-1}^2 \qquad (17.3.6)$$

and

$$SSE/\sigma^2 \sim \chi_{n-v}^2$$

and that SST and SSE are independent. Consequently, we have

$$\frac{SST/((c\sigma_T^2 + \sigma^2)(v - 1))}{SSE/(\sigma^2(n - v))} = \frac{MST/(c\sigma_T^2 + \sigma^2)}{MSE/\sigma^2} \sim \frac{\chi_{v-1}^2/(v - 1)}{\chi_{n-v}^2/(n - v)} \sim F_{v-1,n-v} \,, \qquad (17.3.7)$$

and when $\sigma_T^2 = 0$, then

$$\frac{MST}{MSE} \sim F_{v-1,n-v} \,.$$

Thus, to test $H_0 : \sigma_T^2 = 0$ against $H_A : \sigma_T^2 > 0$, our decision rule is to

$$\text{reject } H_0^T \text{ if } \frac{msT}{msE} > F_{v-1,n-v,\alpha} \qquad (17.3.8)$$

for some chosen value of the significance level α. The test can be set out in an analysis of variance table in the usual way; see Table 17.3. We have included the expected mean squares in the table for easy reference.

Rather than testing whether or not the variance of the population of treatment effects is zero, it may be of more interest to test whether the variance is less than or equal to some proportion of the error

Table 17.3 Analysis of variance table for the random-effects one-way model

Source of variation	Degrees of freedom	Sum of squares	Mean squares	Ratio	Expected mean square
Treatments	$v - 1$	ssT	$\frac{ssT}{v-1}$	$\frac{msT}{msE}$	$c\sigma_T^2 + \sigma^2$
Error	$n - v$	ssE	$\frac{ssE}{n-v}$		σ^2
Total	$n - 1$	$sstot$			
Computational formulae					

$$ssT = \sum_i r_i \bar{y}_{i.}^2 - n\bar{y}_{..}^2 \qquad\qquad ssE = \sum_i \sum_t y_{it}^2 - \sum_i r_i \bar{y}_{i.}^2$$

$$sstot = \sum_i \sum_t y_{it}^2 - n\bar{y}_{..}^2 \qquad\qquad c = \frac{n^2 - \sum r_i^2}{n(v-1)}$$

variance, that is,

$$H_0^{\gamma T} : \sigma_T^2 \le \gamma\sigma^2 \text{ and } H_A^{\gamma T} : \sigma_T^2 > \gamma\sigma^2 ,$$

for some constant γ. From (17.3.7), we see that if $H_0^{\gamma T}$ is true with $\sigma_T^2 = \gamma\sigma^2$, then

$$\frac{MST/(\sigma^2(c\sigma_T^2/\sigma^2 + 1))}{MSE/\sigma^2} = \frac{MST}{MSE(c\gamma + 1)} \sim F_{v-1,n-v} .$$

So, our decision rule (17.3.8) needs only the minor modification of including the constant $(c\gamma + 1)$, that is,

$$\text{reject } H_0^{\gamma T} \text{ if } \frac{msT}{msE} > (c\gamma + 1)F_{v-1,n-v,\alpha} . \tag{17.3.9}$$

If we choose $\gamma = 0$, then the decision rule (17.3.9) reduces to rule (17.3.8) for testing the null hypothesis $H_0^T : \sigma_T^2 = 0$ against its alternative hypothesis $H_A^T : \sigma_T^2 > 0$.

Example 17.3.2 Ice cream experiment, continued

The analysis of variance table for the ice cream experiment of Example 17.3.1 is shown in Table 17.4. If we test the null hypothesis that the variance of melting times in the population of ice creams is negligible against the alternative hypothesis that it is not (that is, $H_0^T : \sigma_T^2 = 0$ versus $H_A^T : \sigma_T^2 > 0$) with a Type I error probability of $\alpha = 0.05$, we would reject H_0^T, since

$$msT/msE = 12.76 > F_{2,30,0.05} = 3.32 ,$$

or equivalently, the *p*-value is less than 0.05.

In such an experiment there will clearly be considerable error variability in the data due to fluctuations of room temperature and the difficulty of determining the exact time at which the ice cream has melted completely. Variability in the melting time of different flavors is unlikely to be of interest to the experimenter unless it is larger than the error variability. Suppose, therefore, instead of testing the hypothesis H_0^T against H_A^T, we test the null hypothesis $H_0^{\gamma T} : \{\sigma_T^2 \le \sigma^2\}$ against $H_A^{\gamma T} : \{\sigma_T^2 > \sigma^2\}$. Since there are $r = 11$ observations on each ice cream, the constant is $c = 11$, and the hypothesis-testing rule (17.3.9) with $\gamma = 1.0$ becomes

$$\text{reject } H_0^{\gamma T} \text{ if } \frac{msT}{msE} > (11 + 1)F_{2,30,\alpha} ,$$

that is,

$$\text{reject } H_0^{\gamma T} \text{ if } 12.76 > 12 \, F_{2,30,\alpha} .$$

Table 17.4   Analysis of variance table for the ice cream experiment

Source of variation	Degrees of freedom	Sum of squares	Mean squares	Ratio	*p*-value
Flavor	2	173009.8788	86504.9394	12.76	0.0001
Error	30	203456.1818	6781.8727		
Total	32	376466.0606			

It can be seen from the table in Appendix A.6 that for any practical choice of α, there is not sufficient evidence to reject the null hypothesis. Thus, although the variation in the melting times of the different flavors is significant, so apparently $\sigma_T^2 > 0$, sufficient evidence has not been gathered to be able to claim that the variation is significantly larger than the error variation in the data. $\square$

17.3.5   Confidence Intervals for Variance Components

We showed in Sect. 17.3.1, p. 618, that the response variable Y_{it} in a random-effects one-way model (17.3.1) has a normal distribution with variance $\sigma^2 + \sigma_T^2$, where σ^2 is the variance of the error variables and σ_T^2 is the variance of the treatment effects in the population. In order to assess the variability of the treatment-effect population, we may wish to calculate a confidence interval for σ_T^2 or, alternatively, for σ_T^2/σ^2 if we want to assess the treatment variability relative to the error variability. Since the latter is the easier calculation, we investigate this first.

Confidence Intervals for σ_T^2/σ^2

From (17.3.7), p. 623, we know that

$$\frac{MST}{MSE(c\sigma_T^2/\sigma^2 + 1)} \sim F_{v-1,n-v}\,, \tag{17.3.10}$$

where $c = (n^2 - \Sigma r_i^2)/(n(v-1))$, and if the r_i are all equal to r, then $c = r$. From this, we can write down an interval in which MST/MSE lies with probability $1 - \alpha$; that is,

$$P\left(F_{v-1,n-v,1-\alpha/2} \le \frac{MST}{MSE(c\sigma_T^2/\sigma^2 + 1)} \le F_{v-1,n-v,\alpha/2} \right) = 1 - \alpha.$$

If we rearrange the left-hand inequality, we find that

$$c\sigma_T^2/\sigma^2 \le \frac{MST}{MSE\, F_{v-1,n-v,1-\alpha/2}} - 1\,,$$

and similarly for the right-hand inequality,

$$c\sigma_T^2/\sigma^2 \ge \frac{MST}{MSE\, F_{v-1,n-v,\alpha/2}} - 1\,.$$

So, replacing the random variables by their observed values, we obtain a $100(1 - \alpha)\%$ confidence interval for σ_T^2/σ^2 as

$$\frac{1}{c}\left[\frac{msT}{msE\, F_{v-1,n-v,\alpha/2}} - 1 \right] \quad \le \frac{\sigma_T^2}{\sigma^2} \le \quad \frac{1}{c}\left[\frac{msT}{msE\, F_{v-1,n-v,1-\alpha/2}} - 1 \right].$$

$$\tag{17.3.11}$$

A drawback of this interval is that if msT is not much larger than msE (or perhaps smaller), then it is possible for the left-hand end of the interval to be negative even though σ_T^2/σ^2 can never be negative. Although we could replace a negative lower bound by zero, we will not do so, since it can result in a

short interval, giving the misleading impression that the experiment was more accurate than it actually was.

For calculation of the interval, remember that $F_{v-1,n-v,\alpha/2}$ denotes the percentile of the $F_{v-1,n-v}$ distribution corresponding to a probability of $\alpha/2$ in the right-hand tail. Also, $F_{v-1,n-v,1-\alpha/2}$ denotes the percentile corresponding to a probability of $\alpha/2$ in the left-hand tail, that is, $1 - \alpha/2$ in the right-hand tail. Since $F_{v-1,n-v,1-\alpha/2}$ is not tabulated in Appendix A.6, it is important to note that

$$F_{v-1,n-v,1-\alpha/2} = (F_{n-v,v-1,\alpha/2})^{-1}. \tag{17.3.12}$$

Example 17.3.3 `Ice cream experiment, continued`

In the ice cream experiment of Examples 17.3.1 and 17.3.2, pp. 621 and 624, the variance σ_T^2 in the melting times (in seconds) of the population of different flavors of ice cream is substantially greater than zero but not substantially greater than the error variance σ^2. A confidence interval for σ_T^2/σ^2 can be obtained using (17.3.11). The values $msT = 86504.9394$, $msE = 6781.8727$, $v = 3$, $c = r = 11$, and $n = 33$ are obtained from Example 17.3.2. From the table in Appendix A.6, we have

$$F_{2,30,.05} = 3.32 \quad \text{and} \quad F_{2,30,.95} = (F_{30,2,.05})^{-1} = (19.5)^{-1} = 0.0513 \,.$$

Therefore, the confidence interval (17.3.11) becomes

$$\frac{1}{11} \left(\frac{86504.9394}{(6781.8727)(3.32)} - 1 \right) \le \frac{\sigma_T^2}{\sigma^2} \le \frac{1}{11} \left(\frac{86504.9394}{(6781.8727)(0.0513)} - 1 \right) ,$$

that is,

$$\sigma_T^2/\sigma^2 \in (0.258, 22.513) \,.$$

This interval is too wide to be of much practical use, since it says that with 95% confidence, σ_T^2 could be 4 times smaller or as much as 22 times bigger than σ^2. However, the result does agree with the test of the null hypothesis $H_0^{\gamma T}$ in Example 17.3.2, since the interval includes the value $\sigma_T^2/\sigma^2 = 1.0$. □

As can be seen from Example 17.3.3, a confidence interval for σ_T^2/σ^2 can be very wide. Not only do we need sufficient numbers of observations on each treatment in the experiment in order to keep a confidence interval narrow, but we also need a sufficiently large selection of treatments to represent the population. In Example 17.3.3, there were only $v = 3$ treatments to represent an entire population of ice cream flavors, and this has contributed to the lack of precision in the experiment. Calculation of sample sizes will be discussed in Sect. 17.4.

Confidence Intervals for σ_T^2

There are various methods of obtaining approximate $100(1 - \alpha)\%$ confidence intervals for σ_T^2. The only method that we shall give here is one that is useful when σ_T^2 is not close to zero and that can be easily adapted when we have more complicated models.

First, remember that an unbiased estimator for σ_T^2 was obtained in Eq. (17.3.5), p. 621, as

$$U = c^{-1}(MST - MSE) , \tag{17.3.13}$$

where $c = (n^2 - \Sigma r_i^2)/(n(v - 1))$, and $c = r$ when the sample sizes are equal. If we can determine the distribution of U, then we can easily find a confidence interval for σ_T^2. We know that for the

random-effects one-way model, $SST/(c\sigma_T^2 + \sigma^2) \sim \chi_{v-1}^2$ and $SSE/\sigma^2 \sim \chi_{n-v}^2$ and that SST and SSE are independent. The exact distribution of U is therefore based on the difference of two chi-squared distributions each multiplied by a constant of unknown value, and this is not a standard tabulated distribution. However, it can be shown that a reasonable approximation to the true distribution of U/σ_T^2 is a chi-squared distribution divided by its degrees of freedom x, where x is estimated by

$$x = \frac{(msT - msE)^2}{msT^2/(v-1) + msE^2/(n-v)} . \tag{17.3.14}$$

In other words, the distribution of $xU/E[U]$ is approximately χ_x^2. This result is related to the Satterthwaite approximation that we used in Sect. 5.6.3, p. 115 (Scheffé 1959, Sect. 7.5, gives the general result). Using this approximation, we can write down the approximate probability statement

$$P\left(\chi_{x,1-\alpha/2}^2 \leq \frac{xU}{\sigma_T^2} \leq \chi_{x,\alpha/2}^2 \right) \approx 1 - \alpha .$$

If we rearrange the left-hand inequality, we obtain

$$\sigma_T^2 \leq \frac{xU}{\chi_{x,1-\alpha/2}^2} ,$$

and if we rearrange the right-hand inequality, we obtain

$$\frac{xU}{\chi_{x,\alpha/2}^2} \leq \sigma_T^2 .$$

Consequently, we obtain an approximate $100(1-\alpha)\%$ confidence interval for σ_T^2 as

$$\frac{xu}{\chi_{x,\alpha/2}^2} \leq \sigma_T^2 \leq \frac{xu}{\chi_{x,1-\alpha/2}^2} , \tag{17.3.15}$$

where u is the observed value of U; that is,

$$u = c^{-1}(msT - msE) . \tag{17.3.16}$$

Example 17.3.4 Ice cream experiment, continued

Suppose we require a 90% confidence interval for the variance of the melting times of the population of ice creams in the ice cream experiment of Examples 17.3.1 and 17.3.2, pp. 621 and 624. Using the information in those examples, we obtain the unbiased estimate (17.3.16) of σ_T^2 as $u = 7247.5526$ seconds^2. The degrees of freedom x of the approximate distribution of U are calculated using (17.3.14), that is,

$$x = \frac{(86504.9394 - 6781.8727)^2}{(86504.9394)^2/2 - (6781.8727)^2/30} \approx 1.7.$$

From Table A.5 we can guess at the approximate values of $\chi_{x,.05}^2$ and $\chi_{x,.95}^2$ as

$$\chi_{1.7,.05}^2 \approx 5.3 \quad \text{and} \quad \chi_{1.7,.95}^2 \approx 0.07 .$$

So a 90% confidence interval for σ_T^2 is roughly

$$\sigma_T^2 \in \left(\frac{(1.7)(7247.5515)}{5.3}, \; \frac{(1.7)(7247.55)}{0.07} \right)$$
$$= (2{,}324.69, \;\; 176{,}011.97)$$

Taking square roots and converting to minutes, we obtain the approximate 90% confidence interval for the standard deviation of melting times as

$$\sigma_T \in (0.8, \; 7.0) \text{ minutes.}$$

Again, this interval is too wide for practical use, due to the small number of flavors examined from the population. □

17.4   Sample Sizes for an Experiment with One Random Effect

For the fixed-effects one-way analysis of variance model, we looked at two different ways of determining sample sizes. The first method (Sect. 3.6) was based on the required power of the hypothesis test for detecting whether two treatment effects differ by more than a chosen quantity Δ. The second method (Sect. 4.5) was based on the required length of confidence intervals for one or more treatment contrasts.

For the random-effects one-way model, we need to determine both the number v of levels of the treatment factor to be observed in the experiment and the number r of observations to be taken on each of these levels. A glance at the formulae (17.3.11) and (17.3.15) shows that a calculation of v and r based on the lengths of confidence intervals will not be straightforward. Both formulae depend not only on the value of msE, but also on that of msT, both of which are unknown prior to the experiment. However, consideration of the variances of the estimators used to develop the confidence intervals helps us determine an appropriate balance between "more treatments" and "more replication."

Consider first the confidence interval for σ_T^2 given in (17.3.15). The confidence interval should be tight if the variance of the unbiased estimator U is small. Assuming equal sample sizes, $U = r^{-1}(MST - MSE)$ has variance

$$\text{Var}(U) = \frac{2n^2(n\sigma_T^2/v + \sigma^2)^2}{v^2(v-1)} + \frac{2n^2\sigma^4}{v^2(n-v)}$$

for $n > v$. (This follows because MST and MSE are independent, $SST/(r\sigma_T^2 + \sigma^2) \sim \chi^2(v-1)$, $SSE/\sigma^2 \sim \chi^2(n-v)$, and the variance of a chi-squared random variable is twice its degrees of freedom.) We want this variance to be as small as possible. Suppose that the total number of observations $n = rv$ is fixed by budget considerations and we require $r \geq 2$ since replication is needed to estimate σ^2, so $v \leq n/2$. To minimize the first term of this variance, we require v as large as possible, corresponding to $v = n/2$. It can be shown that the second term is minimized by taking $v = 2n/3$, or by taking $v = n/2$ if we require equal sample sizes and $r \geq 2$.

In summary, assuming equal replication with $r \geq 2$, the variance of U is minimized by taking $v = n/2$ and $r = 2$. In this case, our estimator would be $U = (1/2)(MST - MSE)$, for which $\text{Var}(U) = 8[(2\sigma_T^4 + \sigma^2)^2/(v-1) + \sigma^4/v]$ is as small as possible given equal replication with $r \geq 2$.

We find a similar requirement resulting from a confidence interval for σ_T^2/σ^2. The mean of an F-distribution with $v - 1$ and $n - v$ degrees of freedom is $(n-v)/(n-v-2)$. So, from (17.3.10),

p. 625, with $c = r$, an unbiased estimator of σ_T^2/σ^2 is given by

$$U = \frac{1}{r}\left[\frac{(n-v-2)}{(n-v)}\frac{MST}{MSE} - 1\right],$$

and a narrow confidence interval should be obtained if we choose v and r to make $\text{Var}(U)$ small. The variance of an F-distribution with m and p degrees of freedom is

$$\frac{2p^2(m+p-2)}{m(p-2)^2(p-4)}.$$

It follows from (17.3.10), the definition of U, and $m = v - 1$ and $p = n - v$ that when the sample sizes are all equal to r,

$$\text{Var}(U) = \left(r\frac{\sigma_T^2}{\sigma^2} + 1\right)^2 \frac{1}{r^2}\left(\frac{2(n-v)^2(n-3)(n-v-2)^2}{(v-1)(n-v-2)^2(n-v-4)(n-v)^2}\right)$$

$$= \left(\frac{\sigma_T^2}{\sigma^2} + \frac{1}{r}\right)^2 \left(\frac{2(n-3)}{(v-1)(n-v-4)}\right).$$

So, if the number of observations n is fixed with $n = rv$ and $r \geq 2$, and if we expect that $\sigma_T^2/\sigma^2 > 1/2$, say, then the squared term $(\sigma_T^2/\sigma^2)^2$ from the first set of parentheses will be most important for determining the size of the variance—more important that the term $(1/r)^2$ or the cross-product term— and we need to minimize its coefficient, which is $2(n-3)/((v-1)(n-v-4))$. This requires that $v = (n-3)/2$, or $v = n/2$ and $r = 2$ in the equireplicate case. On the other hand, in the more unusual case when σ_T^2 is expected to be much smaller than σ^2, then the squared term $(1/r)^2$ from the first set of parentheses will be most important for determining the size of the variance, and we need the minimum value of

$$\frac{1}{r^2}\left(\frac{2(n-3)}{(v-1)(n-v-4)}\right) = \frac{2v^2(n-3)}{n^2(v-1)(n-v-4)},$$

and this occurs when v is as small as possible. However, it is unusual to be interested in σ_T^2/σ^2 if this ratio is expected to be very small, so we discount this case.

In summary, the general recommendation again is to set $v = n/2$ and $r = 2$. The exception is the extraordinary case where one plans such an experiment to study σ_T^2/σ^2 but expects σ_T^2 to be much smaller than σ^2, in which case one should choose v to be small.

We can get a feel for how many observations $n = rv$ are needed in total if we examine the power of the hypothesis test for testing $H_0^{\gamma T} : \sigma_T^2 \leq \gamma\sigma^2$ against the alternative hypothesis $H_A^{\gamma T} : \sigma_T^2 > \gamma\sigma^2$ (for a chosen $\gamma \geq 0$). The decision rule was given in (17.3.9), p. 624, as

$$\text{reject } H_0^{\gamma T} \text{ if } \frac{msT}{msE} > (c\gamma + 1)F_{v-1,n-v,\alpha} = k, \text{ say.} \qquad (17.4.17)$$

What is the probability of rejecting $H_0^{\gamma T}$ if the true value of σ_T^2/σ^2 is Δ? In other words, what is the probability that $MST/MSE > k$, when σ_T^2/σ^2 is equal to Δ? This is the power of the test at the value Δ. We can calculate the power from the knowledge that

$$\frac{MST}{MSE(c\sigma_T^2/\sigma^2 + 1)} \sim F_{v-1,n-v}\,,$$

see (17.3.12), p. 626. If σ_T^2/σ^2 is equal to Δ, then

$$P\left(\frac{MST}{MSE} > k\right) = P\left(\frac{MST}{MSE(c\Delta + 1)} > \frac{k}{c\Delta + 1}\right).$$

Suppose we stipulate that the power must be π when $\sigma_T^2/\sigma^2 = \Delta$. Then, we must have that

$$\frac{k}{c\Delta + 1} = F_{v-1,n-v,\pi}.$$

Remembering from (17.4.17) that $k = (c\gamma + 1)F_{v-1,n-v,\alpha}$, and that $(n - v) = v(r - 1)$ and $c = r$ for equal sample sizes, we obtain the equality

$$\frac{F_{v-1,v(r-1),\alpha}}{F_{v-1,v(r-1),\pi}} = \frac{r\Delta + 1}{r\gamma + 1}\,.$$

So, we need to select γ and α for testing $H_0^{\gamma T}$ together with Δ and π. Then we can try to determine v and r by trial and error as illustrated in Example 17.4.1. Since $F_{v-1,v(r-1),\pi} = (F_{v(r-1),v-1,1-\pi})^{-1}$, we try to find values of v and r such that

$$(F_{v-1,v(r-1),\alpha})(F_{v(r-1),v-1,1-\pi}) \le \frac{r\Delta + 1}{r\gamma + 1}\,. \tag{17.4.18}$$

Example 17.4.1 Ice cream experiment, continued

In Example 17.3.2, p. 624, we were unable to reject the hypothesis $H_0^{\gamma T} : \sigma_T^2 \le \sigma^2$ in favor of the hypothesis $H_A^{\gamma T} : \sigma_T^2 > \sigma^2$ at a significance level of $\alpha = 0.05$. Suppose we wish to repeat this experiment, still with $\gamma = 1.0$ and a Type I error probability of $\alpha = 0.05$. Suppose further that we would like to reject the hypothesis with high probability (say $\pi = 0.95$) if the true value of σ_T^2/σ^2 is at least $\Delta = 2.0$. How many ice cream flavors should we look at and how many observations should we take on each?

From (17.4.18), we need to find v and r such that

$$(F_{v-1,v(r-1),.05})(F_{v(r-1),v-1,.05}) \le (2r + 1)/(r + 1)\,.$$

For the moment set $r = 11$, which is the value used by the experimenter in the ice cream experiment. Then $(2r + 1)/(r + 1) = 23/12 \approx 1.92$, and we have

v	$F_{v-1,10v,.05}$	$F_{10v,v-1,.05}$	Product	$\frac{2r+1}{r+1}$	Action
4	2.84	8.59	24.40	$>$ 1.92	Increase v
100	1.26	1.30	1.64	$<$ 1.92	Decrease v
80	1.30	1.34	1.74	$<$ 1.92	Decrease v
60	1.34	1.41	1.89	$\approx$ 1.92	Stop

So, v around 60 should be reasonable, requiring 660 observations in total, including 60 ice cream varieties.

Let us now reduce r to 3 and compute the required value of v. Then $(2r + 1)/(r + 1) = 1.75$, so

v	$F_{v-1,2v,.05}$	$F_{2v,v-1,.05}$	Product	$\frac{2r+1}{r+1}$	Action
80	1.37	1.39	1.90	> 1.75	Increase v
100	1.32	1.34	1.78	> 1.75	Increase v
105	1.31	1.33	1.75	= 1.75	Stop

So v in the region of 105 would be fine, requiring only $n = 315$ observations. In Exercise 2, the reader is asked to determine whether the use of $r = 2$ would require a smaller total number of observations. To greatly reduce the required number of observations, we would need to relax our requirement of such a high power to reject $H_0^{\gamma^T} : \sigma_T^2 \leq \sigma^2$ when $\sigma_T^2 = 2\sigma^2$. $\square$

17.5   Checking Assumptions on the Model

The simplest way to check the assumptions on the one-way random-effects model is to use residual plots in much the same way as for a fixed-effects one-way model. We need to check that the error assumptions are valid, that is,

$$\epsilon_{it} \sim N(0, \sigma^2), \quad t = 1, \ldots, r_i,$$

for each treatment factor level i ($i = 1, \ldots, v$), and also that the assumptions on the random effect T_i are valid, that is,

$$T_i \sim N(0, \sigma_T^2), \quad i = 1, \ldots, v,$$

and that all random variables are mutually independent.

Checking the error assumptions is straightforward, since we proceed in exactly the same way as for the fixed-effects one-way model. We replace T_i in the model, temporarily, by the fixed effect τ_i. Then the residuals are defined as usual as

$$\widehat{e}_{it} = y_{it} - \widehat{y}_{it} = y_{it} - \bar{y}_{i.} .$$

These are then standardized to obtain the standardized residuals z_{it} with standard deviation 1.0. We plot the standardized residuals versus treatment-factor levels, versus $\widehat{y}_{it}$, versus order, and versus normal scores, as in Chap. 5, to check for outliers, independence, constant variance, and normality. Non-independence between the ϵ_{it}'s and the T_i's is not easy to detect, but unequal variances of the ϵ_{it}'s indicates one form of the problem.

The normality assumption on the random effect T_i can be checked when the sample sizes are equal, unless v is too small. The treatment averages $\overline{Y}_{i.}$ should have a $N(\mu, \sigma_T^2 + \sigma^2/r)$ distribution. So, if we standardize the observed averages $\bar{y}_{i.}$ to have average value zero and sample standard deviation one, and we plot these standardized averages against their corresponding normal scores, we should roughly obtain a straight line—one that cuts the vertical axis at about zero and that has slope about one. It is important to check the normality assumption, since the analysis for random-effects models is not robust to nonnormality of the random effects. We can also use this normal probability plot to check for treatment effect outliers among the observed treatments. In an experiment such as the ice cream experiment, where only $v = 3$ levels of the treatment factor were observed, there is not enough data to be able to examine the distribution of the T_i's in any detail. In Sects. 17.10 and 17.11 we will

illustrate the assumption-checking procedures using the SAS and R software packages, respectively, and the data from the clean wool experiment that was described in Sect. 17.2.2, p. 615.

17.6 Two or More Random Effects

17.6.1 Models and Examples

In the ice cream experiment of Example 17.3.1, p. 621, we modeled the ice cream effect as a random effect, since we were interested in the variability of the melting rates of varieties of a large population of all possible ice creams with similar ingredients. If the experimenter had also been interested in whether or not the container affects the melting time, then she might have randomly selected a number b of containers from the population of all possible containers. If one ice cream melts faster than another ice cream in one container, then it might be safe to assume that it melts faster, and by the same amount, in another container. In other words, the assumption of no ice cream×container interaction might be reasonable. In this case a random two-way main-effects model (with no interaction) would be a possible model; that is,

$$Y_{ijt} = \mu + A_i + B_j + \epsilon_{ijt}, \tag{17.6.19}$$
$$A_i \sim N(0, \sigma_A^2), \quad B_j \sim N(0, \sigma_B^2), \quad \epsilon_{ijt} \sim N(0, \sigma^2),$$
$$A_i\text{'s}, \ B_j\text{'s and } \epsilon_{ijt}\text{'s are all mutually independent}$$
$$t = 1, \ldots, r_{ij}, \ i = 1, \ldots, a, \ j = 1, \ldots, b.$$

where A_i is the effect of the ith ice cream randomly selected from the population of ice creams whose effects on melting times follow a normal distribution with variance σ_A^2 for each container, and where B_j is the effect of the jth container randomly selected from the population of containers whose effects on the melting times follow a normal distribution with variance σ_B^2 for each ice cream. The number of observations r_{ij} to be taken on the (ij)th ice cream–container combination needs to be determined. Normally, we would select the r_{ij}'s to be equal if possible.

Alternatively, it may be expected that a slightly thicker container would show a greater difference in melting times of ice creams than would a thinner container. In other words, an interaction may be expected. In this case, we would add to model (17.6.19) a random effect representing the interaction, as shown in the random-effects two-way complete model (17.6.20):

$$Y_{ijt} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijt} \tag{17.6.20}$$
$$A_i \sim N(0, \sigma_A^2), \quad B_j \sim N(0, \sigma_B^2)$$
$$(AB)_{ij} \sim N(0, \sigma_{AB}^2), \quad \epsilon_{ijt} \sim N(0, \sigma^2)$$
$$A_i\text{'s}, \ B_j\text{'s}, \ (AB)_{ij}\text{'s and } \epsilon_{ijt}\text{'s are mutually independent}$$
$$t = 1, \ldots, r_{ij}, \ i = 1, \ldots, a, \ j = 1, \ldots, b.$$

If σ_{AB}^2 is positive, then there are *AB effects* present—namely, main effects and interactions for the factors A and B. If σ_A^2 or σ_B^2 is positive, then the corresponding main effects are present.

Example 17.6.1 Ammunition experiment

W.A. Thompson, Jr. and J.R. Moore in the 1963 volume of *Technometrics* describe an experiment concerning the muzzle velocity characteristics of ammunition for a field artillery weapon. They describe the ammunition as follows:

Table 17.5 Data for the ammunition experiment

		Charge lot			
		1	2	3	4
Projectile lot	1	63	56	69	78
		78	58	63	79
	2	71	60	64	65
		70	65	68	77
	3	72	58	69	63
		55	55	71	72
	4	70	60	66	73
		64	71	68	79

Source Thompson and Moore (1963). Copyright © 1963 American Statistical Association. Reprinted with permission

> Propelling charges and projectiles for this type of weapon are manufactured and stored separately in a such a way that any charge might be employed by the user to propel any projectile…. Both projectiles and charges are grouped into lots at the time of manufacture, each lot consisting of a large number of individual units assembled during a short period of time using essentially uniform components. Thus, it is hoped that the round to round dispersion [variability] in velocity will be reduced by using charges and projectiles from within lots.

The experiment involved the examination of a random sample of four charge lots (factor A with levels 1, 2, 3, 4) selected at random from a large population of charge lots, and four projectile lots (factor B with levels 1, 2, 3, 4) selected at random from a large population of projectile lots. A weapon surveillance test was conducted using one weapon under uniform ballistic conditions. The muzzle velocities were measured to the nearest foot per second. These are shown in Table 17.5, except that a constant has been added to each recorded velocity.

Since the lots involved in the experiment were randomly selected from large populations, a random-effects two-way complete model (17.6.20) was used in the analysis.   □

In an experiment with more than two random-effects treatment factors, variables representing all of the main effects of the factors and some or all of their interactions would be included in the model in the obvious way. For example, an experiment with five random-effects treatment factors A, B, C, D, G, in which interactions AB, AC, BC, CD, and ABC were thought to be nonnegligible, would be modeled as follows:

$$Y_{ijklmt} = \mu + A_i + B_j + C_k + D_l + G_m$$
$$+ (AB)_{ij} + (AC)_{ik} + (BC)_{jk} + (CD)_{kl} + (ABC)_{ijk} + \epsilon_{ijklmt},$$
$$A_i \sim N(0, \sigma_A^2), \quad B_j \sim N(0, \sigma_B^2), \quad C_k \sim N(0, \sigma_C^2), \quad D_l \sim N(0, \sigma_D^2), \quad G_m \sim N(0, \sigma_G^2),$$
$$(AB)_{ij} \sim N(0, \sigma_{AB}^2), \quad (AC)_{ik} \sim N(0, \sigma_{AC}^2), \quad (BC)_{jk} \sim N(0, \sigma_{BC}^2), \quad (CD)_{kl} \sim N(0, \sigma_{CD}^2),$$
$$(ABC)_{ijk} \sim N(0, \sigma_{ABC}^2), \quad \epsilon_{ijklmt} \sim N(0, \sigma^2),$$

all random variables on the right-hand side

of the model are mutually independent,

$$t = 1, \ldots, r_{ijklm}, \quad i = 1, \ldots, a, \quad j = 1, \ldots, b,$$
$$k = 1, \ldots, c, \quad l = 1, \ldots, d, \quad m = 1, \ldots, g.$$

As for the fixed-effects models, if a high-order interaction is included in the model, then so are all of its "subinteractions" and constituent main effects; that is, if $(ABC)_{ijk}$ is in the model, so are $(AB)_{ij}$, $(AC)_{ik}$, $(BC)_{jk}$, A_i, B_j, and C_k.

17.6.2   Checking Model Assumptions

We may check the error assumptions by replacing temporarily all of the random effects by fixed effects, calculating the standardized residuals, and examining the residual plots in the usual way. Checking the assumptions of each random effect is not easy, since in a two-way or higher-way model there are generally few levels of each treatment factor observed, and the cell averages are not independent. Consequently, we will omit the model checks for the random-effect assumptions, and hope that any severe problems will show up through the analysis of the residuals.

17.6.3   Estimation of σ^2

In Sect. 17.3.2 we found that for the one-way random-effects model, an unbiased estimate of σ^2 was given by msE, where msE was calculated exactly as for the fixed-effects one-way model. Perhaps this should not be surprising, since msE measures the variability in the data that is not accounted for by those sources of variation that were ignored in the experiment. An unbiased estimate for σ^2 in *any* random-effects model can be obtained from its fixed-effects model counterpart.

Example 17.6.2 Unbiased estimate of σ^2

We will show that an unbiased estimate of σ^2 in the random-effects two-way complete model is $msE = ssE/(n - v)$, where

$$ssE = \left[\sum_i \sum_j \sum_t y_{ijt}^2 - \sum_i \sum_j r_{ij} \bar{y}_{ij.}^2 \right],$$

as in (6.4.16) for the fixed-effects two-way complete model. First, note that

$$E[Y_{ijt}] = \mu \ \text{ and } \ \text{Var}(Y_{ijt}) = \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma^2$$

for the random-effects two-way complete model. Also,

$$\bar{Y}_{ij.} = \mu + A_i + B_j + (AB)_{ij} + \sum_t \epsilon_{ijt}/r_{ij}\,.$$

So,

$$E[\bar{Y}_{ij.}] = \mu \ \text{ and } \ \text{Var}(\bar{Y}_{ij.}) = \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma^2/r_{ij}.$$

Thus, the expected value of the random variable SSE is

$$E\,[SSE] = E\left[\sum_i \sum_j \sum_t Y_{ijt}^2 - \sum_i \sum_j r_{ij} \bar{Y}_{ij.}^2 \right]$$

$$= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{t=1}^{r_{ij}} (\text{Var}(Y_{ijt}) + E[Y_{ijt}]^2)$$

$$-\sum_{i=1}^{a}\sum_{j=1}^{b} r_{ij}(\text{Var}(\overline{Y}_{ij\cdot}) + E[\overline{Y}_{ij\cdot}]^2)$$

$$= \left(\sum_{i=1}^{a}\sum_{j=1}^{b} r_{ij}\right)\sigma^2 - \sum_{i=1}^{a}\sum_{j=1}^{b}\sigma^2 = (n - v)\sigma^2,$$

where $n = \Sigma\Sigma r_{ij}$ and $v = ab$. Consequently,

$$E[MSE] = E[SSE/(n - v)] = \sigma^2.$$

$\square$

17.6.4 Estimation of Variance Components

In Sect. 17.3.3, p. 620, we found that for the random-effects one-way model,

$$E[MST] = c\sigma_T^2 + \sigma^2 \ , \quad \text{where } c = \frac{n^2 - \sum r_i^2}{n(v - 1)} \ ,$$

where MST is the mean square for treatments from the fixed-effects one-way model, and where $c = r$ if all the sample sizes are equal. From this, we were able to find an unbiased estimator for σ_T^2, namely $(MST - MSE)/c$.

For more complicated models, we will also be able to find unbiased estimators for the variance components using the fixed-effects mean squares, but each estimator must be calculated individually.

Example 17.6.3 Unbiased estimate of σ_B^2

Suppose an experiment involves three random-effects treatment factors A, B, and D having a, b, and d levels, respectively, and suppose r observations are taken on each of the $v = abd$ combinations. If the only interactions that are expected to be nonnegligible are AB and BD, then, the model is

$$Y_{ijk} = \mu + A_i + B_j + D_k + (AB)_{ij} + (BD)_{jk} + \epsilon_{ijkt},$$

$$t = 1, \ldots, r, \ \ i = 1, \ldots, a, \ \ j = 1, \ldots, b, \ \ k = 1, \ldots, d$$

with the usual assumptions about the distributions of the random treatment effects and error variables.

Suppose we want an unbiased estimator for σ_B^2. We start by investigating $E[MSB]$, where $MSB = SSB/(b - 1)$. Using rule 4, p. 209, the sum of squares for B is

$$ssB = adr\sum_{j=1}^{b}\bar{y}_{\cdot j\cdot}^2 - abdr\bar{y}_{\cdots}^2 \ .$$

Now,

$$E[\overline{Y}_{\cdot j\cdot}] = E[\overline{Y}_{\cdots}] = \mu$$

and

$$\text{Var}[\overline{Y}_{.j..}] = \frac{\sigma_A^2}{a} + \sigma_B^2 + \frac{\sigma_D^2}{d} + \frac{\sigma_{AB}^2}{a} + \frac{\sigma_{BD}^2}{d} + \frac{\sigma^2}{adr} \,,$$

$$\text{Var}[\overline{Y}_{....}] = \frac{\sigma_A^2}{a} + \frac{\sigma_B^2}{b} + \frac{\sigma_D^2}{d} + \frac{\sigma_{AB}^2}{ab} + \frac{\sigma_{BD}^2}{bd} + \frac{\sigma^2}{abdr} \,.$$

Consequently,

$$E[SSB] = adr \sum_j \left[Var(\overline{Y}_{.j..}) + E[\overline{Y}_{.j..}]^2 \right] - abdr \left[Var(\overline{Y}_{....}) + E[\overline{Y}_{....}]^2 \right]$$

$$= adr(b-1)\sigma_B^2 + dr(b-1)\sigma_{AB}^2 + ar(b-1)\sigma_{BD}^2 + (b-1)\sigma^2 \,.$$

So,

$$E[MSB] = adr\sigma_B^2 + dr\sigma_{AB}^2 + ar\sigma_{BD}^2 + \sigma^2 \,. \tag{17.6.21}$$

Thus, if we wish to find an unbiased estimator for σ_B^2, we must find unbiased estimators also for σ_{AB}^2 and σ_{BD}^2. The logical place to look for these is at $E[MS(AB)]$ and $E[MS(BD)]$. We have

$$ss(AB) = dr \sum_{i=1}^a \sum_{j=1}^b \overline{y}_{ij..}^2 - bdr \sum_{i=1}^a \overline{y}_{i...}^2 - adr \sum_{j=1}^b \overline{y}_{.j..}^2 + abdr\overline{y}_{....}^2 \,,$$

$$ss(BD) = ar \sum_{j=1}^b \sum_{k=1}^d \overline{y}_{.jk.}^2 - adr \sum_{j=1}^b \overline{y}_{.j..}^2 - abr \sum_{k=1}^d \overline{y}_{..k.}^2 + abdr\overline{y}_{....}^2 \,,$$

and

$$E[\overline{Y}_{ij..}] = E[\overline{Y}_{i...}] = E[\overline{Y}_{.j..}] = E[\overline{Y}_{.jk.}] = E[\overline{Y}_{..k.}] = [\overline{Y}_{....}] = \mu \,,$$

$$\text{Var}[\overline{Y}_{ij..}] = \sigma_A^2 + \sigma_B^2 + \frac{\sigma_D^2}{d} + \sigma_{AB}^2 + \frac{\sigma_{BD}^2}{d} + \frac{\sigma^2}{dr} \,,$$

$$\text{Var}[\overline{Y}_{i...}] = \sigma_A^2 + \frac{\sigma_B^2}{b} + \frac{\sigma_D^2}{d} + \frac{\sigma_{AB}^2}{b} + \frac{\sigma_{BD}^2}{bd} + \frac{\sigma^2}{bdr} \,,$$

$$\text{Var}[\overline{Y}_{.jk.}] = \frac{\sigma_A^2}{a} + \sigma_B^2 + \sigma_D^2 + \frac{\sigma_{AB}^2}{a} + \sigma_{BD}^2 + \frac{\sigma^2}{ar} \,,$$

$$\text{Var}[\overline{Y}_{..k.}] = \frac{\sigma_A^2}{a} + \frac{\sigma_B^2}{b} + \sigma_D^2 + \frac{\sigma_{AB}^2}{ab} + \frac{\sigma_{BD}^2}{b} + \frac{\sigma^2}{abr} \,,$$

as well as

$$E[SS(AB)] = \left(dr \sum_i \sum_j \overline{Y}_{ij..}^2 - bdr \sum_i \overline{Y}_{i...}^2 \right) - (E[SSB])$$

$$= adr(b-1)\sigma_B^2 + adr(b-1)\sigma_{AB}^2 + ar(b-1)\sigma_{BD}^2 + a(b-1)\sigma^2$$

$$\quad - adr(b-1)\sigma_B^2 - dr(b-1)\sigma_{AB}^2 - ar(b-1)\sigma_{BD}^2 - (b-1)\sigma^2$$

$$= dr(a-1)(b-1)\sigma_{AB}^2 + (a-1)(b-1)\sigma^2 \,.$$

So,

$$E[MS(AB)] = dr\sigma_{AB}^2 + \sigma^2.$$

Similarly,

$$E[MS(BD)] = ar\sigma_{BD}^2 + \sigma^2.$$

Thus, an unbiased estimator for σ_B^2 is

$$U = (MSB - MS(AB) - MS(BD) + MSE)/(adr),$$

and an unbiased estimate for σ_B^2 is therefore

$$u = (msB - ms(AB) - ms(BD) + msE)/(adr).$$

$\square$

Calculation of expected mean squares is quite time-consuming, as was seen in Example 17.6.3. However, when sample sizes are all equal, we can exploit the pattern that emerges in studying such examples. All of the variance components that are involved in $E[MSB]$ in (17.6.21) are those whose random effects include the same subscript as for B in the model. Specifically, B has subscript j in the model, and a j also occurs as subscript in $(AB)_{ij}$, $(BD)_{jk}$, and ϵ_{ijkt}. The constant in front of each variance component is the number of observations taken on each combination of subscripts; that is, there are adr observations on each of the b levels of B, there are dr observations on each of the ab levels of AB, and so on.

A similar pattern can be seen for $E[MS(AB)]$ and $E[MS(BD)]$. This gives us a general rule when sample sizes are equal (which we add to the 16 rules in Chap. 7):

17. To obtain the expected mean square for a main effect or interaction in a random-effects model, first note the subscripts on the term representing that effect in the model. Write down a variance component σ^2 for the effect of interest, for the error, and for every interaction whose term in the model includes the noted set of subscripts. Multiply each variance component except σ^2 by the number of observations taken on each level or combination of levels of the corresponding main effect or interaction. Add up the terms.

17.6.5   Confidence Intervals for Variance Components

In the previous subsection a rule was given for calculating the expected mean square corresponding to each term in the model, when the sample sizes are equal. For unequal sample sizes, the mean squares and expected mean squares are best calculated by a computer program.

From the list of expected mean squares, we can find an unbiased estimator for any given variance component, say σ_*^2. Again, this was illustrated in Example 17.6.3. The estimator can always be a linear combination of mean squares, which, in general, we can write as $U = \Sigma k_i(MS)_i$, where k_i is the constant in front of the ith mean square in the linear combination. Then, an approximation to the distribution of xU/σ_*^2 is a chi-squared distribution with x degrees of freedom, where

$$x = \frac{(\Sigma k_i(ms)_i)^2}{\Sigma k_i^2(ms)_i^2/x_i} \tag{17.6.22}$$

and where x_i is the number of degrees of freedom corresponding to the ith mean square and $(ms)_i$ is the observed value of the ith mean square in the linear combination. An example of this formula was given in Sect. 17.3.5, p. 625, for the one-way model. A more complicated example is given below.

Example 17.6.4 `Calculation of degrees of freedom`

We continue Example 17.6.3, which involved a random-effects model with five random effects A_i, B_j, and D_k (corresponding to main effects of factors A, B, and D), and $(AB)_{ij}$ and $(BD)_{jk}$ (corresponding to interactions AB and BD). An unbiased estimator for σ_B^2 was shown to be

$$U = \Sigma k_i(MS)_i = MSB/(adr) - MS(AB)/(adr) - MS(BD)/(adr) + MSE/(adr).$$

An approximation to the distribution of xU/σ_B^2 is a χ_x^2 distribution, where x is given by (17.6.22), that is,

$$x = \frac{[(msB - ms(AB) - ms(BD) + msE)/(adr)]^2}{\frac{msB^2}{(adr)^2(b-1)} + \frac{ms(AB)^2}{(adr)^2(a-1)(b-1)} + \frac{ms(BD)^2}{(adr)^2(b-1)(d-1)} + \frac{msE^2}{(adr)^2 df}}$$

$$= \frac{[msB - ms(AB) - ms(BD) + msE]^2}{\frac{msB^2}{(b-1)} + \frac{ms(AB)^2}{(a-1)(b-1)} + \frac{ms(BD)^2}{(b-1)(d-1)} + \frac{msE^2}{df}},$$

where df is the number of degrees of freedom for error, which can be obtained, as usual, by subtraction. In this example, df is equal to

$$\begin{aligned} df &= (abdr - 1) - (a - 1) - (b - 1) - (d - 1) \\ &\quad - (a - 1)(b - 1) - (b - 1)(d - 1) \\ &= ab(dr - 1) - b(d - 1) + 1 \, . \end{aligned}$$

$\square$

Once we know an approximate distribution for a variance component estimator, we can easily write down a probability statement and convert it to a confidence interval. Suppose that $U = \Sigma k_i(MS)_i$ is an unbiased estimator for σ_*^2 and that xU/σ_*^2 has approximately a χ_x^2 distribution; then

$$P\left(\chi_{x,1-\alpha/2}^2 \leq xU/\sigma_*^2 \leq \chi_{x,\alpha/2}^2\right) \approx 1 - \alpha.$$

Then, if we observe the value of U to be $u = \Sigma k_i(ms)_i$, by manipulating the two inequalities in the probability statement we can obtain the following approximate $100(1 - \alpha)\%$ confidence interval:

$$\frac{xu}{\chi_{x,\alpha/2}^2} \leq \sigma_*^2 \leq \frac{xu}{\chi_{x,1-\alpha/2}^2} \, , \tag{17.6.23}$$

where x is calculated as in (17.6.22). If the estimate u is negative or the calculated degrees of freedom x is extremely small, then this approximate confidence interval procedure should not be used.

Example 17.6.5 `Ammunition experiment, continued`

The ammunition experiment was described in Example 17.6.1, p. 632, and the data were given in Table 17.5. A random-effects two-way complete model (17.6.20) was used. The mean squares for this

Table 17.6 Two-way analysis of variance table for the ammunition experiment

Source of variation	Degrees of freedom	Sum of squares	Mean squares	Expected mean square
Charge (A)	3	669.12	223.04	$8\sigma_A^2 + 2\sigma_{AB}^2 + \sigma^2$
Projectile (B)	3	92.12	30.71	$8\sigma_B^2 + 2\sigma_{AB}^2 + \sigma^2$
Interaction (AB)	9	257.63	28.63	$2\sigma_{AB}^2 + \sigma^2$
Error	16	516.00	32.25	σ^2
Total	31	1534.87		

model are calculated in exactly the same way as for the fixed-effects two-way complete model, and these are shown in the analysis of variance table, Table 17.6. Also listed in the table are the expected mean squares calculated as in rule 17, p. 637.

For example, to calculate the expected mean square for A, we note that the subscript for the term A_i in the model is i, and also that i is included among the subscripts of the terms $(AB)_{ij}$ and ϵ_{ijt}. This means that the expected mean square must include the three variance components

$$\sigma_A^2, \quad \sigma_{AB}^2, \quad \text{and} \quad \sigma^2 \,.$$

The constant in front of σ_A^2 is 8, the number of observations on each charge lot, whereas the constant in front of σ_{AB}^2 is 2, the number of observations on each combination of charge lot and projectile lot.

The expected mean square for AB, $E[MS(AB)] = 2\sigma_{AB}^2 + \sigma^2$, contains only two terms, since only the two terms $(AB)_{ij}$ and ϵ_{ijt} in the model contain both i and j as subscripts. An unbiased estimator for σ_A^2 is given by $U = (MSA - MS(AB))/8$. Also, xU/σ_A^2 has approximately a χ_x^2 distribution, where x is calculated as in (17.6.22). Thus, an unbiased estimate of σ_A^2 from this experiment is

$$u \;=\; (msA - ms(AB))/8 \;=\; (223.04 - 28.63)/8 = 24.30 \,,$$

and the number of degrees of freedom of the associated chi-squared distribution is

$$x = \frac{24.30^2}{\frac{223.04^2}{(8^2)(3)} + \frac{28.63^2}{(8^2)(9)}} = 2.27 \,.$$

Therefore, an approximate 90% confidence interval (17.6.23) for σ_A^2, the variance of velocities arising from the population of charge lots, is

$$\frac{(2.27)(24.3)}{\chi_{2.27,0.05}^2} \;\le\; \sigma_A^2 \;\le\; \frac{(2.27)(24.3)}{\chi_{2.27,0.95}^2} \,,$$

and since $\chi_{2.27,0.05}^2 \approx 6.5$ and $\chi_{2.27,0.05}^2 \approx 0.17$, the approximate 90% confidence interval, in units of (feet per second)^2, is

$$8.49 \;\le\; \sigma_A^2 \;\le\; 324.48 \,.$$

An approximate 90% confidence interval for the standard deviation of the velocity (in feet per second), obtained by taking square roots, is

$$2.91 \;\le\; \sigma_A \;\le\; 18.01 \,.$$

Fig. 17.2 Plot of average
velocity against charge lot
by projectile lot for the
ammunition experiment
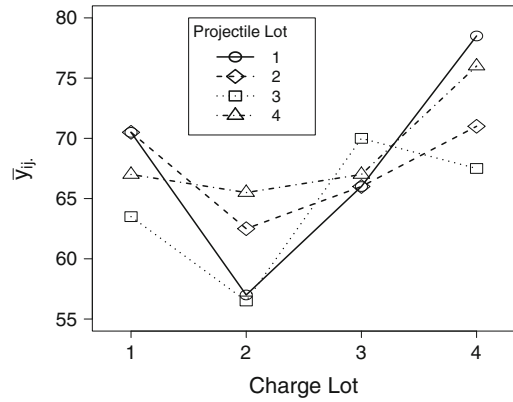
Before leaving this example, we note that an unbiased estimate for σ^2_{AB} calculated in this way is actually negative, since

$$u \; = \; (ms(AB) - msE)/2 \; = \; -1.81 \,,$$

and the calculation for x, the number of degrees of freedom of the associated χ^2 distribution, is

$$x \; = \; \frac{(-1.81)^2}{\frac{28.63^2}{(2^2)(9)} + \frac{32.25^2}{(2^2)(16)}} \; = \; 0.084 \,.$$

Thus, we are not able to say anything sensible about the variance of the interaction, other than that it appears to be very small. The interaction plot in Fig. 17.2 for the lots included in the experiment supports this conclusion.   □

17.6.6 Hypothesis Tests for Variance Components

In order to focus the discussion, we will use the random-effects model of Example 17.6.3, p. 635; that is,

$$Y_{ijkt} = \mu + A_i + B_j + D_k + (AB)_{ij} + (BD)_{jk} + \epsilon_{ijkt} \,,$$

$$t = 1, \ldots, r \, ; \;\; i = 1, \ldots, a \, ; \;\; j = 1, \ldots, b \, ; \;\; k = 1, \ldots, d \,,$$

together with the usual assumptions about the distributions of the random variables. Some of the expected mean squares for this model were calculated in Example 17.6.3 and are listed, together with the remaining mean squares, in Table 17.7.

Testing the hypothesis $H_0^{AB} : \{\sigma^2_{AB} = 0\}$ against its alternative hypothesis $H_A^{AB} : \{\sigma^2_{AB} > 0\}$ is straightforward, since the corresponding expected mean square looks very similar to the situation that we had in the one-way model. If H_0^{AB} is true, then the numerator of the ratio $ms(AB)/msE$ is expected to be σ^2, the same as the denominator. Otherwise, the numerator is expected to be larger. Consequently, the decision rule is

$$\text{reject } H_0^{AB} \text{ if } \frac{ms(AB)}{msE} > F_{(a-1)(b-1),df,\alpha}$$

Table 17.7   Expected mean squares and degrees of freedom for a random-effects three-way model with two interactions

Effect	Degrees of freedom	Expected mean square
A	$a - 1$	$bdr\sigma_A^2 + dr\sigma_{AB}^2 + \sigma^2$
B	$b - 1$	$adr\sigma_B^2 + dr\sigma_{AB}^2 + ar\sigma_{BD}^2 + \sigma^2$
D	$d - 1$	$abr\sigma_D^2 + ar\sigma_{BD}^2 + \sigma^2$
AB	$(a - 1)(b - 1)$	$dr\sigma_{AB}^2 + \sigma^2$
BD	$(b - 1)(d - 1)$	$ar\sigma_{BD}^2 + \sigma^2$
Error	df	σ^2

as usual, where the number of error degrees of freedom is

$$df = ab(dr - 1) - b(d - 1) + 1 .$$

We could modify this test as in (17.3.9), p. 624, so that the decision rule for testing $H_0^{\gamma AB} : \{\sigma_{AB}^2 \leq \gamma\sigma^2\}$ against $H_A^{\gamma AB} : \{\sigma_{AB}^2 > \gamma\sigma^2\}$ is

$$\text{reject } H_0^{\gamma AB} \text{ if } \frac{ms(AB)}{msE} > (1 + dr\gamma)F_{(a-1)(b-1), df, \alpha}.$$

We have similar tests for H_0^{BD} against H_A^{BD}, and $H_0^{\gamma BD}$ against $H_A^{\gamma BD}$.

Testing $H_0^A : \sigma_A^2 = 0$ against $H_A^A : \sigma_A^2 > 0$ is more complicated. Until now, we have used the same test statistics as we used in the fixed-effects case. But if we try to use msA/msE to test H_0^A, we have a problem. If H_0^A is true, so that $\sigma_A^2 = 0$, the expected value of the numerator is $E[MSA] = dr\sigma_{AB}^2 + \sigma^2$, while that of the denominator is $E[msE] = \sigma^2$. This suggests two things:

(i) we should use $ms(AB)$ as the denominator, not msE, and
(ii) we should question whether it makes sense to test H_0^A if the interaction AB is significant.

The second point is, of course, exactly the same point that arose in the fixed-effects model, and the answer is usually "no, it makes no sense." Consequently, we generally test a main effect only when that factor is not involved in any significant interactions. Nevertheless, we shall still use the interaction mean square as the denominator in case an incorrect decision was made regarding the interaction. Consequently, the decision rule for testing H_0^A against H_A^A is

$$\text{reject } H_0^A \text{ if } MSA/MS(AB) > F_{a-1, (a-1)(b-1), \alpha}.$$

Notice that the second set of degrees of freedom for the F-distribution is the degrees of freedom corresponding to the denominator of the ratio. The test for $H_0^D : \{\sigma_D^2 = 0\}$ is similar.

Obtaining a suitable denominator for testing the null hypothesis $H_0^B : \{\sigma_B^2 = 0\}$ versus the alternative hypothesis $H_0^B : \{\sigma_B^2 > 0\}$ is harder again. If H_0^B is true, so that $\sigma_B^2 = 0$, then the expected value of MSB is

$$E[MSB] = dr\sigma_{AB}^2 + ar\sigma_{BD}^2 + \sigma^2 .$$

We would generally want to test this hypothesis only if we believed that the interactions AB and BD were both negligible. Yet to be on the safe side, we would like a denominator with the same expected value. It can be verified that

$$E[U] = E[MS(AB) + MS(BD) - MSE] = dr\sigma_{AB}^2 + ar\sigma_{BD}^2 + \sigma^2.$$

As in Sect. 17.3.5, $xU/(dr\sigma_{AB}^2 + ar\sigma_{BD}^2 + \sigma^2)$ has approximately a chi-squared distribution with degrees of freedom x calculated as in (17.6.22), p. 637; that is,

$$x = \frac{[ms(AB) + ms(BD) - msE]^2}{\frac{[ms(AB)]^2}{(a-1)(b-1)} + \frac{[ms(BD)]^2}{(b-1)(d-1)} + \frac{[msE]^2}{df}}.$$

Therefore, if H_0 is true, msB/U has approximately an $F_{b-1,x}$ distribution. So to test H_0^B against H_A^B, the decision rule is

$$\text{reject } H_0^B \text{ if } \frac{msB}{ms(AB) + ms(BD) - msE} > F_{(b-1),x,\alpha}.$$

Example 17.6.6 Ammunition experiment, continued

An unbiased estimate for the variance of the muzzle velocities due to the population of charge lots (factor A) was calculated to be $u = 24.3$ (feet per second)^2 in Example 17.6.5 for the ammunition experiment. A question is whether this value could be due to random error or whether the variance is really sizable; that is, we wish to test the hypothesis $H_0^A : \{\sigma_A^2 = 0\}$ against the alternative hypothesis $H_A^A : \{\sigma_A^2 > 0\}$. The interaction variability was found to be very small in Example 17.6.5, so the main-effect hypothesis makes sense. The expected mean squares for A and AB are listed in Table 17.6 as

$$E[MSA] = 8\sigma_A^2 + 2\sigma_{AB}^2 + \sigma^2$$

and

$$E[MS(AB)] = 2\sigma_{AB}^2 + \sigma^2,$$

with 3 and 9 corresponding degrees of freedom, respectively. The decision rule, therefore, is

$$\text{reject } H_0^A \text{ if } \frac{msA}{ms(AB)} > F_{3,9,\alpha}.$$

If we select a Type I error probability of $\alpha = 0.05$, then $F_{3,9,0.05} = 3.86$. Since $msA/ms(AB) = 223.04/28.63 = 7.79$, we can conclude that $\sigma_A^2 > 0$.   □

17.6.7 Sample Sizes

If we test main effects and interactions only when the higher-order interactions involving those factors are negligible, then we can adapt (17.4.18) by changing the degrees of freedom to match those in the decision rule being used.

17.7 Mixed Models

Models that contain both random and fixed treatment effects are called *mixed models*. The analysis of random effects proceeds in exactly the same way as described in the previous sections. All that is needed is a way to write down the expected mean squares. The fixed effects can be analyzed as in Chaps. 3–7, except that, here, too, we may need to replace the mean square for error by a different

appropriate mean square. We show how to calculate the expected mean squares for a mixed model in Sect. 17.7.1.

An interaction between two or more factors any of which has random effects will be regarded as a random effect, since the combination of levels observed in the experiment depends upon the random selection of levels of those factors that have random effects.

17.7.1 Expected Mean Squares and Hypothesis Tests

Expected mean squares can be obtained for a mixed model when the sample sizes are equal by modifying rule 17 on p. 637. We start by writing out the expected mean squares as though all the factors were random. We then collect all of the fixed effects and list them together as one "quadratic form." The quadratic form is a function of fixed-effect parameters such as $\alpha_i^* = \alpha_i + (\overline{\alpha\beta})_{i.}$ (see Example 17.7.1) that typically feature in fixed-effects models.

As an example, consider a model containing the main effects of factors A, B, and D and the interactions AB and BD. Suppose that factors A and B have fixed effects, so that all of their levels of interest are observed in the experiment, and factor D has random effects, so that its levels form a large population of which only a random selection are observed in the experiment. Then interaction AB is a fixed effect, but interaction BD is a random effect.

We use α_i to represent the effect of the ith level of A, β_j to represent the effect of the jth level of B, and $(\alpha\beta)_{ij}$ to represent their interaction. The effect of the kth randomly selected level of D is represented by the random variable D_k, and the effect of the interaction between the jth specifically selected level of B and the kth randomly selected level of D is denoted by the random variable $(\beta D)_{jk}$. The model is then as follows:

$$Y_{ijkt} = \mu + \alpha_i + \beta_j + D_k + (\alpha\beta)_{ij} + (\beta D)_{jk} + \epsilon_{ijkt} , \qquad (17.7.24)$$
$$D_k \sim N(0, \sigma_D^2), \quad (\beta D)_{jk} \sim N(0, \sigma_{BD}^2), \quad \epsilon_{ijkt} \sim N(0, \sigma^2) ,$$
$$D_k\text{'s}, \ (\beta D)_{jk}\text{'s and}, \ \epsilon_{ijkt}\text{'s are all mutually independent,}$$
$$t = 1, \ldots, r, \ i = 1, \ldots, a, \ j = 1, \ldots, b, \ k = 1, \ldots, d.$$

The expected mean squares for the corresponding random-effects model were calculated in Example 17.6.3 and are reproduced in the second column of Table 17.8. The expected mean squares for the above mixed model are given in the third column of Table 17.8 and are obtained by collecting the terms in the expected mean squares corresponding to the fixed effects into one quadratic form.

Table 17.8 Expected mean squares for a three-way mixed model

Effect	For random-effects model	For mixed model
A	$bdr\sigma_A^2 + dr\sigma_{AB}^2 + \sigma^2$	$Q(A, AB) + \sigma^2$
B	$adr\sigma_B^2 + dr\sigma_{AB}^2 + ar\sigma_{BD}^2 + \sigma^2$	$Q(B, AB) + ar\sigma_{BD}^2 + \sigma^2$
D	$abr\sigma_D^2 + ar\sigma_{BD}^2 + \sigma^2$	$abr\sigma_D^2 + ar\sigma_{BD}^2 + \sigma^2$
AB	$dr\sigma_{AB}^2 + \sigma^2$	$Q(AB) + \sigma^2$
BD	$ar\sigma_{BD}^2 + \sigma^2$	$ar\sigma_{BD}^2 + \sigma^2$
Error	σ^2	σ^2

The expected mean squares can all be verified by direct calculation. We illustrate the calculation for B in the following example. The term $Q(B, AB)$ in $E[MSB]$ corresponds to a quadratic (i.e., squared) function of $\beta_j^* = \beta_j + \overline{(\alpha\beta)}_{.j}$, a quantity that we are used to dealing with in fixed-effects models.

Example 17.7.1 Calculation of expected mean squares

Consider an experiment with two fixed-effects treatment factors A and B, and one random-effects treatment factor D, and suppose that (17.7.24) is thought to be a reasonable model. Using rule 4 of Chap. 7, the fixed-effect sum of squares for B is

$$SSB = adr \sum_{j=1}^{b} \overline{Y}_{.j..}^2 - abdr\overline{Y}_{....}^2 .$$

Now,

$$\overline{Y}_{.j..} = \sum_{i=1}^{a} \sum_{k=1}^{d} \sum_{t=1}^{r} Y_{ijkt}/adr$$
$$= \mu + \overline{\alpha}_. + \beta_j + \overline{D}_. + \overline{(\alpha\beta)}_{.j} + \overline{(\beta D)}_{j.} + \overline{\epsilon}_{.j..} .$$

So,

$$E[\overline{Y}_{.j..}] = \mu + \overline{\alpha}_. + \beta_j + \overline{(\alpha\beta)}_{.j} \text{ and } \mathrm{Var}(\overline{Y}_{.j..}) = \frac{\sigma_D^2}{d} + \frac{\sigma_{BD}^2}{d} + \frac{\sigma^2}{adr} .$$

Similarly,

$$E[\overline{Y}_{....}] = \mu + \overline{\alpha}_. + \overline{\beta}_. + \overline{(\alpha\beta)}_{..} \text{ and } \mathrm{Var}(\overline{Y}_{....}) = \frac{\sigma_D^2}{d} + \frac{\sigma_{BD}^2}{bd} + \frac{\sigma^2}{abdr} .$$

Using the facts that $MSB = SSB/(b-1)$ and $E[X^2] = \mathrm{Var}(X) + (E[X])^2$, we obtain

$$E[MSB] = \frac{adr}{(b-1)} \sum_{j=1}^{b} \left(\frac{\sigma_D^2}{d} + \frac{\sigma_{BD}^2}{d} + \frac{\sigma^2}{adr} + [\mu + \overline{\alpha}_. + \beta_j + \overline{(\alpha\beta)}_{.j}]^2 \right)$$
$$- \frac{abdr}{(b-1)} \left(\frac{\sigma_D^2}{d} + \frac{\sigma_{BD}^2}{bd} + \frac{\sigma^2}{abdr} + [\mu + \overline{\alpha}_. + \overline{\beta}_. + \overline{(\alpha\beta)}_{..}]^2 \right)$$
$$= ar\sigma_{BD}^2 + \sigma^2 + Q(B, AB) ,$$

where

$$Q(B, AB) = \frac{adr}{(b-1)} \sum_j \left[(\beta_j + \overline{(\alpha\beta)}_{.j}) - (\overline{\beta}_. + \overline{(\alpha\beta)}_{..}) \right]^2 .$$

$\square$

Notice that in Example 17.7.1, the quadratic form $Q(B, AB)$ is equal to zero when all the $\beta_j^* = \beta_j + \overline{(\alpha\beta)}_{.j}$ are equal. We can make use of this fact when looking for an appropriate denominator for the test ratio for testing $H_0^B : \{\beta_j + \overline{(\alpha\beta)}_{.j} \text{ are all equal}\}$. If this hypothesis is true, then

Table 17.9 Test ratios for a three-way mixed model

Effect	$E[MS]$	Ratio
A	$Q(A, AB) + \sigma^2$	msA/msE
B	$Q(B, AB) + ar\sigma_{BD}^2 + \sigma^2$	$msB/ms(BD)$
D	$abr\sigma_D^2 + ar\sigma_{BD}^2 + \sigma^2$	$msD/ms(BD)$
AB	$Q(AB) + \sigma^2$	$ms(AB)/msE$
BD	$ar\sigma_{BD}^2 + \sigma^2$	$ms(BD)/msE$

$$E[MSB] = ar\sigma_{BD}^2 + \sigma^2 \,.$$

Consequently, a sensible denominator would be *MS(BD)*, which has the same expected value (see Table 17.8). Thus, the decision rule for testing H_0^B against the alternative hypothesis that the β_j^* are not all equal is

$$\text{reject } H_0^B \text{ if } \frac{msB}{ms(BD)} > F_{(b-1),(b-1)(d-1),\alpha}.$$

From Table 17.8 we can construct tests for the other relevant hypotheses in a similar manner. For example, to test the hypothesis

$$H_0^{AB} : \{(\alpha\beta)_{ij} - \overline{(\alpha\beta)}_{i.} - \overline{(\alpha\beta)}_{.j} + \overline{(\alpha\beta)}_{..} = 0 \,, \text{ for all } i, j\}$$

against the alternative hypothesis that the interaction contrasts are not all zero, the decision rule is

$$\text{reject } H_0^{AB} \text{ if } \frac{ms(AB)}{msE} > F_{(a-1)(b-1),df,\alpha} \,,$$

where *df* is the number of error degrees of freedom.

To test the hypothesis $H_0^D : \{\sigma_D^2 = 0\}$ against the alternative hypothesis $H_A^D : \{\sigma_D^2 > 0\}$, the decision rule is

$$\text{reject } H_0^D \text{ if } \frac{msD}{ms(BD)} > F_{d-1,(b-1)(d-1),\alpha}.$$

The test ratios are summarized in Table 17.9. Generally, we would not test a main-effect or interaction hypothesis unless all higher-order interactions involving these same factors were believed to be negligible. For some mixed models, as for random-effects models, the appropriate denominator for the test statistic may not be listed among the expected mean squares for the factors in the model. In this case, it would be necessary to calculate it, and the corresponding degrees of freedom, using (17.6.22), p. 637.

17.7.2 Confidence Intervals in Mixed Models

Confidence Intervals for Fixed Effects

For fixed effects in a mixed model with equal sample sizes, we may use all of the rules of Sect. 7.3, p. 209, exactly as if there were no random effects in the model, *except that we replace msE by the same mean square that was identified for hypothesis testing*—namely, used in the denominator of the test ratio—and the error degrees of freedom are also replaced. The necessity of doing this replacement

is highlighted in Example 17.7.2. Apart from this, we may use the Bonferroni, Scheffé, Tukey, and Dunnett methods of multiple comparisons in the usual way. When the sample sizes are unequal, computing least squares estimates and appropriate standard errors is more complicated. Appropriate methods will be illustrated in Chaps. 18 and 19 using PROC MIXED in SAS software and using the lmer and lsmeans functions in R.

Example 17.7.2 `Calculation of confidence intervals`

Consider an experiment with two fixed-effects treatment factors and a third treatment factor with random effects, for which the following model is thought to be reasonable (this is the same model that has been discussed throughout this subsection):

$$Y_{ijkt} = \mu + \alpha_i + \beta_j + D_k + (\alpha\beta)_{ij} + (\beta D)_{jk} + \epsilon_{ijkt}.$$

The fixed part of the model is

$$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

which looks exactly like one of the two-way analysis of variance models that was studied in Chap. 6.

Suppose we need confidence intervals for pairwise comparisons in the levels of A and of B. Then, as usual, the least squares estimates for pairwise differences are

$$\widehat{\alpha}_i^* - \widehat{\alpha}_p^* = \left(\widehat{\alpha}_i + \widehat{(\alpha\beta)}_{i.}\right) - \left(\widehat{\alpha}_p - \widehat{(\alpha\beta)}_{p.}\right) = \bar{y}_{i...} - \bar{y}_{p...}$$

and

$$\widehat{\beta}_j^* - \widehat{\beta}_u^* = \left(\widehat{\beta}_j + \widehat{(\alpha\beta)}_{.j}\right) - \left(\widehat{\beta}_u + \widehat{(\alpha\beta)}_{.u}\right) = \bar{y}_{.j..} - \bar{y}_{.u..}.$$

Tables 17.8 (p. 643) and 17.9 (p. 645) suggest that *msE* should be used in the formulae for confidence intervals for $\alpha_i^* - \alpha_p^*$, as usual, but that *ms(BD)* should be used in place of *msE* in the formulae for confidence intervals for $\beta_j^* - \beta_u^*$. All confidence intervals are of the form

$$(\text{least squares estimate}) \pm (w) \times (\text{standard error}).$$

The standard error is the square root of the estimated variance of the least squares estimator. Now,

$$\text{Var}(Y_{ijkt}) = \sigma_D^2 + \sigma_{BD}^2 + \sigma^2 \quad \text{and} \quad \text{Var}(\bar{Y}_{.j..}) = \frac{\sigma_D^2}{d} + \frac{\sigma_{BD}^2}{d} + \frac{\sigma^2}{adr}.$$

The Y_{ijkt}'s are not independent. Observations on the same level of D are correlated. If two observations are taken on the same levels of B and D, we have

$$\text{Cov}(Y_{ijkt}, Y_{pjks}) = \sigma_D^2 + \sigma_{BD}^2.$$

If two observations are taken on the same level of D, but different levels of B, then

$$\text{Cov}(Y_{ijkt}, Y_{puks}) = \sigma_D^2.$$

All other pairs of response variables are independent. Consequently,

$$\text{Cov}(\overline{Y}_{.j..}, \ \overline{Y}_{.u..}) = \frac{1}{a^2 d^2 r^2}\left[\sum_{i=1}^{a}\sum_{p=1}^{a}\sum_{k=1}^{d}\sum_{t=1}^{r}\sum_{s=1}^{r}\text{Cov}(Y_{ijkt}, Y_{puks})\right]$$

$$= \frac{1}{a^2 d^2 r^2}\left[a^2 d r^2 \sigma_D^2\right]$$

$$= \frac{\sigma_D^2}{d},$$

and

$$\text{Var}(\overline{Y}_{.j..} - \overline{Y}_{.u..}) = \text{Var}(\overline{Y}_{.j..}) + \text{Var}(\overline{Y}_{.u..}) - 2\text{Cov}(\overline{Y}_{.j..}, \ \overline{Y}_{.u..})$$

$$= 2\left(\frac{\sigma_D^2}{d} + \frac{\sigma_{BD}^2}{d} + \frac{\sigma^2}{adr}\right) - 2\left(\frac{\sigma_D^2}{d}\right)$$

$$= \frac{2}{adr}\left(ar\sigma_{BD}^2 + \sigma^2\right),$$

which is of the form $(\Sigma c_i^2/(adr))(ar\sigma_{BD}^2 + \sigma^2)$. Thus, we need to estimate $(ar\sigma_{BD}^2 + \sigma^2)$ rather than σ^2, and an unbiased estimate is given by $ms(BD)$. So, the standard error for $\hat{\beta}_j^* - \hat{\beta}_u^* = \overline{Y}_{.j..} - \overline{Y}_{.u..}$ is $((2/(adr))\,ms(BD))^{1/2}$ with corresponding degrees of freedom $(b-1)(d-1)$.   □

In some models, the necessary mean square will not be listed in the expected mean squares table, and (17.6.22), p. 637, and methods discussed there will need to be used to find an approximate mean square and degrees of freedom.

Confidence Intervals for Variance Components

In obtaining confidence intervals for variance components, only the random part of the model is used, or, equivalently, only the mean squares corresponding to random effects. Consequently, the formulae of Sect. 17.6.5 are used exactly as described for random-effects models.

17.8   Rules for Analysis of Random-Effects and Mixed Models

Rules 1–7 of Sect. 7.3, p. 209, are valid for calculating degrees of freedom, sums of squares, and mean squares in random-effects and mixed models as well as in fixed-effects models. In addition, rules 8–16 are valid for analyzing fixed effects, except that σ^2 and msE may need to be replaced. Rules 17–22 below summarize the results of this chapter. Rule 17 is an expanded version of rule 17 on p. 637.

17.8.1   Rules—Equal Sample Sizes

17. To obtain the expected mean square for a particular main effect or interaction, first make a note of the subscripts on the term representing that particular effect in the model. Write down variance components for the effect of interest, for the error, and for every interaction whose term in the

model includes the noted set of subscripts. Gather up all variance components corresponding to fixed effects into one quadratic form Q. Multiply any remaining variance component except σ^2 by the number of observations taken on each level or combination of levels of the corresponding effect (main effect or interaction). Add up the terms.

18. To obtain the denominator of the test statistic for testing the null hypothesis that a main effect or interaction effect is zero, write down the expected mean square for the effect of interest (see rule 17). Cross out the term that would be zero if the null hypothesis were true. The denominator of the test statistic is the mean square, or linear combination of mean squares, u, whose expected value is equal to the remaining expression.

19. For a random effect, let $U = \Sigma k_i MS_i$ be the mean square or linear combination of mean squares whose expected value is equal to the variance component corresponding to the random effect. An exact or approximate $100(1 - \alpha)\%$ confidence interval for this variance component is

$$\left(\frac{xu}{\chi^2_{x,\alpha/2}} \ , \ \frac{xu}{\chi^2_{x,1-\alpha/2}} \right),$$

where

$$x = \frac{[\Sigma k_i(ms_i)]^2}{\Sigma [k_i(ms_i)]^2/x_i}$$

and where u is the observed value of U, ms_i is the observed value of MS_i, and x_i is the number of degrees of freedom corresponding to ms_i.

20. For a fixed effect, confidence intervals are obtained as in rule 14, p. 211, except that msE is replaced by the denominator u from rule 18, and the number of error degrees of freedom is replaced by x in rule 19.

21. For a fixed effect, the decision rule for testing the hypothesis that the effect is zero is the same as that in rule 8, p. 210, for fixed-effects models, except that msE is replaced by the denominator u from rule 18, and the number of error degrees of freedom is replaced by x in rule 19.

22. For a random effect, the decision rule for testing the hypothesis H_0 that the corresponding variance component is zero against the alternative hypothesis that it is not zero is

$$\text{reject } H_0 \text{ if } \frac{ms}{u} > F_{\nu,x,\alpha} \,,$$

where ms is the mean square for the effect of interest and ν the corresponding degrees of freedom, u is the observed value of the denominator as in rule 18, and x is the corresponding degrees of freedom calculated as in rule 19.

17.8.2   Controversy (Optional)

Before proceeding, we should mention that some other textbooks may present slightly different tables of expected mean squares. For example, the expected mean square for D in Table 17.8, which we have calculated as

$$E[MSD] = abr\sigma_D^2 + ar\sigma_{BD}^2 + \sigma^2,$$

may in other texts be listed as

$$E[MSD] = abr\sigma_D^2 + \sigma^2 \,.$$

This alternative listing occurs when constraints are placed on the model parameters involving fixed-effects factors, and it suggests use of the denominator msE rather than $ms(BD)$ in testing $H_0^D : \{\sigma_D^2 = 0\}$ against $H_A^D : \{\sigma_D^2 > 0\}$. A number of articles in the statistical literature have been written advocating one denominator rather than the other, and there still appears to be no consensus.

If we follow the line of reasoning that we have followed to this point, that normally we will examine main effects only when there is no interaction, then some of the controversy disappears. If σ_{BD}^2 is really zero, then $E[MSD] = abr\sigma_D^2 + \sigma^2$ in both cases. Of course, due to variability of the data and uncertainty about whether or not σ_{BD}^2 is really zero (or close to it), we still have to make the choice in practice. We have recommended using $ms(BD)$ as the denominator if the objective is to test $H_0^D : \sigma_D^2 = 0$. However, if interest is really in testing

$$H_0^{D+BD} : \{\sigma_D^2 + b^{-1}\sigma_{BD}^2 = 0\},$$

or equivalently

$$H_0^{D+BD} : \{\sigma_D^2 = \sigma_{BD}^2 = 0\},$$

then we would use msE as the denominator.

The controversy originally arose from the formulation of the model. In our example, the model was given in (17.7.24), p. 643, and the controversy surrounds the random effect $(\beta D)_{jk}$. We have modeled this as a normally distributed random variable. Some authors add to the model the restriction $\Sigma_j (\beta D)_{jk} = 0$, and this leads to the canceling of the term in σ_{BD}^2 when the expected mean square of D is calculated.

Hocking (1996, p. 569) shows that under this restriction, the hypothesis H_0^D is actually our hypothesis H_0^{D+BD}. An explanation for this is as follows. If constraints are placed on the parameters, then the $(\beta D)_{jk}$ effects truly represent interaction effects, and σ_{BD}^2 measures precisely variability in BD-interaction effects. However, if no constraints are placed on the parameters, then σ_{BD}^2 being positive implies the presence of main effects of B and D as well as the presence of BD-interaction effects. In other words, the parameters $(\beta D)_{jk}$ represent "BD effects" in model (17.7.24), though we have referred to them as BD-interaction effects. Thus, under our model (17.7.24), the hypothesis $H_0^{D+BD} : \{\sigma_D^2 = \sigma_{BD}^2 = 0\}$ is that there are no main effects of D (or BD interactions). Also, there are no BD interactions if $\sigma_{BD}^2 = 0$, and there are no main effects of B (or BD interactions) if $\beta_1 = \beta_2 = \cdots = \beta_b$ and $\sigma_{BD}^2 = 0$. From this viewpoint, the hypothesis $H_0^D : \sigma_D^2 = 0$ is that there are no main effects of D if σ_{BD}^2 is believed to be zero; otherwise, it is the hypothesis that main effects of D are no less negligible than BD interactions.

Since there are problems inherent in placing restrictions on the model parameters, we prefer not to do so, and we prefer to use the set of expected mean squares in Table 17.7. If the parameters in the model are properly interpreted, then there is no controversy, and the appropriate test is determined by what is most sensible for the experiment at hand.

17.9   Block Designs and Random Block Effects

In certain types of experiments, it is extremely common for the levels of a blocking factor to be randomly selected. For example, in medical, psychological, educational, or pharmaceutical experiments, blocks frequently represent subjects that have been selected at random from a large population of similar subjects. In agricultural experiments, blocks may represent different fields selected from a large variable population of fields. In industrial experiments, different machine operators may represent different levels of the blocking factor and may be similar to a random sample from a large population of possible operators. Raw material may be delivered to the factory in batches, a random selection of which are used as blocks in the experiment.

Since we are not interested in the blocking factor itself, its designation as random rather than fixed will affect the analysis only if the model includes a block×treatment interaction. For example, suppose that factor D in Table 17.8 represents a random-effects blocking factor, and that A and B are two fixed-effects treatment factors. The analysis of factor A, which has no interaction with D, is unaffected by the designation of D as a random effect. However, the analysis of factor B, which interacts with blocks, *is* affected, since *msE* in hypothesis tests and confidence intervals for contrasts in the levels of B will be replaced by *ms(BD)*.

Example 17.9.1 `Temperature experiment`

The temperature experiment was run by M. Bowe, J. Cooper, J. Donato, S. Giust, and H. Schieman in 1994 to compare the times required for three different digital thermometers (factor A at $a = 3$ levels) to register body temperature at two different sites—in the mouth and under the arm—(factor B at $b = 2$ levels). Thus, there were six treatment combinations. Four subjects were selected at random from the American statistics graduate students at The Ohio State University, and each treatment combination was measured once for each subject. The experiment was designed as a randomized complete block design, with subjects representing blocks. The recorded times are shown in Table 17.10.

The four subjects used in the experiment are not themselves of interest. Of more interest is how the thermometers react on average over a large population of subjects. The population of American statistics graduate students at the university is large, but not infinite. However, the four subjects used in the experiment are, hopefully, representative of all possible American graduate students, and it is reasonable to model the subject (block) effect as a random effect.

Since subjects vary in body heat, it is possible that factor B (site) might interact with subject. It is also possible that different thermometers might act differently at the two different sites. Consequently the following model might be reasonable for this experiment.

$$Y_{hij} = \mu + S_h + \alpha_i + \beta_j + (\alpha\beta)_{ij} + (S\beta)_{hj} + \epsilon_{hij} , \qquad (17.9.25)$$
$$h = 1, 2, 3, 4, \quad i = 1, 2, 3, \quad j = 1, 2,$$
$$S_h \sim N(0, \sigma_S^2), \quad (S\beta)_{hj} \sim N(0, \sigma_{SB}^2), \quad \epsilon_{hij} \sim N(0, \sigma^2) ,$$
$$S_h\text{'s}, \ (S\beta)_{hj}\text{'s and } \epsilon_{hij}\text{'s are all mutually independent} ,$$

$$(17.9.26)$$

where all random variables on the right-hand side of the model are mutually independent, and where S_h represents the effect of the hth randomly selected subject (block), α_i represents the effect of the ith specifically selected thermometer, and β_j represents the effect of the ith specifically selected site. This model is similar to mixed model (17.7.24) with S_h replacing D_k. Consequently, the expected mean squares will be similar to those in Table 17.8, p. 643. The analysis of variance table is shown in Table 17.11.

Table 17.10 Data (in seconds) for the temperature experiment

Subject	Treatment combination					
	11	12	21	22	31	32
1	62.16	61.53	154.42	310.46	95.98	225.65
2	65.63	63.70	132.30	284.64	98.50	241.63
3	63.12	61.34	105.52	315.61	110.05	364.07
4	61.51	61.54	94.88	294.16	107.93	304.58

Table 17.11 Analysis of variance table for the mixed model temperature experiment

Source of variation	Degrees of freedom	Mean square	p-value	Expected mean square
Subject (block)	3	570.04	–	–
Thermometer (A)	2	52879.34	0.0001	$Q(A, AB) + \sigma^2$
Site (B)	1	86029.60	0.0035	$Q(B, AB) + 3\sigma_{SB}^2 + \sigma^2$
Therm*Site (AB)	2	21897.23	0.0001	$Q(AB) + \sigma^2$
Subject*Site (SB)	3	1210.67	0.2625	$3\sigma_{SB}^2 + \sigma^2$
Error	12	802.57		σ^2
Total	23			

We start by testing the two interaction hypotheses. To test the hypothesis $H_0^{SB} : \{\sigma_{SB}^2 = 0\}$, that the subject by site interaction variance is negligible, against the alternative hypothesis that it is not negligible, using a significance level of 0.01 (so that the overall significance level will be at most 0.05), we

$$\text{reject } H_0^{SB} \text{ if } \frac{ms(SB)}{msE} > F_{3,12,0.01} = 5.95 .$$

Since $ms(SB)/msE = 1.51$, there is not sufficient evidence to conclude that the interaction variance is greater than zero (equivalently, the p-value is greater than 0.01). Before we can examine the site main effect, however, we also need to look at the thermometer by site interaction.

To test the hypothesis

$$H_0^{AB} : \{(\alpha\beta)_{ij} - (\alpha\beta)_{ip} - (\alpha\beta)_{uj} + (\alpha\beta)_{up}, \text{ for all } i, j, u, p\}$$

against the alternative hypothesis that the interaction is not negligible, we

$$\text{reject } H_0^{AB} \text{ if } \frac{ms(AB)}{msE} > F_{2,12,0.01} = 6.93 .$$

Since $ms(AB)/msE = 27.28$, we reject H_0^{AB} and conclude that there is a thermometer$\times$site interaction. Thus, it is unlikely that the thermometer and site main effects are of interest. However, for illustration purposes, we ask whether the *average* time taken for these three digital thermometers to register is the same whether used in the mouth or under the arm. Thus, we will test the hypothesis

$$H_0^B : \{\beta_1 + (\overline{\alpha\beta})_{.1} = \beta_2 + (\overline{\alpha\beta})_{.2}\} .$$

To test this hypothesis at significance level 0.01, we

$$\text{reject } H_0^B \text{ if } \frac{msB}{ms(SB)} > F_{1,3,0.01} = 29.5 .$$

Since $msB/ms(SB) = 71.06$, we reject H_0^B and conclude that it does make a difference in registering temperature (on average for these three thermometers) as to whether the thermometer is used in the mouth or under the arm. This conclusion is made on average over the three thermometers and over the whole population of similar graduate students. □

17.10   Using SAS Software

Section 17.10.1 illustrates the use of SAS software to check model assumptions on random effects. Then in Sect. 17.10.2, the analysis of mixed models using `PROC GLM` and `PROC MIXED` is illustrated, followed by an example of analysis of covariance to deal with a quadratic time trend. Finally in Sect. 17.10.3, the SAS `DATA` step and functions are used to do the sample size calculations of Example 17.4.1.

17.10.1   Checking Assumptions on the Model

Using the data of Table 17.1, p. 617, for the clean wool experiment, we illustrate some methods of checking model assumptions for a random-effects one-way model. The experimenters took observations on $r = 4$ cores of wool from each of $v = 7$ randomly selected wool bales.

We let the random variable T_i represent the true clean content of the ith randomly selected bale of wool from the shipment, and let $Y_{it} = T_i + \epsilon_{it}$ represent the observed clean content of the tth core (observation) from the ith bale, where the error variable ϵ_{it} includes the deviation from the true average clean content of the tth core from the ith bale, the measurement error, environmental conditions, etc.

First, we check the error assumptions by calculating and plotting the standardized residuals obtained as though the bale effects were fixed. The standardized residuals are calculated in the usual way and plotted against the levels of the treatment factor and the predicted values (see Sect. 5.8, p. 119). The latter plot, obtained by `PROC SGPLOT`, is shown in Fig. 17.3.

The most noticeable feature is that bale 1 gives rise to one very large standardized residual (an outlier). This means one of several things: Perhaps the data value is in error, so that this value is an outlier, or perhaps bale 1 is extremely more variable than the other bales in the population, or perhaps the error variables are not normally distributed. Let us suppose that we could go back to the original experimenters and that indeed, something unusual happened at this point during the time at which the observations were taken. If so, we could exclude this value. The new residual plot is shown in Fig. 17.4.

All standardized residuals now lie within the expected range for normally distributed errors. The plot gives us quite a lot of information about our sample of bales and possibly about the shipment of bales from which they were drawn. First, the average clean content of bale 1 is around 53, considerably

Fig. 17.3 Residuals versus predicted values by bale type for the clean wool experiment
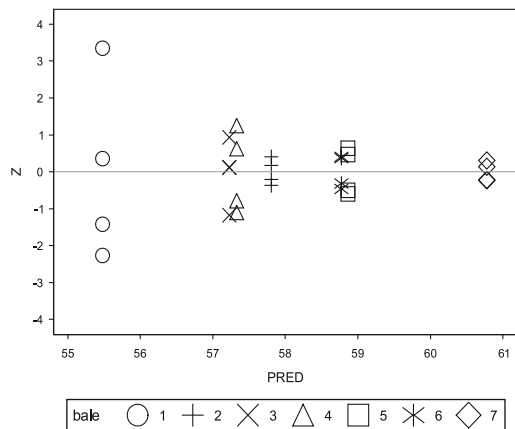
Fig. 17.4 Residuals versus predicted values for the clean wool experiment, excluding the outlier
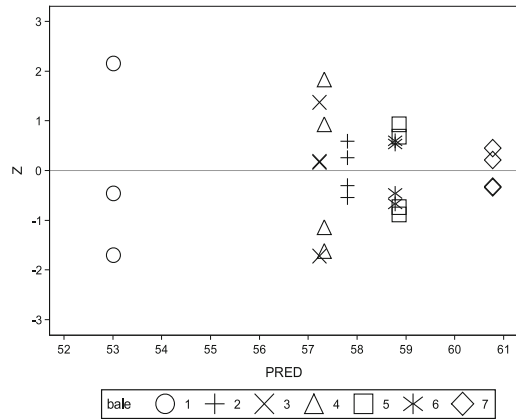

Table 17.12 SAS program to plot standardized treatment averages against their normal scores

```
PROC SORT; BY BALE;
PROC MEANS NOPRINT; BY BALE;
  VAR CONTENT;
  OUTPUT OUT = WOOL2 MEAN = AVCONT;
PROC STANDARD STD = 1.0 MEAN = 0.0;
  VAR AVCONT;
PROC RANK NORMAL = BLOM;
  VAR AVCONT;
  RANKS NSCORE;
PROC SGPLOT;
  SCATTER X = NSCORE Y = AVCONT / GROUP = BALE;
```

below the others. This was the bale that had the supposed outlier. One might suspect that this bale either did not come from the same shipment or was contaminated at some point before being measured. On the other hand, the shipment may contain a number of "rogue bales," and this ought to be investigated. At the other end of the range, we see that bale 7 had the highest clean content and was least variable. Perhaps this is not too surprising, since a bale with 100% clean content would probably show no variability in the measurements taken on it. Thus, one might suspect that our model that includes normally distributed errors is not ideal for this situation. However, the plot of standardized residuals against normal scores does not show any anomalies (figure not shown).

In a one-way random-effects model, we can check the assumption that the treatment effects have a normal distribution by making a normal probability plot of the standardized treatment averages $\overline{Y}_{i.}$ against their normal scores. (This cannot be done for models with more than one random effect, since the treatment averages are not independent.) For the clean wool experiment, the normal probability plot is obtained by means of the statements in Table 17.12, and the resulting plot is shown in Fig. 17.5. If the normality assumption for the population of bales is satisfied, the standardized bale averages should roughly lie along a line (with slope 1.0) through (0, 0). In Fig. 17.5, we see that this is roughly the case.

In summary, the random-effects one-way model with the standard distribution assumptions does not fit these data too well, since variances apparently are not constant or there is an outlier. Nevertheless, we have established that the population of bales in this shipment is extremely variable. Selected bales 1 and 7 appear to be somewhat different from the other five selected bales. Perhaps the shipment is

Fig. 17.5 Normal
probability plot of the
standardized treatment
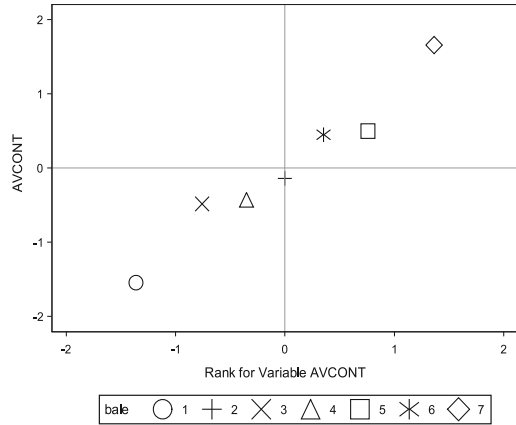averages for the clean wool
experiment

Table 17.13   SAS program for the temperature experiment

```
DATA TEMPR;
  INPUT THERM SITE SUBJ TIME;
  LINES;
  1 1 1  62.16
  1 2 1  61.53
  : : :    :
  3 2 4 304.58
;
PROC GLM;
  ODS EXCLUDE LSMeanCL;
  CLASS THERM SITE SUBJ;
  MODEL TIME = SUBJ THERM SITE THERM*SITE SUBJ*SITE;
  RANDOM SUBJ SUBJ*SITE / TEST;
  CONTRAST 'SITE1-SITE2' SITE 1 -1 / E = SUBJ*SITE;
  LSMEANS SITE / CL PDIFF E = SUBJ*SITE;
```

made up of dissimilar subpopulations (perhaps from different sources). This should be checked, since it may give a clue as to how to improve the wool clean content in the future.

17.10.2   Estimation and Hypothesis Testing

PROC GLM

Analysis of variance tables for random-effects and mixed models are obtained using PROC GLM in exactly the same way as for fixed-effects models. The additional expected mean squares column can be obtained very easily by inserting a RANDOM statement immediately after the model statement. All random effects should be listed in the RANDOM statement, as shown, for example, in Table 17.13 for the temperature experiment of Example 17.9.1, p. 650. The denominators, calculated as explained throughout this chapter, can be obtained by adding the option TEST to the RANDOM statement, as shown in the following example. The actual denominators are printed out as well as the p-values.

The output is shown in Fig. 17.6. The first few lines reproduce the expected mean squares that were calculated by hand in Table 17.7, p. 641. The remainder of the output gives the TYPE III sums of

squares, but instead of calculating the usual test ratios with *msE* as the denominator, the TEST option on the RANDOM statement has caused the denominator *ms*(subj×site) to be used where appropriate.

All pairwise comparisons of fixed effects can be estimated using the LSMEANS statement, which allows the correct variance estimator to be specified. For example, the statement

```
LSMEANS SITE / CL PDIFF E = SUBJ*SITE;
```

will use the subj×site interaction mean square as the variance estimate for comparing sites, rather than the error mean square. The LSMEANS statement provides a confidence interval for the pairwise comparison of sites, and also a *p*-value for testing equality of the two site effects. The ODS statement is used to exclude printing of the confidence intervals (limits) for site level means, since the standard errors would be incorrectly estimated using *ms*(subj×site); the reader is asked in Exercise 9 to verify that $\text{Var}(\overline{Y}_{.j}) = (3\sigma_S^2 + 3\sigma_{SB}^2 + \sigma^2)/12$ for the *j*th site under model (17.9.25), p. 650.

Any contrast for the fixed effects can be estimated as usual using the ESTIMATE statement, though the standard error is computed using *MSE* whether or not this is appropriate. Confidence intervals can be calculated by hand and the mean squared error replaced by the denominator used in the test procedures if necessary. For testing individual contrasts, the CONTRAST statement can be used and the required denominator can be specified. For example, the statement

```
CONTRAST 'SITE1-SITE2' SITE 1 -1 / E = SUBJ*SITE;
```

will use the subj×site interaction mean square as the variance estimate for comparing sites as appropriate.

The deficiencies of PROC GLM in computing standard errors are overcome by PROC MIXED, introduced next.
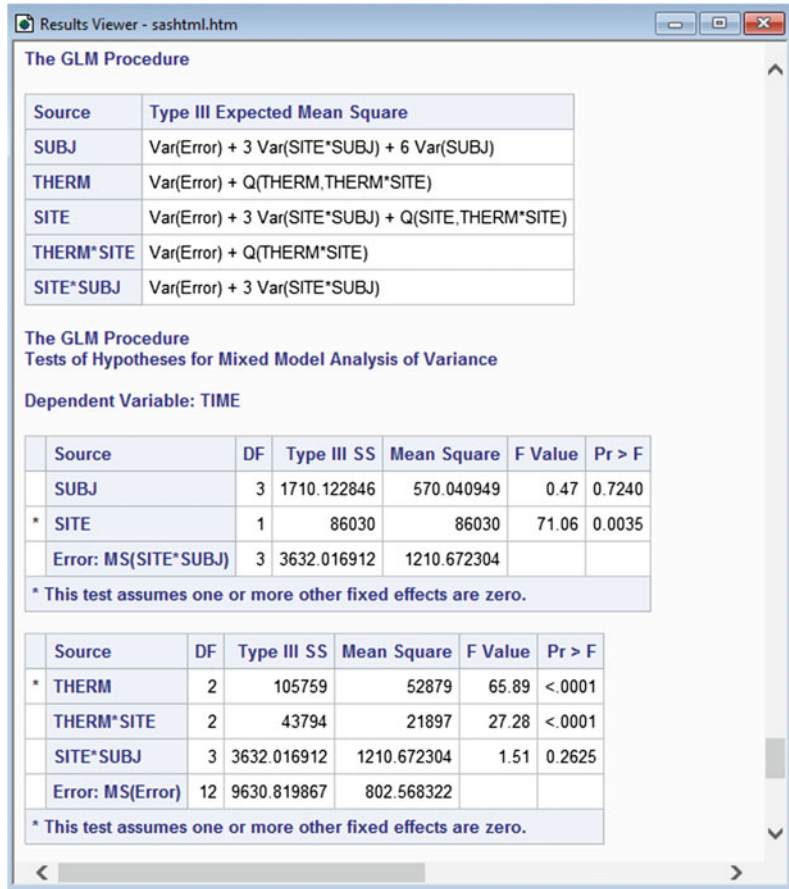
PROC MIXED

The SAS package includes an alternative procedure PROC MIXED, explicitly designed to cope with random-effects and mixed models. The statements that generate the same set of information as in Fig. 17.6 are

```
PROC MIXED METHOD = TYPE3;
  CLASS THERM SITE SUBJ;
  MODEL TIME = THERM SITE THERM*SITE / DDFM = SAT;
  RANDOM SUBJ SUBJ*SITE;
  LSMEANS SITE / CL DIFF;
```

For PROC MIXED, the MODEL statement only contains fixed effects, with random effects specified in the RANDOM statement. The option METHOD=TYPE3 causes the model to be fit by the method of least squares, as used by PROC GLM. Estimates of the variance components are also calculated by this procedure and, under the option METHOD=TYPE3, they match those one could compute by hand from GLM output. An important advantage of PROC MIXED is that standard errors of means and contrasts are automatically correctly estimated under mixed models, even if composite estimates are needed, as is the case here for site treatment means, for example. The DDFM=SAT option in the MODEL statement causes Satterthwaite's approximation to be used to compute degrees of freedom associated with any composite variance estimates, as may be needed for either *F*-statistic denominators or standard error estimates.

Throughout this chapter we have discussed the analysis of mixed and random effects models using the *analysis of variance approach*—namely, fitting the model by least squares as if all effects were fixed, obtaining a corresponding analysis of variance table, then using the expected mean squares to obtain unbiased estimates of the variance components and to determine appropriate *F*-statistic denominators and standard error estimates. The variance component estimates so obtained are called *analysis of variance estimates*. For balanced data under normality, the least squares estimates of any

Fig. 17.6 SAS software
analysis of variance for the
temperature experiment

The GLM Procedure

Source	Type III Expected Mean Square
SUBJ	Var(Error) + 3 Var(SITE*SUBJ) + 6 Var(SUBJ)
THERM	Var(Error) + Q(THERM,THERM*SITE)
SITE	Var(Error) + 3 Var(SITE*SUBJ) + Q(SITE,THERM*SITE)
THERM*SITE	Var(Error) + Q(THERM*SITE)
SITE*SUBJ	Var(Error) + 3 Var(SITE*SUBJ)

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: TIME

	Source	DF	Type III SS	Mean Square	F Value	Pr > F
	SUBJ	3	1710.122846	570.040949	0.47	0.7240
*	SITE	1	86030	86030	71.06	0.0035
	Error: MS(SITE*SUBJ)	3	3632.016912	1210.672304		

* This test assumes one or more other fixed effects are zero.

	Source	DF	Type III SS	Mean Square	F Value	Pr > F
*	THERM	2	105759	52879	65.89	<.0001
	THERM*SITE	2	43794	21897	27.28	<.0001
	SITE*SUBJ	3	3632.016912	1210.672304	1.51	0.2625
	Error: MS(Error)	12	9630.819867	802.568322		

* This test assumes one or more other fixed effects are zero.

estimable fixed effects are best linear unbiased estimates, and the analysis of variance estimates of variance components are minimum variance unbiased estimates, but the variance component estimates can be negative, as happens in the above example where the subjects variance component estimate is $\hat{\sigma}_S^2 = [ms(\text{SUBJ}) - ms(\text{SITE*SUBJ})]/6 \approx -106.77$.

There are other more sophisticated statistical methods for estimating variance components that prevent the estimates from ever being negative, and that are generally preferable for unbalanced data. One such approach involves estimation of the variance components by *restricted maximum likelihood* (ReML). This approach, implemented in PROC MIXED, will be discussed in Chap. 19.

Covariates

Before leaving this section, we will examine a more complicated model. The plot of standardized residuals against order of observation by flavor for the ice cream experiment (Example 17.3.1, p. 621) is shown in Fig. 17.7. This plot suggests that there may be a quadratic time trend in the data.

We define two extra variables X and X2 in the DATA statement as follows,

```
DATA ICE;
   INPUT FLAVOR MELTTIME ORDER;
   X=ORDER-16.5;
   X2=X*X;
   LINES;
   1 924 1
   :  :  :
```

and we add these variables to the model statement, so the code for `PROC GLM` becomes

```
PROC GLM;
  CLASS FLAVOR;
  MODEL MELTTIME = X X2 FLAVOR;
  RANDOM FLAVOR / TEST;
```

The variable `X` is just the same as `ORDER`, except that we have subtracted the average order 16.5. This helps to reduce computational problems in the model fitting. The Type III sums of squares and the expected mean squares are shown in Fig. 17.8. We see that the quadratic effect of time order is quite substantial and that from the list of expected mean squares, our estimate of the variance of melting times due to flavor (`var(FLAVOR)`) must be calculated as

$$\hat{\sigma}_T^2 = \frac{94179.139 - 4497.426}{9.6478} = 9359.62 \text{ seconds}^2$$

or $\hat{\sigma}_T = 96.75$ seconds, which is a little larger than the estimate of $\hat{\sigma}_T = 85.13$ seconds that we obtained in Example 17.3.4, p. 627. Examination of the residuals in the new model shows that the error assumptions are fairly well satisfied. In Exercise 8, the reader is asked to recalculate the confidence intervals for σ_T^2 and σ_T^2/σ^2 using the new model.

17.10.3   Sample Size Calculations

In this section, a program in the SAS `DATA` step is used to do the sample size calculations of Example 17.4.1, p. 630, for the ice cream experiment. Recall, in testing the hypothesis $H_0^{\gamma T} : \sigma_T^2 \le \sigma^2$ against the hypothesis $H_A^{\gamma T} : \sigma_T^2 > \sigma^2$ at a significance level of $\alpha = 0.05$, the goal is to be able to reject the null hypothesis with probability $\pi = 0.95$ if the true value of σ_T^2/σ^2 is at least $\Delta = 2.0$. In Example 17.4.1, trial and error was used to determine the number v of ice cream flavors we should look at if we took $r = 11$ or $r = 3$ observations on each. Recall, to achieve the desired power for given r, we need to find v such that

$$(F_{v-1,v(r-1),.05})(F_{v(r-1),v-1,.05}) \le (2r+1)/(r+1).$$

The calculations are illustrated in the SAS program in Table 17.14, and the corresponding output is displayed in Fig. 17.9. In the SAS program, for each value $r = 11$ and $r = 3$, the SAS function `FINV`

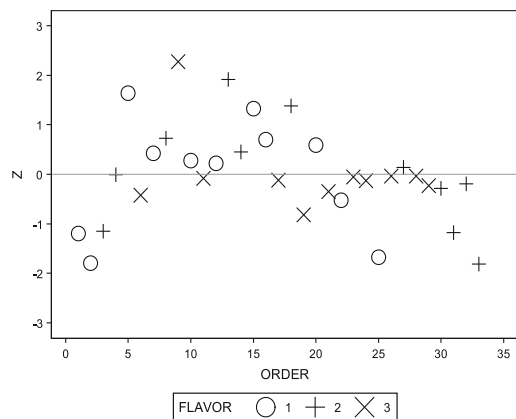**Fig. 17.7** Plot of the standardized residuals against order of observation by flavor for the ice cream experiment

Fig. 17.8 SAS software
analysis of variance for the
ice cream experiment

Table 17.14 SAS program doing sample size calculations for the ice cream experiment

```
DATA POWER;
  ALPHA=0.05; POWER=0.95;
  DO R=11,3;
    RATIO = (2*R+1)/(R+1);
    RESULT="Need more data";
    V=1;
    DO WHILE (RESULT="Need more data" and V < 201);
      V=V+1;
      DF1=V-1; DF2=V*(R-1);
      F1 = FINV(1-ALPHA,DF1,DF2); * Compute F(DF1,DF2,ALPHA);
      F2 = FINV(POWER,DF2,DF1);   * Compute F(DF2,DF1,1-POWER);
      PRODUCT=F1*F2;
      IF PRODUCT < RATIO THEN RESULT="Enough data";
    END;
    OUTPUT; * Output results to data set;
  END; * End R loop;
;
PROC PRINT;
  VAR R V RESULT POWER ALPHA F1 F2 PRODUCT RATIO DF1 DF2;
```

is used to compute the quantile values $F_{v-1,v(r-1),.05}$ and $F_{v(r-1),v-1,.05}$ for each value $v = 2, 3, \ldots,$
either stopping when the above inequality is satisfied, giving the required value of v and the result
"Enough data", or reaching $v = 200$ and giving the result "Need more data". The results displayed in
Fig. 17.9 are a bit more precise than those given in Example 17.4.1.

Fig. 17.9 SAS software sample size calculations for the ice cream experiment

17.11   Using R Software

Section 17.11.1 illustrates the use of R to check model assumptions on random effects. Then in Sect. 17.11.2, the analysis of fixed effects in mixed models using $\texttt{aov}$ is illustrated. Finally in Sect. 17.11.3, an R function is defined to do the sample size calculations of Example 17.4.1.

17.11.1   Checking Assumptions on the Model

Using the data of Table 17.1, p. 617, for the clean wool experiment, we illustrate some methods of checking model assumptions for a random-effects one-way model. The experimenters took observations on $r = 4$ cores of wool from each of $v = 7$ randomly selected wool bales.

We let the random variable T_i represent the true clean content of the ith randomly selected bale of wool from the shipment, and let $Y_{it} = T_i + \epsilon_{it}$ represent the observed clean content of the tth core (observation) from the ith bale, where the error variable ϵ_{it} includes the deviation from the true average clean content of the tth core from the ith bale, the measurement error, environmental conditions, etc.

First, we check the error assumptions by calculating and plotting the standardized residuals obtained as though the bale effects were fixed. The standardized residuals are calculated in the usual way and plotted against the levels of the treatment factor and the predicted values (see Sect. 5.9, p. 126). The latter plot, obtained from the statements

```
plot(z ~ ypred, data=wool.data, ylab="Standardized Residuals", las=1,
     type="n") # Suppress plotting of circles
text(z ~ ypred, bale, cex=0.75, data=wool.data) # Plot bale number
mtext("Plotting symbol is bale", side=3, adj=1, line=1) # Margin text
abline(h=0)
```

is shown in Fig. 17.10. The plotted symbols indicate the bales from which the residuals arose.

The most noticeable feature is that bale 1 gives rise to one very large standardized residual (an outlier). This means one of several things: Perhaps the data value is in error, so that this value is an outlier, or perhaps bale 1 is extremely more variable than the other bales in the population, or perhaps the error variables are not normally distributed. Let us suppose that we could go back to the original experimenters and that indeed, something unusual happened at this point during the time at which the observations were taken. If so, we could exclude this value. The new residual plot is shown in Fig. 17.11.

All standardized residuals now lie within the expected range for normally distributed errors. The plot gives us quite a lot of information about our sample of bales and possibly about the shipment of bales from which they were drawn. First, the average clean content of bale 1 is around 53, considerably

below the others. This was the bale that had the supposed outlier. One might suspect that this bale either did not come from the same shipment or was contaminated at some point before being measured. On the other hand, the shipment may contain a number of "rogue bales," and this ought to be investigated. At the other end of the range, we see that bale 7 had the highest clean content and was least variable. Perhaps this is not too surprising, since a bale with 100% clean content would probably show no variability in the measurements taken on it. Thus, one might suspect that our model that includes normally distributed errors is not ideal for this situation. However, the plot of standardized residuals against normal scores does not show any anomalies (figure not shown).

In a one-way random-effects model, we can check the assumption that the treatment effects have a normal distribution by making a normal probability plot of the standardized treatment averages $\overline{Y}_{i.}$ against their normal scores. (This cannot be done for models with more than one random effect, since the treatment averages are not independent.) For the clean wool experiment, the normal probability plot is obtained by means of the statements in Table 17.15, and the resulting plot is shown in Fig. 17.12. If the normality assumption for the population of bales is satisfied, the standardized bale averages should roughly lie along a line (with slope 1.0) through (0, 0). In Fig. 17.12, we see that this is roughly the case.

In summary, the random-effects one-way model with the standard distribution assumptions does not fit these data too well, since variances apparently are not constant or there is an outlier. Nevertheless, we have established that the population of bales in this shipment is extremely variable. Selected bales 1 and 7 appear to be somewhat different from the other five selected bales. Perhaps the shipment is

Fig. 17.10 Residuals versus predicted values for the clean wool experiment
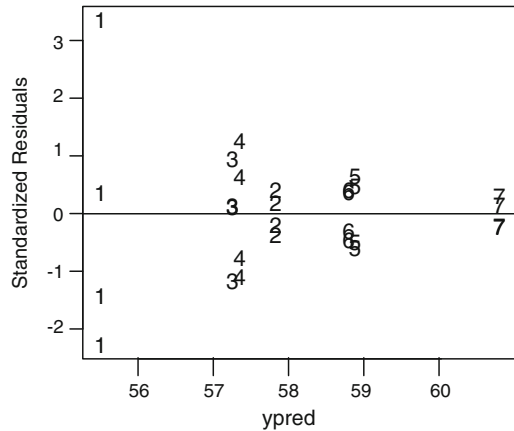

Fig. 17.11 Residuals versus predicted values for the clean wool experiment, excluding the outlier
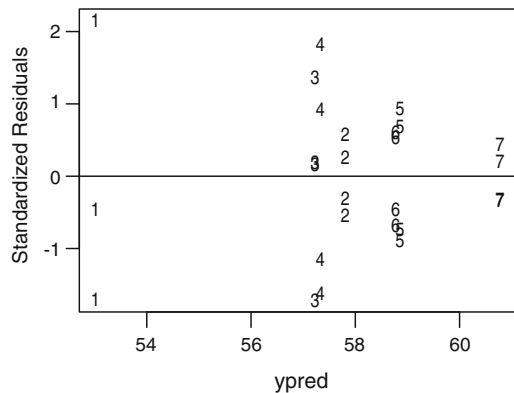
Table 17.15 R program to plot standardized treatment averages against their normal scores

```
AvgContent = by(wool.data$y, wool.data$bale, mean)  # Col of means (by bale)
# Standardized treatment means
StdzdAvg = (AvgContent - mean(AvgContent))/sd(AvgContent - mean(AvgContent))
# Normal scores
nscore = qqnorm(StdzdAvg)$x
plot(StdzdAvg ~ nscore, ylab="Standardized Trtmt Means", las=1)
   qqline(StdzdAvg)  # Line through 1st and 3rd quantile points
```

Fig. 17.12 Normal probability plot of the standardized treatment averages for the clean wool experiment
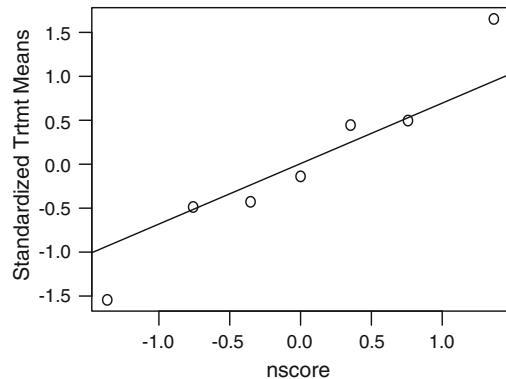
made up of dissimilar subpopulations (perhaps from different sources). This should be checked, since it may give a clue as to how to improve the wool clean content in the future.

17.11.2   Estimation and Hypothesis Testing

Analysis of Fixed Effects in Mixed Models

Analysis of variance F-tests for fixed effects in mixed models are obtained in essentially the same way as for fixed-effects models. The `aov` function is used to fit the linear model by least squares, but with the following changes: random effects are entered into the model as error terms, and the `summary` command is used to generate the analysis of variance F tests for fixed effects.

Sample R code is shown in Table 17.16 for the temperature experiment of Example 17.9.1, p. 650. F-tests for fixed effects are generated by the second block of code. In the model specified for the `aov` function, the term

```
Error(fSubj + fSubj:fSite)
```

calls the `Error` function in R, modeling subject main effects and subject–site interactions as random effects, and allowing their use as error terms for tests and for estimating standard errors for confidence intervals for means and fixed effect contrasts as appropriate. Each model can include only one `Error` function call, but the call may include multiple random effects, with two in this example. The expected mean squares are not displayed, but correct F-tests are generated for fixed effects. In particular, the denominators, calculated as explained throughout this chapter, are automatically obtained by modeling the random effects as error terms as in this sample code. Tests for fixed effects are provided, including the appropriate test statistic denominator used.

Table 17.16   R program for the temperature experiment

```
tempr.data = read.table("data/temperature.txt", header=T)
tempr.data = within(tempr.data, {
   fTherm = factor(Therm); fSite = factor(Site); fSubj = factor(Subj) })
head(tempr.data, 3)

# Least squares anova, specifying random effects as error terms
options(contrasts = c("contr.sum", "contr.poly"))
model1 = aov(Time ~ fTherm + fSite + fTherm:fSite
                  + Error(fSubj + fSubj:fSite), data=tempr.data)
summary(model1)

# Means and contrasts: estimates, CIs, tests
library(lsmeans)
lsmSite = lsmeans(model1, ~ fSite) # Compute and save lsmeans
confint(lsmSite, level=0.95) # Display lsmeans and 95% CIs
# Pairwise comparison
summary(contrast(lsmSite, method="pairwise"),
        infer=c(T,T), level=0.95, side="two-sided")
```

Corresponding output is shown in Table 17.17. The first block of output code shows information for subjects, though it is not used for any tests. The second block of output displays the F test for the main effect of site, using the mean square for subject–site interaction as the denominator of the F statistic. In the third block of output, thermometer main effects and thermometer–site interactions are each tested using the usual mean squared error as the denominator of the corresponding F statistics.

Treatment means and pairwise comparisons of fixed effects can be estimated using the `lsmeans` function of the `lsmeans` package, as illustrated by the last block of code in Table 17.16. The correct variance estimators are automatically chosen. For example, for confidence limits for site level means, the reader is asked in Exercise 9 to verify that $\text{Var}(\overline{Y}_{..j}) = (3\sigma_S^2 + 3\sigma_{SB}^2 + \sigma^2)/12$ for the jth site under model (17.9.25), p. 650. R computes the corresponding estimate, using Satterthwaite's method to compute the corresponding number of degrees of freedom, and provides corresponding lower and upper 95% confidence limits.

The last two lines of code in Table 17.16 concern the site main-effect contrast, providing information for estimation and testing this pairwise comparison. The corresponding output, displayed last in Table 17.17, correctly uses the subject–site interaction mean square to estimate the standard error for comparing sites. If site had more than two levels, one could apply Tukey's method, for example, by including the option `adjust="tukey"` in the `contrast` function as follows.

```
summary(contrast(lsmSite, method="pairwise", adjust="tukey"),
        infer=c(T,T), level=0.95, side="two-sided")
```

Also, any contrast for the fixed effects can be estimated using the following more generic syntax, and the list could be expanded to included multiple named contrasts separated by commas.

```
summary(contrast(lsmSite, list(Pairwise=c( 1, -1))),
        infer=c(T,T), level=0.95, side="two-sided
```

In each case, R will automatically compute an appropriate standard error estimate, which for the pairwise site contrast involves the subject–site interaction mean square.

Table 17.17 R analysis of fixed effects for the temperature experiment

```
> model1 = aov(Time ~ fTherm + fSite + fTherm:fSite
+               + Error(fSubj + fSubj:fSite), data=tempr.data)
> summary(model1)

Error: fSubj
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  3   1710     570

Error: fSubj:fSite
          Df Sum Sq Mean Sq F value Pr(>F)
fSite      1  86030   86030    71.1 0.0035
Residuals  3   3632    1211

Error: Within
            Df Sum Sq Mean Sq F value  Pr(>F)
fTherm       2 105759   52879    65.9 3.4e-07
fTherm:fSite 2  43794   21897    27.3 3.4e-05
Residuals   12   9631     803

> # Means and contrasts: estimates, CIs, tests
> library(lsmeans)
> lsmSite = lsmeans(model1, ~ fSite) # Compute and save lsmeans
NOTE: Results may be misleading due to involvement in interactions
> confint(lsmSite, level=0.95) # Display lsmeans and 95% CIs

 fSite lsmean     SE   df lower.CL upper.CL
 1      96.00 8.6137 5.31   74.244   117.76
 2     215.74 8.6137 5.31  193.986   237.50

Results are averaged over the levels of: fTherm
Confidence level used: 0.95

> # Pairwise comparison
> summary(contrast(lsmSite, method="pairwise"),
+         infer=c(T,T), level=0.95, side="two-sided")

 contrast estimate     SE df lower.CL upper.CL t.ratio p.value
 1 - 2     -119.74 14.205  3  -164.95  -74.536   -8.43  0.0035

Results are averaged over the levels of: fTherm
Confidence level used: 0.95
```

Analysis of Random Effects

Tests for random effects are not provided by the code discussed above for analysis fixed effects in mixed models. Still, the necessary means squares are provided, or they can be obtained by fitting a model treating all effects, including random effects, as fixed. So, based on expected mean squares, one could conduct the appropriate tests of random effects by hand for any balanced mixed- or random-effects model. The following section mentions another approach.

Other Approaches

Throughout this chapter we have discussed the analysis of mixed and random effects models using the *analysis of variance approach*—namely, fitting the model by least squares as if all effects were fixed, obtaining a corresponding analysis of variance table, then using the expected mean squares to obtain unbiased estimates of the variance components and to determine appropriate F-statistic denominators and standard error estimates. The variance component estimates so obtained are called *analysis of variance estimates*. For balanced data under normality, the least squares estimates of any estimable fixed effects are best linear unbiased estimates, and the analysis of variance estimates of variance components are minimum variance unbiased estimates, but the variance component estimates can be negative, as happens in the recently discussed temperature experiment, where the subjects variance component estimate is $\hat{\sigma}_S^2 = [ms(\text{fSubj}) - ms(\text{fSubj:fSite})]/6 \approx -106.83$.

There are other more sophisticated statistical methods for estimating variance components that prevent the estimates from ever being negative, and that are generally preferable for unbalanced data. One such approach involves estimation of the variance components by *restricted maximum likelihood* (ReML). This approach will be discussed in Chap. 18.

17.11.3   Sample Size Calculations

In this section, an R function is defined to do the sample size calculations of Example 17.4.1, p. 630, for the ice cream experiment. Recall, in testing the hypothesis $H_0^{\gamma T} : \sigma_T^2 \leq \sigma^2$ against the hypothesis $H_A^{\gamma T} : \sigma_T^2 > \sigma^2$ at a significance level of $\alpha = 0.05$, the goal is to be able to reject the null hypothesis with probability $\pi = 0.95$ if the true value of σ_T^2/σ^2 is at least $\Delta = 2.0$. In Example 17.4.1, trial and error was used to determine the number v of ice cream flavors we should look at if we took $r = 11$ or $r = 3$ observations on each. Recall, to achieve the desired power for given r, we need to find v such that

$$(F_{v-1,v(r-1),.05})(F_{v(r-1),v-1,.05}) \leq (2r+1)/(r+1) \,.$$

Table 17.18 contains the R program and corresponding output. A user-defined function, `compute.v.given.r`, is defined to do the computations. This function inputs an r value and the specified significance level and power. Given this information, the new function uses the standard R function `qf` to compute the quantile values $F_{v-1,v(r-1),.05}$ and $F_{v(r-1),v-1,.05}$ for each value $v = 2, 3, \ldots$, either stopping when the above inequality is satisfied, giving the required value of v and the result "Power = 0.95" (for power = 0.95, say), or reaching $v = 200$ and giving the result "Power < 0.95". Once done, the function returns all pertinent information. After the new function `compute.v.given.r` has been defined in Table 17.18, it is called first for $r = 11$ and then for $r = 3$, each time using significance level 0.05 and power 0.95. The results returned by the function calls, displayed in the bottom of Table 17.18, are a bit more precise than those given in Example 17.4.1.

Table 17.18 R function doing sample size calculations for the ice cream experiment

```
> # Create a user-defined function for sample size calculations
> compute.v.given.r = function(r=5,alpha=0.05,power=0.95){
+     # Initialize variables
+     ratio = (2*r+1)/(r+1);
+     result="Power <"
+     v=1;
+     while((result=="Power <")&(v<201)){
+         v=v+1;
+         df1=v-1; df2=v*(r-1);
+         F1 = qf(1-alpha,df1,df2) # Compute F(df1,df2,alpha)
+         F2 = qf(power,df2,df1)   # Compute F(df2,df1,1-power)
+         product=F1*F2
+         if (product<ratio) {result="Power ="}
+     } # End while loop, either finding v or power less than target
+     data.frame(r,v,result,power,alpha,F1,F2,product,ratio,df1,df2)
+ } # end function
>
> compute.v.given.r(r=11,alpha=0.05,power=0.95) # Call the function
   r  v  result power alpha     F1     F2 product  ratio df1 df2
1 11 58 Power =  0.95  0.05 1.3501 1.4193  1.9162 1.9167  57 580
>
> compute.v.given.r(r=3,alpha=0.05,power=0.95) # Call the function
   r   v  result power alpha     F1     F2 product ratio df1 df2
1 3 106 Power =  0.95  0.05 1.3113 1.3314  1.7459  1.75 105 212
```

Table 17.19 Data for the alcohol experiment

Bottle	Concentration (mg/ml)			
1	1.4357	1.4348	1.4336	1.4309
2	1.4244	1.4232	1.4213	1.4256
3	1.4153	1.4137	1.4176	1.4164
4	1.4331	1.4325	1.4312	1.4297
5	1.4252	1.4261	1.4293	1.4272
6	1.4179	1.4217	1.4191	1.4204

Exercises

1. **Alcohol experiment**

   Solutions of alcohol are used for calibrating Breathalyzers. The data in Table 17.19 show the alcohol concentrations (mg/ml) of samples of alcohol solutions taken from six bottles of alcohol solution randomly selected from a large batch. Concentrations are determined by gas chromatography.

   (a) Check the assumptions on the random-effects one-way model for these data.
   (b) Calculate a 95% upper confidence bound for the error variance.
   (c) Calculate a 95% confidence interval for the variance of the alcohol concentrations in the population of bottles in this large batch.

(d) Test the hypothesis that the variance of the alcohol concentrations is at most five times the error variance versus the alternative hypothesis that it is not. Use a significance level of $\alpha = 0.05$.

2. **Ice cream experiment, continued**

   As in Example 17.4.1, p. 630, suppose the ice cream experiment is to be repeated, with $\gamma = 1.0$ and with a Type I error probability of $\alpha = .05$. Suppose that we would like to reject the null hypothesis $H_0^{\gamma T} : \{\sigma_T^2 \le \sigma^2\}$ with probability $\pi = 0.95$ if the true value of σ_T^2/σ^2 is greater than $\Delta = 2.0$. How many ice cream flavors should be included in the experiment if $r = 2$ observations are to be taken on each? How many observations are needed? Is this an improvement over the result in the example for $r = 3$? (Note: $F_{150,150,.05} = 1.309$, $F_{160,160,.05} = 1.298$, $F_{170,170,.05} = 1.288$, $F_{180,180,.05} = 1.279$).

3. **Random effects model**

   Consider the following random-effects model:

   $$
   \begin{aligned}
   Y_{ijkmt} &= \mu + A_i + B_j + C_k + D_m \\
   &\quad + (AB)_{ij} + (BC)_{jk} + (BD)_{jm} + \epsilon_{ijkmt}, \\
   i &= 1, \ldots, a, \ \ j = 1, \ldots, b, \ \ k = 1, \ldots, c, \\
   m &= 1, \ldots, d, \ \ t = 1, \ldots, r, \\
   A_i &\sim N(0, \sigma_A^2), \ \ B_j \sim N(0, \sigma_B^2), \ \ C_k \sim N(0, \sigma_C^2), \\
   D_m &\sim N(0, \sigma_D^2), \ \ (AB)_{ij} \sim N(0, \sigma_{AB}^2), \ \ (BC)_{jk} \sim N(0, \sigma_{BC}^2), \\
   (BD)_{jm} &\sim N(0, \sigma_{BD}^2), \ \ \epsilon_{ijkmt} \sim N(0, \sigma^2),
   \end{aligned}
   $$

   where all random variables on the right hand side of the model are mutually independent.

   (a) Write out the expected mean squares for all main effects and interactions in the model.
   (b) How would you test the null hypothesis $H_0^A : \{\sigma_A^2 = 0\}$ against the alternative hypothesis $H_A^A : \{\sigma_A^2 > 0\}$?
   (c) How would you test the null hypothesis $H_0^B : \{\sigma_B^2 = 0\}$ against the alternative hypothesis $H_A^B : \{\sigma_B^2 > 0\}$?
   (d) Give formulae for unbiased estimates of σ_{BD}^2 and σ_B^2.
   (e) Give formulae for individual 95% confidence intervals for σ_{BD}^2 and σ_B^2. What is the overall confidence level?

4. **Buttermilk biscuit experiment**

   The buttermilk biscuit experiment was run by Stacie Taylor in 1995 to find out which brands of refrigerated buttermilk biscuit give rise to the fluffiest biscuits. Three brands were examined (factor A, 3 levels, fixed effect), all of which had claims to be light, fluffy, or flaky in their advertising campaigns. The biscuits were baked on a baking tray for 7 min in the center of an oven set to 425°F. Since only six biscuits could be baked at a time, the experiment was run as a general complete block design with blocks of size $k = 6$.

   (a) Use a mixed model with interaction to represent the data, where the random effect represents the block (run of the oven) and the fixed effect represents the biscuit brand. Write out the model including all of the assumptions.

Table 17.20 Treatments and percentage change in height for the buttermilk biscuit experiment

Block	Position					
	1	2	3	4	5	6
1	2 (150.0)	1 (188.2)	2 (177.8)	3 (166.7)	3 (187.5)	1 (182.4)
2	1 (183.3)	2 (183.3)	2 (183.3)	3 (176.5)	1 (160.0)	3 (187.5)
3	1 (178.9)	3 (182.4)	2 (193.8)	3 (176.5)	2 (188.9)	1 (188.9)
4	2 (177.8)	1 (145.5)	3 (155.0)	1 (173.7)	3 (200.0)	2 (187.5)
5	1 (205.6)	3 (188.2)	3 (142.9)	2 (161.9)	2 (177.8)	1 (159.1)

Table 17.21 Data for the candle experiment (seconds)

Person	Color							
	Red		White		Blue		Yellow	
1	989	1032	1044	979	1011	951	974	998
	1077	1019	987	1031	928	1022	1033	1041
2	899	912	847	880	899	800	886	859
	911	943	879	830	820	812	901	907
3	898	840	840	952	909	790	950	992
	955	1005	961	915	871	905	920	890
4	993	957	987	960	864	925	949	973
	1005	982	920	1001	824	790	978	938

(b) The data collected by the experimenter are shown in Table 17.20. As far as possible, check the assumptions on the model for these data.

(c) Write out the expected mean squares for all terms in the model.

(d) Draw a block×brand interaction plot for those blocks observed in the experiment.

(e) Test the hypothesis that the variance in height of the biscuits due the population of block×brand interactions is negligible against the alternative hypothesis that it is not negligible. Interpret your conclusions in terms of the plot in part (d).

(f) Calculate a set of 95% simultaneous confidence intervals for the pairwise comparisons between the brands. State your conclusions.

5. **Candle experiment**

An experiment to determine whether different colored candles (red, white, blue, yellow) burn at different speeds was conducted by Hsing-Chuan Tsai, Mei-Chiao Yang, Derek Wheeler, and Tom Schultz in 1989. Each experimenter collected four observations on each color in a random order, and "experimenter" was used as a blocking factor. Thus, the design was a general complete block design with $v = 4, k = 16, b = 4$, and $s = 4$. The resulting burning times (in seconds) are shown in Table 17.21. A pilot experiment indicated that treatments and blocks do interact. The candles used in the experiment were cake candles made by a single manufacturer. Analyze the experiment as though the experimenters represent a random sample from a large population of people who might use these candles in practice. Use a two-way mixed model with interaction.

6. **Golf ball experiment**

   An experiment was planned by Tim Kelaghan in 1995 to examine whether different brands of golf balls travel on average the same distances when hit by amateur golfers. The experiment was planned with a specific selection of $v = 3$ golf balls and some number b of golfers to be determined. The experiment was to be run as a general complete block design with fixed treatment effects and random golfer effects. Since the golfer is aware of which brand of ball he or she is hitting, there may well be a golfer×brand interaction. However, the differences between brands averaged over the interaction is important here.

   A small pilot experiment was conducted. There were only two golfers, and each hit $s = 6$ balls of each brand in a random order. Mis-hits were ignored. The distances that the balls traveled were recorded in yards and are shown in Table 17.22.

   (a) Use the pilot experiment data to calculate a 95% upper bound for the error variance σ^2.
   (b) The experimenter wanted the main experiment to be able to calculate a set of simultaneous 95% confidence intervals for the pairwise differences in the brands, and he wanted the widths of these intervals to be at most 20 yards.
   Assuming that each golfer would hit about 18 balls in total, as in the pilot experiment, how many randomly selected golfers would be needed?

7. **Mixed model**

   Consider the following mixed model:

   $$\begin{aligned}
   Y_{ijkmt} = {} & \mu + \alpha_i + B_j + C_k + \delta_m + (\alpha B)_{ij} + (\alpha \delta)_{im} \\
   & + (B\delta)_{jm} + (C\delta)_{km} + (\alpha B\delta)_{ijm} + \epsilon_{ijkmt}, \\
   & i = 1, \ldots, a, \ \ j = 1, \ldots, b, \ \ k = 1, \ldots, c, \\
   & m = 1, \ldots, d, \ \ t = 1, \ldots, r, \\
   B_j \sim {} & N(0, \sigma_B^2), \ \ C_k \sim N(0, \sigma_C^2), \ \ (\alpha B)_{ij} \sim N(0, \sigma_{AB}^2), \\
   & (B\delta)_{jm} \sim N(0, \sigma_{BD}^2), \ \ (C\delta)_{km} \sim N(0, \sigma_{CD}^2), \\
   & (\alpha B\delta)_{ijm} \sim N(0, \sigma_{ABD}^2), \ \ \epsilon_{ijkmt} \sim N(0, \sigma^2),
   \end{aligned}$$

   where α_i and δ_m are fixed effects, all other effects are random effects, and all random variables on the right-hand side of the model are mutually independent.

Table 17.22 Distances (in yards) traveled by balls in the golf experiment

Golfer	Brand	Distance					
		1	2	3	4	5	6
1	1	209	204	179	230	233	245
	2	188	211	242	222	187	233
	3	219	204	247	215	197	161
2	1	240	207	192	190	226	188
	2	216	195	240	215	219	238
	3	195	221	205	192	183	230

(a) Write out the expected mean squares for all main effects and interactions in the model.
(b) How would you test the hypothesis $H_0 : \{\delta_m + \overline{(\alpha\delta)}_{.m}$ all equal$\}$ against the alternative hypothesis that these parameters are not all equal?
(c) Give a formula for an unbiased estimate of σ_B^2.
(d) Give a formula for a 95% confidence interval for σ_B^2.

8. **Ice cream experiment, continued**

The ice cream experiment was described in Example 17.3.1, p. 621, and was analyzed in Examples 17.3.2–17.3.4 and 17.4.1. In Sect. 17.10, p. 657, a new model was suggested that involved a quadratic time trend.

(a) What could account for a quadratic time trend?
(b) Investigate the assumptions on the models with and without the quadratic time trend.
(c) Redo the analyses of Examples 17.3.3 and 17.3.4 for the new model and compare your answers with the original model.
(d) Which model do you prefer and why?

9. **Temperature experiment, continued**

The temperature experiment was described and analyzed in Example 17.9.1, with corresponding SAS and R software analysis provided in Sects. 17.10.2 and 17.11.2, respectively. The experiment was to compare the times required for three different digital thermometers (factor A at three levels) to register body temperature at two different sites—in the mouth and under the arm—(factor B at two levels). A randomized complete block design was used, with each of the six treatment combinations observed once on each of four subjects (factor S at four levels). Using the mixed model (17.9.25), p. 650, consider estimation of the treatment mean $\bar{\mu}_{..j} = \mu + \bar{\alpha}_{.} + \beta_j + \overline{(\alpha\beta)}_{.j}$ for the jth site.

(a) Verify that $\text{Var}(\overline{Y}_{..j}) = (3\sigma_S^2 + 3\sigma_{SB}^2 + \sigma^2)/12$.
(b) Provide an unbiased estimate of $\text{Var}(\overline{Y}_{..j})$.
(c) Compute the number of degrees of freedom associated with the estimate in part (b).
(d) Given that $\bar{y}_{..1} = 96.00$, construct a 95% confidence interval for $\bar{\mu}_{..1}$.

Nested Models 18

18.1 Introduction

A factor is said to be *nested* within a second factor if each of its levels is observed in conjunction with just one level of the second factor. An example can be obtained from the clean wool experiment that was discussed in the last chapter. There, the objective of the experiment was to examine the variability of the "clean content" among bales of wool in a large shipment. Several bales were selected for examination, and several cores were taken from each bale and measured. Each core was taken from only one bale, so the cores (levels of the first factor) are observed in conjunction with only one bale (level of the second factor). In the above language, the cores are *nested within the bales*. In the original experiment, there was only one observation taken on each core. The variability of the different cores could not, therefore, be distinguished from measurement error, and their effects were not included explicitly in the model. Had there been more than one observation per core, we could have included in the model separate effects due to bales, cores nested within bales, and experimental error.

In this chapter we discuss how to recognize nested factors, how to formulate the associated models, and how to analyze the effects in these models. Many of the analysis techniques are similar to those in the previous chapter.

In the next section we discuss some examples of hypothetical experiments involving nested effects, and possible models to represent the data. In Sect. 18.3, we find the estimable contrasts for fixed-effects nested models and develop tests of hypotheses and confidence intervals for these. The more usual setting where the nested effects are random effects is discussed in Sect. 18.4 and, where possible, we borrow the formulae from the fixed effects setting as we did in Chap. 17. The rules of Chaps. 7 and 17 for finding degrees of freedom, sums of squares and expected mean squares and variance components are then extended to encompass nested models. The analysis of nested models using the SAS and R computer packages is discussed in Sects. 18.5 and 18.6, respectively.

18.2 Examples and Models

Nested factors are usually, but not always, random effects, and they are usually, but not always, blocking factors. In the following examples, we give a selection of different situations involving random effects and suggest some reasonable models to represent the data.

Example 18.2.1 `Machine head experiment`

Hicks (1956) describes a simple experiment to study the differences in the strain readings (the response) of four different heads on each of five different machines. The heads on each machine were supposedly all doing the same job and should have given rise to similar (nonvariable) readings.

Since each head was observed on only one machine, the heads were "nested within machines," giving twenty heads in total. Four observations were taken on each head. The usual two-way analysis of variance model is not appropriate here, since it would read

$$Y_{ijt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijt} \,,$$
$$\epsilon_{ijt} \sim N(0, \sigma^2) \,,$$
$$\epsilon_{ijt}\text{'s are mutually independent,}$$
$$t = 1, \ldots, 4; \quad i = 1, \ldots, 5; \quad j = 1, \ldots, 4,$$

where, α_i is the effect of the ith machine, β_j is the effect of the jth head, and $(\alpha\beta)_{ij}$ is the extra effect of observing the ith machine and jth head together. This suggests that every head is observed on every machine, which was not the case. Instead, we need a notation that will clearly indicate the nested nature of the factors. One popular notation, which we shall adopt here, is to replace $\beta_j + (\alpha\beta)_{ij}$ by $\beta_{j(i)}$, where the parentheses indicate that we are looking at the head that happens to be numbered as the *jth head on the ith machine*. The *two-way nested model* is then

$$Y_{ijt} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijt} \,, \tag{18.2.1}$$
$$\epsilon_{ijt} \sim N(0, \sigma^2) \,,$$
$$\epsilon_{ijt}\text{'s are mutually independent,}$$
$$t = 1, \ldots, 4; \quad i = 1, \ldots, 5; \quad j = 1, \ldots, 4.$$

We note in passing that the response Y_{ijt} could also be written as a nested effect $Y_{t(ij)}$, since this represents the tth observation that is specific to the (ij)th machine head. However, since this representation is not crucial to the analysis, we will continue to use the notation Y_{ijt} that we have used so far throughout the book.

One final consideration is whether the machine effects and head effects should be fixed or random. Let us first suppose that the five machines are the only machines of this type in the factory and that they are not due for replacement. The experimenter would then be interested in these five machines specifically, and their effects on the response would be modeled as fixed effects. Let us alternatively suppose that machine heads wear out and are continually being replaced. The experimenter would then be interested in the population of heads from which the particular twenty in the experiment were drawn. Consequently, the nested head effect would be modeled as a random effect. The model would be written as

$$Y_{ijt} = \mu + \alpha_i + B_{j(i)} + \epsilon_{ijt} \,, \tag{18.2.2}$$
$$\epsilon_{ijt} \sim N(0, \sigma^2) \,, \quad B_{j(i)} \sim N(0, \sigma^2_{B(A)}) \,,$$
$$\epsilon_{ijt}\text{'s and } B_{j(i)}\text{'s are all mutually independent,}$$
$$t = 1, \ldots, 4; \quad i = 1, \ldots, 5; \quad j = 1, \ldots, 4,$$

where α_i is the effect of the ith machine, and $\sigma^2_{B(A)}$ is the variance of responses from the population of machine heads that could be fitted on these five machines. Notice that all random variables on the right-hand side of the model are assumed to be mutually independent. □

In the previous example there were two treatment factors, one of whose levels were nested within those of the other. In the following experiment, there are two blocking factors, which are nested one within the other.

Example 18.2.2 `Efficiency experiment`

An experiment was run in 1997 by Carina Dalton, Greg Krzys, Scott O'Dee, and Brad Welch to examine the assertion that "a person works more efficiently when there is no one looking over his or her shoulder." Twelve subjects were recruited for the experiment, and three of these were assigned to each of the four experimenters. Each subject was asked to complete a simple task—crossing through every occurrence of the letter "e" on a page of prose. There were two levels of the treatment factor. Level 1 required the assigned experimenter to look over the subject's shoulder while the task was being completed, and level 2 required the experimenter to be elsewhere in the room absorbed in a book. The response was the time taken to complete the task. Each subject was assigned both treatments, but in a randomized order.

The blocking factor in this experiment was subject. However, the subjects each worked with only one experimenter, and so the subject effects were nested within the experimenter effects.

The subjects were graduate students at The Ohio State University. Although they were not selected according to the rules of a simple random sample, let us suppose that they were a reasonable representation of that population. Let us also suppose that the variation among the techniques of the experimenters, who were also graduate students, was representative of a population of student experimenters. It might also be reasonable to assume that some subjects may be more perturbed than others about an experimenter watching them complete the task. In this case, we might wish to include a subject–treatment interaction in the model. However, there is only one observation per subject per treatment, so the subject–treatment interaction could not be distinguished from the random error.

A second possible model would be to include an experimenter–treatment interaction instead of a subject–treatment interaction. Such an interaction might occur if the actions of the four experimenters were not all identical. In this case the model would be

$$Y_{hqi} = \mu + E_h + S_{q(h)} + \alpha_i + (\alpha E)_{hi} + \epsilon_{hqi} ,$$
$$\epsilon_{hqi} \sim N(0, \sigma^2) , \quad S_{q(h)} \sim N(0, \sigma^2_{S(E)}) , \quad (\alpha E)_{hi} \sim N(0, \sigma^2_{EA}) ,$$
$$\epsilon_{hqi}\text{'s}, E_h\text{'s}, S_{q(h)}\text{'s and } (\alpha E)_{hi}\text{'s are all mutually independent,}$$
$$h = 1, \ldots, 4; \quad i = 1, 2, 3; \quad i = 1, 2 .$$

where E_h is the effect of the hth randomly selected experimenter, $S_{q(h)}$ is the effect of the qth randomly selected subject assigned to the hth experimenter, α_i is the effect of the ith treatment, and $(\alpha E)_{hi}$ is the random effect representing the interaction between the hth experimenter and the ith treatment.

Lastly, we may also wish to include a time-order effect in the model, since the subjects may have been able to complete the task faster on the second occasion just due to familiarity. So we could add the extra term γx_{hqi}, where x_{hqi} is 1 or 2 according to whether the (hq)th subject is assigned treatment i on the first or second occasion. □

18.3 Analysis of Nested Fixed Effects

18.3.1 Least Squares Estimates

Consider first the simplest possible fixed-effects nested model—the two-way nested model (18.2.1) that was suggested for the machine head experiment of Example 18.2.1; that is,

$$Y_{ijt} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijt} ,$$
$$\epsilon_{ijt} \sim N(0, \sigma^2) ,$$
$$\epsilon_{ijt}\text{'s are mutually independent,}$$
$$t = 1, \ldots, r_{ij}; \quad i = 1, \ldots, a; \quad j = 1, \ldots, b .$$

The error assumptions are examined in the same way as in Chap. 5 for the one-way analysis of variance model. In any model, the estimable contrasts are functions of the expected values of the response variables (see, for example, Sect. 3.4.1, p. 34). In the present model, $E[Y_{ijt}]$ is equal to

$$E[Y_{ijt}] = \mu + \alpha_i + \beta_{j(i)} .$$

If we take an average over the subscripts t and j, we find that a comparison of the levels of A averaged over the levels of B is estimable; that is, we can estimate pairwise comparisons such as

$$\left[\alpha_i + \overline{\beta}_{.(i)}\right] - \left[\alpha_s + \overline{\beta}_{.(s)}\right] ,$$

and we can estimate general contrasts such as

$$\sum_{i=1}^{a} c_i \left[\alpha_i + \overline{\beta}_{.(i)}\right] , \quad \text{with} \sum_{i=1}^{a} c_i = 0 .$$

We can also compare the effects of those levels of B that were observed in conjunction with the *same* level of A; that is,

$$\left[\alpha_i + \beta_{j(i)}\right] - \left[\alpha_i + \beta_{u(i)}\right] = \beta_{j(i)} - \beta_{u(i)} ,$$

or, in general,

$$\sum_{j=1}^{b} d_j \beta_{j(i)} , \quad \text{with} \sum_{j=1}^{b} d_j = 0 , \quad \text{for any given } i .$$

To obtain the least squares estimators of estimable contrasts, we use the method of least squares to find parameter estimates that minimize the sum of squared errors

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{t=1}^{r_{ij}} e_{ijt}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{t=1}^{r_{ij}} \left(y_{ijt} - \mu - \alpha_i - \beta_{j(i)}\right)^2 .$$

Readers with a knowledge of calculus may verify (see Exercise 7) that the least squares estimate of $\mu + \alpha_i + \beta_{j(i)}$ is $\overline{y}_{ij.}$. Consequently, the least squares estimator of

$$\sum_{i=1}^{a} c_i \left[\alpha_i + \overline{\beta}_{.(i)} \right] \quad \text{is} \quad \sum_{i=1}^{a} c_i \overline{Y}_{i..}$$

with $\Sigma c_i = 0$. The corresponding variance is $\Sigma c_i^2 \sigma^2 / r_{i.}$. Similarly, the least squares estimator of

$$\sum_{j=1}^{b} d_j \beta_{j(i)} \quad \text{is} \quad \sum_{j=1}^{b} d_j \overline{Y}_{ij.} \quad \text{for any } i$$

with $\Sigma d_j = 0$. The corresponding variance is $\Sigma d_j^2 \sigma^2 / r_{ij}$.

All of these formulae can easily be adapted to the case where B has a different number of levels for each level of A by replacing b by b_i.

18.3.2   Estimation of σ^2

The error sum of squares is

$$
\begin{aligned}
ssE &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{t=1}^{r_{ij}} \left(y_{ijt} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_{j(i)} \right)^2 \\
&= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{t=1}^{r_{ij}} \left(y_{ijt} - \overline{y}_{ij.} \right)^2 \quad\quad\quad\quad (18.3.3) \\
&= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{t=1}^{r_{ij}} y_{ijt}^2 - \sum_{i=1}^{a} \sum_{j=1}^{b} r_{ij} \overline{y}_{ij.}^2 . \quad\quad\quad (18.3.4)
\end{aligned}
$$

A comparison with the formulae in Sect. 6.4 shows that everything that we have written so far about the fixed-effects two-way nested model could have been deduced from the fixed-effects two-way complete model after replacing $\beta_j + (\alpha\beta)_{ij}$ by $\beta_{j(i)}$. Therefore, we may also deduce that the error mean square, $msE = ssE/(n - v)$, gives an unbiased estimate for σ^2, and the corresponding random variable MSE has a chi-squared distribution with $n - v$ degrees of freedom (where $n = r_{..}$ and $v = ab$).

18.3.3   Confidence Intervals

We may obtain a $100(1 - \alpha)\%$ confidence bound for σ^2 from the information in the previous subsection; that is,

$$\sigma^2 \leq \frac{ssE}{\chi_{n-v, 1-\alpha}^2} .$$

The derivation of the bound was explained in Sect. 3.4.6.

Confidence intervals for $\Sigma c_i (\alpha_i + \overline{\beta}_{.(i)})$ and for $\Sigma d_j \beta_{j(i)}$ may be obtained using the relevant methods from Chap. 4 together with the formulae

$$\Sigma c_i \overline{Y}_{i..} \ \pm \ w \sqrt{\sum_i \left(\frac{c_i^2}{r_i} \right) msE}$$

and

$$\Sigma d_j \overline{Y}_{ij.} \ \pm \ w \sqrt{\sum_j \left(\frac{d_j^2}{r_{ij}} \right) msE}.$$

18.3.4   Hypothesis Testing

We may obtain a test of the null hypothesis that the levels of B have the same effect on the response within every given level of A, that is,

$$H_0^{B(A)} : \{\beta_{1(i)} = \beta_{2(i)} = \ldots = \beta_{b(i)}, \text{ for every } i = 1, \ldots, a\},$$

against the alternative hypothesis $H_A^{B(A)} : \{H_0^{B(A)}$ is not true$\}$ by comparing the sum of squares for error (18.3.3) in the fixed-effects two-way nested model with the sum of squares for error in the reduced (one-way) model. The reduced model is

$$Y_{ijt} = \mu^* + \alpha_i + \epsilon_{ijt},$$

and the error sum of squares is given by (3.4.4), p. 39, with an extra subscript; that is,

$$ssE_0 = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{t=1}^{r_{ij}} (y_{ijt} - \overline{y}_{i..})^2.$$

The numerator of the test statistic is then

$$msB(A) = \frac{ssB(A)}{a(b-1)},$$

where the number of degrees of freedom for $B(A)$ is obtained as the difference between the error degrees of freedom in the reduced and full models; that is,

$$(n - a) - (n - v) \ = \ v - a \ = \ ab - a \ = \ a(b - 1),$$

and where

$$\begin{aligned}
ssB(A) &= ssE_0 - ssE \\
&= \sum_i \sum_j \sum_t (y_{ijt} - \overline{y}_{i..})^2 - \sum_i \sum_j \sum_t (y_{ijt} - \overline{y}_{ij.})^2 \\
&= \sum_i \sum_j \sum_t y_{ijt}^2 - \sum_i r_{i.} \overline{y}_{i..}^2 - \sum_i \sum_j \sum_t y_{ijt}^2 + \sum_i \sum_j r_{ij} \overline{y}_{ij.}^2 \\
&= \sum_i \sum_j r_{ij} \overline{y}_{ij.}^2 - \sum_i r_{i.} \overline{y}_{i..}^2
\end{aligned} \tag{18.3.5}$$

$$= \sum_i \sum_j r_{ij}(\overline{y}_{ij.} - \overline{y}_{i..})^2 \,. \tag{18.3.6}$$

The decision rule for testing $H_0^{B(A)}$ versus $H_A^{B(A)}$ at significance level α is

$$\text{reject } H_0^{B(A)} \quad \text{if} \quad \frac{ssB(A)/a(b-1)}{ssE/(n-ab)} \; > \; F_{a(b-1),n-ab,\alpha} \,.$$

Similarly, the decision rule for testing

$$H_0^A : \{\alpha_i + \overline{\beta}_{.(i)} \text{ all equal}\}$$

against the alternative hypothesis $H_A^A : \{H_0^A \text{ is false}\}$ is

$$\text{reject } H_0^A \text{ if } \frac{ssA/(a-1)}{ssE/(n-ab)} \; > \; F_{a-1,n-ab,\alpha} \,,$$

where

$$ssA = \sum_i r_{i.}(\overline{y}_{i..} - \overline{y}_{...})^2 = \sum_i r_{i.}\overline{y}_{i..}^2 - n\overline{y}_{...}^2 \,.$$

Notice that $ssB(A)$ in the two-way nested model is equal to $ssB + ssAB$ in the two-way complete model. Also, the degrees of freedom for $B(A)$ in the nested model can be obtained as the sum of the degrees of freedom for B and AB in the complete model; that is,

$$(b-1) + (b-1)(a-1) \; = \; a(b-1) \,.$$

This link between the nested model and the corresponding complete model means that when the sample sizes are equal, we can obtain all the formulae we need from the rules in Chap. 7. This remains true for more complicated models also. For example, if we take the nested model

$$Y_{ijkt} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{k(ij)} + \epsilon_{ijkt} \,,$$

we have the following equivalences with the terms of the three-way complete model:

$$\beta_{j(i)} = \beta_j + (\alpha\beta)_{ij} \,,$$
$$\gamma_{k(ij)} = \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \,;$$

so, for example, the sum of squares for $C(AB)$ is

$$\begin{aligned}
ssC(AB) &= ssC + ssAC + ssBC + ssABC \\
&= \sum_i \sum_j \sum_k r_{ijk}\overline{y}_{ijk.}^2 - \sum_i \sum_j r_{ij.}\overline{y}_{ij..}^2 \\
&= \sum_i \sum_j \sum_k r_{ijk}(\overline{y}_{ijk.} - \overline{y}_{ij..})^2 \,,
\end{aligned}$$

with degrees of freedom

$$(c-1) + (a-1)(c-1) + (b-1)(c-1) + (a-1)(b-1)(c-1) = ab(c-1).$$

As with the crossed model, the degrees of freedom for *C(AB)* give a clue to the subscripts needed in the formula for the sum of squares for *C(AB)*; that is, the degrees of freedom $ab(c-1) = abc - ab$ suggest that the sum of squares for *C(AB)* must contain the terms $\overline{y}_{ijk.}$ and $\overline{y}_{ij..}$, the latter with a minus sign.

To obtain the degrees of freedom corresponding to any effect, we notice that the degrees of freedom for *A* are the same as in the crossed model; that is, $(a-1)$. The degrees of freedom for *B(A)* are $(b-1) + (a-1)(b-1) = a(b-1)$, and those for *C(AB)* are $ab(c-1)$. Thus we see a pattern. The number of degrees of freedom is the product of the numbers of levels corresponding to the factors in parentheses and one less than the numbers of levels corresponding to the factors not in parentheses. We may now modify rules 1 and 2 in Sect. 7.3 for equal sample sizes listed below. We also include rules 3 and 4 here for easy reference, although these remain the same.

1. Write down the name of the main effect or interaction of interest and the corresponding number of levels and subscripts. Include parentheses to denote nesting of factors.
2. The number of degrees of freedom ν for any effect is the product of the numbers of levels corresponding to the factors in parentheses and one less than the numbers of levels corresponding to the factors not in parentheses.
3. Multiply out the number of degrees of freedom and replace each letter with the corresponding subscripts.
4. The sum of squares for testing the hypothesis that a main effect or an interaction is negligible is obtained as follows. Use each group of subscripts in rule 3 as the subscripts of a term $\overline{y}$, averaging over all subscripts not present and keeping the same signs. Put the resulting estimate in parentheses, square it and sum over all possible subscripts. To expand the parentheses, square each term in the parentheses, keep the same signs, and sum over all possible subscripts.

The other rules remain the same. In particular, confidence intervals for $\sum_{i=1}^{a} c_i \left(\alpha_i + \overline{\beta}_{.(i)} \right)$ and for $\sum_{j=1}^{b} d_j \beta_{j(i)}$ may be calculated using the usual multiple-comparison techniques of Chap. 4.

Example 18.3.1 `Plastic experiment`

Consider the following hypothetical experiment in which a manufacturer of molded plastic wishes to replace a standard ingredient by a cheaper alternative. The two ingredients form the two levels of the treatment factor to be studied. The manufacturing company has factories in three different parts of the country, and since different climates may affect the product differently, the experiment is to take place in each of the three locations. Within each factory, two operators oversee two machines each. The experiment will be run during the usual downtime of the machines.

A possible model for the experiment is

$$Y_{ijkut} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{k(ij)} + \tau_u + (\tau\gamma)_{uk(ij)} + \epsilon_{ijkut},$$
$$\epsilon_{ijkut} \sim N(0, \sigma^2),$$
$$\epsilon_{ijkut}\text{'s are mutually independent,}$$
$$t = 1, \ldots, r; \quad i = 1, 2, 3; \quad j = 1, 2; \quad k = 1, 2; \quad u = 1, 2;$$

where α_i is the effect of the ith location, $\beta_{j(i)}$ is the effect of the jth operator at the ith location, $\gamma_{k(ij)}$ is the effect of the kth machine that is looked after by the jth operator at the ith location, τ_u is the effect of the uth treatment, $(\tau\gamma)_{uk(ij)}$ is the interaction effect between the uth treatment and (ijk)th

Table 18.1 Degrees of freedom and sums of squares

Effect	Degrees of freedom	Sum of squares
A	$a - 1 = 2$	$bcdr \Sigma_i \bar{y}_{i\dots}^2 - abcdr \bar{y}_{\dots}^2$
$B(A)$	$a(b - 1) = 3$	$cdr \Sigma_i \Sigma_j \bar{y}_{ij\dots}^2 - bcdr \Sigma_i \bar{y}_{i\dots}^2$
$C(AB)$	$ab(c - 1) = 6$	$dr \Sigma_i \Sigma_j \Sigma_k \bar{y}_{ijk\cdot\cdot}^2 - cdr \Sigma_i \Sigma_j \bar{y}_{ij\dots}^2$
Trt	$d - 1 = 1$	$abcr \Sigma_u \bar{y}_{\dots u\cdot}^2 - abcdr \bar{y}_{\dots}^2$
Trt $\times$ $C(AB)$	$ab(c - 1)(d - 1) = 6$	$r \Sigma_i \Sigma_j \Sigma_k \Sigma_u \bar{y}_{ijku}^2 - dr \Sigma_i \Sigma_j \Sigma_k \bar{y}_{ijk\cdot\cdot}^2$
		$\qquad -cr \Sigma_i \Sigma_j \Sigma_u \bar{y}_{ij\cdot u\cdot}^2 + cdr \Sigma_i \Sigma_j \bar{y}_{ij\dots}^2$
Error	$24r - 19$, by subtraction	Obtain by subtraction
Total	$n - 1 = 24r - 1$	$\Sigma_i \Sigma_j \Sigma_k \Sigma_u \Sigma_t y_{ijkut}^2 - abcdr \bar{y}_{\dots}^2$

machine, Y_{ijkut} is the tth response (strength measurement) on the uth treatment and (ijk)th machine, and ϵ_{ijkut} is the corresponding random error, assumed to have a normal distribution with mean 0 and variance σ^2. We also assume that the error variables are mutually independent.

The degrees of freedom and sums of squares for each effect are obtained from rules 1–4 listed above this example and are shown in Table 18.1. Using the formula for confidence intervals in Sect. 18.3.3, we may obtain a confidence interval for $\Sigma_u h_u \tau_u$ as

$$\sum_u h_u \bar{y}_{\dots u\cdot} \pm w \sqrt{\sum_u \frac{h_u^2}{12r} msE}.$$

$\square$

18.4   Analysis of Nested Random Effects

18.4.1   Expected Mean Squares

In Chap. 17 we found that we could modify many of the formulae arising from the fixed-effect crossed models to obtain confidence intervals and hypothesis tests for variance components in the corresponding random-effects models. To find the denominators for the hypothesis tests and to find the estimates for variance components, all we need to do is to calculate the expected values of the mean squares arising from the corresponding fixed-effect models. For equal sample sizes, expected mean squares can be obtained using rule 17 of Sect. 17.8.1, p. 647, exactly as for the random- and mixed-effects crossed models. The rules 18–22 for calculating test ratios and confidence intervals also follow exactly as for the crossed models. We will illustrate these via the model that was suggested for the machine head experiment in Example 18.2.1. A suggested model was

$$Y_{ijt} = \mu + \alpha_i + B_{j(i)} + \epsilon_{ijt},$$
$$\epsilon_{ijt} \sim N(0, \sigma^2), \quad B_{j(i)} \sim N(0, \sigma_{B(A)}^2),$$
$$\epsilon_{ijt}\text{'s and } B_{j(i)}\text{'s are all mutually independent},$$
$$t = 1, \dots, 4; \quad i = 1, \dots, 5; \quad j = 1, \dots, 4,$$

Table 18.2 Analysis of variance table for a mixed-effects two-way nested model

Source of variation	Deg. of freedom	Sum of squares	Mean square	Expected mean square
A	$(a-1)$	ssA	msA	$Q(\alpha_i) + r\sigma^2_{B(A)} + \sigma^2$
B(A)	$a(b-1)$	$ssB(A)$	$msB(A)$	$r\sigma^2_{B(A)} + \sigma^2$
Error	$ab(r-1)$	ssE	msE	σ^2
Total	$abr-1$	$sstot$		
Formulae for equal sample sizes				

$ssA = br\,\Sigma_i \bar{y}^2_{i..} - abr\bar{y}^2_{...}$ $\qquad\qquad ssB(A) = r\,\Sigma_i \Sigma_j \bar{y}^2_{ij.} - br\,\Sigma_i \bar{y}^2_{i..}$

$ssE = \Sigma_i \Sigma_j \Sigma_t y^2_{ijt} - r\,\Sigma_i \Sigma_j \bar{y}^2_{ij.}$ $\qquad\qquad sstot = \Sigma_i \Sigma_j \Sigma_t y^2_{ijt} - abr\bar{y}^2_{...}$

where α_i is the effect of the ith machine, $B_{j(i)}$ is the effect of the jth randomly selected head on the ith machine, and $\sigma^2_{B(A)}$ is the variance of responses from the population of machine heads that could be fitted on these five machines. The error assumptions are examined in the same way as in Chap. 5 for the one-way analysis of variance model. More sophisticated techniques of checking other assumptions on a mixed model with nested effects are discussed by Beckman et al. (1987).

The degrees of freedom and sums of squares for the fixed-effects two-way nested model were calculated in Sect. 18.3. These are listed, for equal sample sizes, in Table 18.2.

We first verify that the fixed-effects mean square for error also provides an unbiased estimate for σ^2 in the mixed effects two-way nested model. From Sect. 18.3.2, we know that $MSE = SSE/(ab(r-1))$, where

$$SSE = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{t=1}^{r} Y^2_{ijt} - r \sum_{i=1}^{a} \sum_{j=1}^{b} \bar{Y}^2_{ij.}\,.$$

For the mixed-effects two-way nested model (18.2.2), we have

$$E[Y_{ijt}] = E[\bar{Y}_{ij.}] = E[\bar{Y}_{i..}] = \mu + \alpha_i \quad \text{and } E[\bar{Y}_{...}] = \mu + \bar{\alpha}_.\,.$$

Also,

$$\mathrm{Var}(Y_{ijt}) = \sigma^2_{B(A)} + \sigma^2$$

and

$$\mathrm{Var}(\bar{Y}_{ij.}) = \mathrm{Var}\left(\mu + \alpha_i + B_{j(i)} + \frac{1}{r}\sum_{t=1}^{r} \epsilon_{ijt}\right) = \sigma^2_{B(A)} + \frac{\sigma^2}{r}\,.$$

So,

$$\begin{aligned}
E[SSE] &= \left[abr(\sigma^2_{B(A)} + \sigma^2) + br\sum_i(\mu + \alpha_i)^2\right] \\
&\quad - \left[abr\left(\sigma^2_{B(A)} + \frac{\sigma^2}{r}\right) + br\sum_i(\mu + \alpha_i)^2\right] \\
&= ab(r-1)\sigma^2\,.
\end{aligned}$$

So, $E[MSE] = \sigma^2$ as required. We also have that

$$\text{Var}(\overline{Y}_{i..}) = \text{Var}\left(\mu + \alpha_i + \frac{1}{b}\sum_{j=1}^{b} B_{j(i)} + \frac{1}{br}\sum_{j=1}^{b}\sum_{t=1}^{r} \epsilon_{ijt}\right)$$

$$= \frac{\sigma_{B(A)}^2}{b} + \frac{\sigma^2}{br}.$$

Similarly,

$$\text{Var}(\overline{Y}_{...}) = \frac{\sigma_{B(A)}^2}{ab} + \frac{\sigma^2}{abr}.$$

Consequently, the expected value of the sum of squares for A is

$$E[SSA] = E\left[br\sum_{i=1}^{a} \overline{Y}_{i..}^2 - abr\overline{Y}_{...}^2\right]$$

$$= \left[abr\left(\frac{\sigma_{B(A)}^2}{b} + \frac{\sigma^2}{br}\right) + br\sum_i (\mu + \alpha_i)^2\right]$$

$$- \left[abr\left(\frac{\sigma_{B(A)}^2}{ab} + \frac{\sigma^2}{abr}\right) + abr\sum_i (\mu + \overline{\alpha}_.)^2\right]$$

$$= r(a-1)\sigma_{B(A)}^2 + (a-1)\sigma^2 + br\sum_i (\alpha_i - \overline{\alpha}_.)^2.$$

Then, since $MSA = SSA/(a-1)$, we have

$$E[MSA] = \frac{br}{a-1}\sum_i (\alpha_i - \overline{\alpha}_.)^2 + r\sigma_{B(A)}^2 + \sigma^2$$

$$= Q(\alpha_i) + r\sigma_{B(A)}^2 + \sigma^2.$$

Similarly, the expected value of the sum of squares for B nested within A is

$$E[SSB(A)] = E\left[r\sum_{i=1}^{a}\sum_{j=1}^{b} \overline{Y}_{ij.}^2 - br\sum_{i=1}^{a} \overline{Y}_{i..}^2\right]$$

$$= \left[abr\left(\sigma_{B(A)}^2 + \frac{\sigma^2}{r}\right) + br\sum_i (\mu + \alpha_i)^2\right]$$

$$- \left[abr\left(\frac{\sigma_{B(A)}^2}{b} + \frac{\sigma^2}{br}\right) + br\sum_i (\mu + \alpha_i)^2\right]$$

$$= ar(b-1)\sigma_{B(A)}^2 + a(b-1)\sigma^2,$$

and, since $MSB(A) = SSB(A)/(a(b-1))$, we have

$$E[MSB(A)] = r\sigma_{B(A)}^2 + \sigma^2.$$

These expected mean squares are listed in the last column of Table 18.2, and we may verify that they can all be obtained from rule 17 of Chap. 17. This rule, which applies also to more complicated mixed-effects nested models, says

17. To obtain the expected mean square for a particular main effect or interaction, first make a note of the subscripts on the term representing that particular effect in the model. Write down variance components for the effect of interest, for the error, and for every interaction whose term in the model includes the noted set of subscripts. Gather up all variance components corresponding to fixed effects into one quadratic form Q. Multiply any remaining variance component except σ^2 by the number of observations taken on each level or combination of levels of the corresponding effect (main effect or interaction). Add up the terms.

18.4.2   Estimation of Variance Components

The rules for obtaining confidence intervals for fixed effects or variance components also remain the same as those in Chap. 17 for non-nested models. Thus, we may obtain a confidence interval for a variance component in a mixed-effects nested model as follows:

19. For a random effect, let $U = \Sigma k_i MS_i$ be the mean square or linear combination of mean squares whose expected value is equal to the variance component corresponding to the random effect. An exact or approximate $100(1 - \alpha)\%$ confidence interval for this variance component is

$$\left(\frac{xu}{\chi^2_{x,\alpha/2}} \; , \; \frac{xu}{\chi^2_{x,1-\alpha/2}} \right) ,$$

where

$$x = \frac{[\Sigma k_i (ms_i)]^2}{\Sigma [k_i (ms_i)]^2 / x_i} ,$$

and where u is the observed value of U, ms_i is the observed value of MS_i, and x_i is the number of degrees of freedom corresponding to ms_i.

For example, for the mixed-effects two-way nested model, we may estimate the variability of the response due to the effect of B within A as

$$u = \frac{msB(A) - msE}{r} .$$

Then, using rule 19, we can obtain a $100(1 - \alpha)\%$ confidence interval for $\sigma^2_{B(A)}$ as

$$\left(\frac{xu}{\chi^2_{x,\alpha/2}} \; , \; \frac{xu}{\chi^2_{x,1-\alpha/2}} \right) ,$$

where

$$x = \frac{(msB(A) - msE)^2}{\frac{msB(A)^2}{a(b-1)} + \frac{msE^2}{ab(r-1)}} .$$

18.4.3 Hypothesis Testing

Hypothesis testing rules are also obtained from the rules in Chap. 17:

18. To obtain the denominator of the test statistic for testing the null hypothesis that a main effect or interaction effect is zero, write down the expected mean square for the effect of interest (see rule 17). Cross out the term that would be zero if the null hypothesis were true. The denominator of the test statistic is the mean square, or linear combination of mean squares, u, whose expected value is equal to the remaining expression.

21. For a fixed effect, the decision rule for testing the hypothesis that the effect is zero is the same as that in rule 8, p. 210, for fixed-effects models except that msE is replaced by the denominator u from rule 18 and the number of error degrees of freedom is replaced by x in rule 19.

22. For a random effect, the decision rule for testing the hypothesis H_0 that the corresponding variance component is zero against the alternative hypothesis that it is not zero is

$$\text{reject } H_0 \text{ if } \frac{ms}{u} > F_{\nu,x,\alpha},$$

where ms is the mean square for the effect of interest and ν the corresponding degrees of freedom, u is the observed value of the denominator as in rule 18, and x is the corresponding degrees of freedom calculated as in rule 19.

For example, using the information in the expected mean squares column of Table 18.2 for the mixed-effects two-way nested model, the decision rule for testing the null hypothesis $H_0^{B(A)} : \{\sigma_{B(A)}^2 = 0\}$ of no variability in the effect of B within each level of A against the alternative hypothesis $H_A^{B(A)} : \{\sigma_{B(A)}^2 > 0\}$ is

$$\text{reject } H_0^{B(A)} \quad \text{if} \quad \frac{msB(A)}{msE} \quad > \quad F_{a(b-1),ab(r-1),\alpha}, \tag{18.4.7}$$

at chosen significance level α.

To test the hypothesis $H_0^A : \{\alpha_1 = \alpha_2 = \cdots = \alpha_a\}$ that the machine effects are the same averaged over their four heads, the decision rule at significance level α is

$$\text{reject } H_0^A \text{ if } \frac{msA}{msB(A)} > F_{a-1,a(b-1),\alpha}. \tag{18.4.8}$$

18.4.4 Some Examples

Example 18.4.1 Machine head experiment, continued

The data for the machine head experiment are listed in Table 18.3, and the analysis of variance table is shown in Table 18.4. One can show that the p-value for testing the hypothesis of no machine differences is 0.67, and we would conclude no difference in the effect on strain readings of the five machines. The test of the null hypothesis that the variance $\sigma_{B(A)}^2$ of the population of possible heads fitted to the machines is zero has p-value 0.065. Only if our choice of significance level is greater than this value would we conclude nonzero variability among the heads.

Table 18.3   Data for the machine head experiment

Mach.	Head 1				Head 2				Head 3				Head 4			
1	6	2	0	8	13	3	9	8	1	10	0	6	7	4	7	9
2	10	9	7	12	2	1	1	10	4	1	7	9	0	3	4	1
3	0	0	5	5	10	11	6	7	8	5	0	7	7	2	5	4
4	11	0	6	4	5	10	8	3	1	8	9	4	0	8	6	5
5	1	4	7	9	6	7	0	3	3	0	2	2	3	7	4	0

Source Hicks (1956). Copyright © 1956 American Society for Quality. Reprinted with permission

Table 18.4   Analysis of variance table for the machine head experiment

Effect	d.f.	Sum of squares	Mean square	Expected mean square	Ratio
Machine	4	45.075	11.2688	$Q(\alpha_i) + 4\sigma^2_{B(A)} + \sigma^2$	0.5975
Head (mach.)	15	282.875	18.8583	$4\sigma^2_{B(A)} + \sigma^2$	1.7625
Error	60	642.000	10.7000	σ^2	
Total	79	969.950			

An unbiased estimate of $\sigma^2_{B(A)}$ is given by

$$\frac{msB(A) - msE}{r} = \frac{18.8583 - 10.7000}{4} = 2.0396\,,$$

and since,

$$x = \frac{(2.0396)^2}{\frac{(18.8583/4)^2}{15} + \frac{(10.70/4)^2}{60}} = 2.598\,,$$

a 90% confidence interval for $\sigma^2_{B(A)}$ is given by

$$\left(\frac{(2.598)(2.0396)}{\chi^2_{2.598,.05}}\,,\ \frac{(2.598)(2.0396)}{\chi^2_{2.598,.95}}\right) \approx \left(\frac{5.299}{6.90}\,,\ \frac{5.299}{0.22}\right) = (0.77\,,\ 24.09)$$

measured in squared units of strain.   □

Example 18.4.2   Soil experiment

Consider an experiment to compare analyses of soil samples with four treatment factors A, B, C, and D, where

A is "method of analysis" and involves $a = 2$ specifically selected methods.
B is "laboratory" and involves $b = 4$ specifically selected labs.
C is "operator conducting the analysis" and there are $c = 3$ randomly selected operators in each lab.
D is "location from which soil was taken" and involves $d = 3$ randomly selected locations.

Suppose the model is

$$Y_{ijkut} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + C_{k(j)} + (\alpha C)_{ik(j)} + D_u$$
$$+ (\alpha D)_{iu} + (\beta D)_{ju} + (\alpha\beta D)_{iju} + \epsilon_{ijkut} ,$$

$$C_{k(j)} \sim N(0, \sigma^2_{C(B)}) ; \quad (\alpha C)_{ik(j)} \sim N(0, \sigma^2_{AC(B)}) ; \quad D_u \sim N(0, \sigma^2_D) ; \quad (\alpha D)_{iu} \sim N(0, \sigma^2_{AD})$$
$$(\beta D)_{ju} \sim N(0, \sigma^2_{BD}) ; \quad (\alpha\beta D)_{iju} \sim N(0, \sigma^2_{ABD}) ; \quad \epsilon_{ijkut} \sim N(0, \sigma^2)$$
$$i = 1, 2 ; \quad j = 1, 2, 3, 4 ; \quad k = 1, 2, 3 ; \quad u = 1, 2, 3 ; \quad t = 1, 2 ;$$

where α_i is the effect of the ith method of analysis, β_j is the effect of the jth laboratory, and $(\alpha\beta)_{ij}$ is the effect of their interaction; $C_{k(j)}$ is the effect of the kth randomly selected operator in the kth laboratory and $(\alpha C)_{ik(j)}$ is the operator $\times$ analysis method interaction; D_u is the effect of the uth randomly selected location from which the soil was selected and $(\alpha D)_{iu}$, $(\beta D)_{ju}$ and $(\alpha\beta D)_{iju}$ are respectively the interactions of the uth soil location with the ith method of analysis, the uth soil location with the jth laboratory, and the three-factor interaction of the uth soil location, ith method of analysis, and jth laboratory. Two observations are taken on each soil sample via each method of analysis by each operator.

The degrees of freedom, sums of squares, and expected mean squares for this model are obtained using rules 17–21 in Sects. 18.4.1–18.4.3 and are shown in Table 18.5.

The decision rule for testing the null hypothesis $H_0^{ABD} : \{\sigma^2_{ABD} = 0\}$ against the alternative hypothesis $H_A^{ABD} : \{\sigma^2_{ABD} > 0\}$ is given by

$$\text{reject } H_0^{ABD} \text{ if } \frac{msABD}{msE} > F_{6,104,\alpha} .$$

If this hypothesis is not rejected, we may wish to examine the AB, AD, and BD interactions. The decision rule for testing the null hypothesis $H_0^{BD} : \{\sigma^2_{BD} = 0\}$ against the alternative hypothesis $H_A^{BD} : \{\sigma^2_{BD} > 0\}$ is given by

$$\text{reject } H_0^{BD} \text{ if } \frac{msBD}{msABD} > F_{6,6,\alpha} .$$

The test for the AD interaction is similar, utilizing the test statistic $msAD/msABD$. To obtain a suitable denominator for testing

$$H_0^{AB} : \{(\alpha\beta)_{ij} - \overline{(\alpha\beta)}_{i.} - \overline{(\alpha\beta)}_{.j} + \overline{(\alpha\beta)}_{..} = 0, \quad \text{for all } i, j\}$$

against the alternative hypothesis that the interaction is not zero, we need the denominator of the test statistic to be an unbiased estimator for

$$6\sigma^2_{AC(B)} + 6\sigma^2_{ABD} + \sigma^2 .$$

Such an estimator is

$$U = MS(AC(B)) + MS(ABD) - MSE .$$

This has approximately a χ^2_x distribution with

Table 18.5 Degrees of freedom, sums of squares, and expected mean squares for the soil experiment

Effect	Degrees of freedom	Expected mean square
A	$a - 1 = 1$	$Q(\alpha, \alpha\beta) + 6\sigma^2_{AC(B)} + 24\sigma^2_{AD} + 6\sigma^2_{ABD} + \sigma^2$
B	$b - 1 = 3$	$Q(\beta, \alpha\beta) + 12\sigma^2_{C(B)} + 6\sigma^2_{AC(B)} + 12\sigma^2_{BD} + 6\sigma^2_{ABD} + \sigma^2$
AB	$(a - 1)(b - 1) = 3$	$Q(\alpha\beta) + 6\sigma^2_{AC(B)} + 6\sigma^2_{ABD} + \sigma^2$
$C(B)$	$b(c - 1) = 8$	$12\sigma^2_{C(B)} + 6\sigma^2_{AC(B)} + \sigma^2$
$AC(B)$	$(a - 1)b(c - 1) = 8$	$6\sigma^2_{AC(B)} + \sigma^2$
D	$d - 1 = 2$	$48\sigma^2_D + 24\sigma^2_{AD} + 12\sigma^2_{BD} + 6\sigma^2_{ABD} + \sigma^2$
AD	$(a - 1)(d - 1) = 2$	$24\sigma^2_{AD} + 6\sigma^2_{ABD} + \sigma^2$
BD	$(b - 1)(d - 1) = 6$	$12\sigma^2_{BD} + 6\sigma^2_{ABD} + \sigma^2$
ABD	$(a - 1)(b - 1)(d - 1) = 6$	$6\sigma^2_{ABD} + \sigma^2$
Error	subtraction $= 104$	σ^2
Total	$n - 1 = 143$	
Formulae		

$$ssA = 72\Sigma_i \bar{y}^2_{i...} - 144\bar{y}^2_{....}$$

$$ssB = 36\Sigma_j \bar{y}^2_{.j..} - 144\bar{y}^2_{....}$$

$$ssAB = 18\Sigma_i \Sigma_j \bar{y}^2_{ij..} - 72\Sigma_i \bar{y}^2_{i...} - 36\Sigma_j \bar{y}^2_{.j..} + 144\bar{y}^2_{....}$$

$$ssC(B) = 12\Sigma_j \Sigma_k \bar{y}^2_{.jk.} - 36\Sigma_j \bar{y}^2_{.j..}$$

$$ssAC(B) = 6\Sigma_i \Sigma_j \Sigma_k \bar{y}^2_{ijk.} - 18\Sigma_i \Sigma_j \bar{y}^2_{ij..} - 12\Sigma_j \Sigma_k \bar{y}^2_{.jk.} + 36\Sigma_j \bar{y}^2_{.j..}$$

$$ssD = 48\Sigma_u \bar{y}^2_{...u.} - 144\bar{y}^2_{....}$$

$$ssAD = 24\Sigma_i \Sigma_u \bar{y}^2_{i..u.} - 72\Sigma_i \bar{y}^2_{i...} - 48\Sigma_u \bar{y}^2_{...u.} + 144\bar{y}^2_{....}$$

$$ssBD = 12\Sigma_j \Sigma_u \bar{y}^2_{.j.u.} - 36\Sigma_j \bar{y}^2_{.j..} - 48\Sigma_u \bar{y}^2_{...u.} + 144\bar{y}^2_{....}$$

$$ssABD = 6\Sigma_i \Sigma_j \Sigma_u \bar{y}^2_{ij.u.} - 18\Sigma_i \Sigma_j \bar{y}^2_{ij..} - 24\Sigma_i \Sigma_u \bar{y}^2_{i..u.} - 12\Sigma_j \Sigma_u \bar{y}^2_{.j.u.}$$
$$+ 72\Sigma_i \bar{y}^2_{i...} + 36\Sigma_j \bar{y}^2_{.j..} + 48\Sigma_u \bar{y}^2_{...u.} - 144\bar{y}^2_{....}$$

$$sstot = \Sigma_i \Sigma_j \Sigma_k \Sigma_u \Sigma_t y^2_{ijkut} - 144\bar{y}^2_{....}$$

ssE is obtained by subtraction

$$x = \frac{[MS(AC(B)) + MS(ABD) - MSE]^2}{\frac{MSAC(B)^2}{8} + \frac{MS(ABD)^2}{6} + \frac{MSE^2}{104}} .$$

Thus the decision rule for testing H_0^{AB} against H_A^{AB} is

$$\text{reject } H_0^{AB} \text{ if } \frac{msAB}{U} > F_{3, x, \alpha} .$$

An unbiased estimate of σ^2_{BD} is

$$U = \frac{msBD - msABD}{12} .$$

This has approximately a χ^2_x distribution, where

$$x = \frac{u^2}{\frac{(msBD/12)^2}{6} + \frac{(msABD/12)^2}{6}} \,,$$

and an approximate 95% confidence interval for σ_{BD}^2 is

$$\left(u/\chi_{x,\alpha/2}^2, \ u/\chi_{x,1-\alpha/2}^2\right) \,.$$

Any one of the main effects A, B, or $C(B)$ can be investigated if the interactions involving the corresponding factor are all negligible. The relevant formulae can be obtained along the same lines as those described above. □

18.5   Using SAS Software

The SAS procedure PROC GLM can handle nested effects when they are described in the MODEL statement using notation of the form $B(A)$. The RANDOM statement is used to obtain expected mean squares. The procedure PROC MIXED, which was described briefly in Chap. 17, can also be used. We will illustrate these procedures via the experiment in Sect. 18.5.1.

18.5.1   Voltage Experiment

An experiment was described by David Desmond in the 1954 issue of *Applied Statistics* on reducing the variability of voltage regulators fitted to motor cars. The voltage regulator was required to operate within a range of 15.8–16.4 volts. When the experiment took place, records showed that about 18% of regulators required readjustment during inspection, and sometimes this figure rose to 50%. Despite the inspection procedure, some of the regulators reaching customers were still outside the specification limits, and complaints from customers were considered to be excessive.

The experiment was run in order to measure the variability in the regulator setting operation. Measurements were taken on 64 voltage regulators at each of four testing stations. The 64 regulators were selected at random from several different setting stations. In Table 18.6, we have reproduced the data for six of these setting stations, corresponding to 40 voltage regulators. Since the regulators were selected at random, we model their effects as random effects nested within setting station. For purposes of illustration, we consider the four testing stations and six setting stations as the only stations of interest and model them as fixed effects. In the original article, these were modeled as random effects.

The effect of testing station is crossed with the effect of setting station and with regulator. A model to describe the data can be written as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + C_{k(j)} + \epsilon_{ijk} \,,$$
$$\epsilon_{ijk} \sim N(0, \sigma^2) \,, \quad C_{k(j)} \sim N(0, \sigma_{C(B)}^2) \,,$$
$$\epsilon_{ijk}\text{'s and } C_{k(j)}\text{'s are all mutually independent}$$
$$i = 1, \ldots, 4; \quad j = 1, \ldots, 6; \quad k = 1, \ldots, r_j \,,$$

where α_i is the effect of the ith testing station, β_j is the effect of the jth setting station, and $C_{k(j)}$ is the effect of the kth randomly selected regulator from the jth setting station.

There is no reason to suspect that the testing stations would differ in their comparative results for different regulators, so there is no reason to expect a regulator × testing station interaction. Since there

Table 18.6 Voltages for the voltage experiment

Set. sta. (B)	Regulator (C)	Testing station (A)				Set. sta. (B)	Regulator (C)	Testing station (A)			
		1	2	3	4			1	2	3	4
1	1	16.5	16.5	16.6	16.6	4	1	16.1	16.0	16.0	16.2
	2	15.8	16.7	16.2	16.3		2	16.5	16.1	16.5	16.7
	3	16.2	16.5	15.8	16.1		3	16.2	17.0	16.4	16.7
	4	16.3	16.5	16.3	16.6		4	15.8	16.1	16.2	16.2
	5	16.2	16.1	16.3	16.5		5	16.2	16.1	16.4	16.2
	6	16.9	17.0	17.0	17.0		6	16.0	16.2	16.2	16.1
	7	16.0	16.2	16.0	16.0		7	16.0	16.0	16.1	16.0
	8	16.0	16.0	16.1	16.0	5	1	15.5	15.6	15.4	15.8
2	1	16.0	16.1	16.0	16.1		2	15.8	16.2	16.0	16.2
	2	*15.4*	16.4	16.8	16.7		3	16.2	*15.4*	16.1	16.3
	3	16.1	16.4	16.3	16.3		4	16.2	16.2	16.0	16.1
	4	15.9	16.1	16.0	16.0		5	16.1	16.2	16.3	16.2
							6	16.1	16.1	16.0	16.1
3	1	16.0	16.0	15.9	16.3	6	1	15.5	15.5	15.3	15.6
	2	15.8	16.0	16.3	16.0		2	16.0	15.6	15.7	16.2
	3	15.7	16.2	15.3	15.8		3	16.0	16.4	16.2	16.2
	4	16.2	16.4	16.4	16.6		4	15.8	16.5	16.2	16.2
	5	16.0	16.1	16.0	15.9		5	15.9	16.1	15.9	16.0
	6	16.1	16.1	16.1	16.1		6	15.9	16.1	15.8	15.7
	7	16.1	16.0	16.1	16.0		7	16.0	16.4	16.0	16.0
							8	16.1	16.2	16.2	16.1

Source Desmond (1954). Copyright © 1956 Blackwell Publishers. Reprinted with permission

is only one observation per regulator–testing station combination, we would not be able to distinguish such an interaction from experimental error. A SAS program for analyzing this model is shown in Table 18.7.

PROC GLM

A plot of the standardized residuals (not shown) highlights two rather large outliers. The two outlying observations are those highlighted in italics in Table 18.6, and we notice that they are from different regulators and different testing stations. If these outliers are removed, the output shown in Fig. 18.1 is obtained. The TEST option produces the correct denominators for the tests of $H_0^A : \{\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4\}$, $H_0^B : \{\beta_1 = \beta_2 = \cdots = \beta_6\}$ and $H_0^{C(B)} : \{\sigma^2_{C(B)} = 0\}$. If we select an overall significance level of $\alpha = 0.06$ for the three tests and do each test at level $\alpha^* = 0.02$, we see that there is a significant difference between testing stations, but not between setting stations. Also, the variance of the regulators within setting stations appears to be significantly different from zero. Mind you, the F-test for setting stations is somewhat approximate, not only because the denominator is a composite variance estimator, but also because the treatment type III mean squares may be slightly dependent as a consequence of the removal of the two outliers, causing a slight dependence of the numerator and denominator of the F-statistic.

Unbiased estimates of σ^2 and $\sigma^2_{C(B)}$ can be obtained from the listed expected mean squares as $\hat{\sigma}^2 = msE = 0.0268$ and

Table 18.7   SAS program to analyze a mixed-effects nested model

```
DATA VLT;
 * Input setting station (B), regulator (C),
 *       testing station (A), and voltage;
 INPUT B C A VOLTG;
 LINES;
  1 1 1 16.5
  1 1 2 16.5
  : : :   :
  6 8 4 16.1
;
* Plot standardized residuals versus predicted values for all data;
PROC GLM;
  CLASS A B C;
  MODEL VOLTG = A B C(B);
  RANDOM C(B) / TEST;
  LSMEANS A / PDIFF = ALL CL ADJUST = TUKEY;
  LSMEANS B / PDIFF = ALL CL ADJUST = TUKEY E = C(B);
  OUTPUT OUT = RESIDS PREDICTED = PRED RESIDUAL = Z;
PROC STANDARD STD=1.0;
  VAR Z;
PROC PLOT; * or use PROC SGPLOT;
  PLOT Z*PRED = A Z*PRED = B Z*PRED = C / VREF = 0 VPOS = 19 HPOS = 50;
* Analysis without two outliers;
DATA VLT2; SET VLT;
  IF B = 2 AND C = 2 AND A = 1 THEN DELETE;
  IF B = 5 AND C = 3 AND A = 2 THEN DELETE;
PROC GLM;
  CLASS A B C;
  MODEL VOLTG = A B C(B);
  RANDOM C(B) / TEST;
  LSMEANS A / PDIFF = ALL CL ADJUST = TUKEY;
  * The following should be approximately correct;
  LSMEANS B / PDIFF = ALL CL ADJUST = TUKEY E = C(B);
PROC MIXED METHOD = TYPE3;
  CLASS A B C;
  MODEL VOLTG = A B / DDFM = SAT;
  RANDOM C(B);
  LSMEANS A B / CL PDIFF ADJUST = BON;
```

$$\hat{\sigma}_{C(B)}^2 = \frac{msC(B) - msE}{3.9499} = \frac{0.2405 - 0.0268}{3.9499} = 0.0541 \,,$$

respectively. Thus the variability of the regulator strain readings is estimated to be about twice as large as the experimental error.

A 90% confidence interval for $\sigma_{C(B)}^2/\sigma^2$ can be obtained by adapting the formula (17.3.11) as follows:

$$\frac{1}{c}\left[\frac{msC(B)}{msE\ F_{\nu_1,\nu_2,\alpha/2}} - 1\right] \ \le\ \frac{\sigma_{C(B)}^2}{\sigma^2} \le\ \frac{1}{c}\left[\frac{msC(B)}{msE\ F_{\nu_1,\nu_2,1-\alpha/2}} - 1\right]$$

$$= \ \frac{1}{c}\left[\frac{0.2405}{0.0268\ F_{34,115,0.05}} - 1\right] \ \le\ \frac{\sigma_{C(B)}^2}{\sigma^2} \le\ \frac{1}{c}\left[\frac{0.2405}{0.0268\ F_{34,115,0.95}} - 1\right]\,.$$

Fig. 18.1 SAS output for the voltage experiment
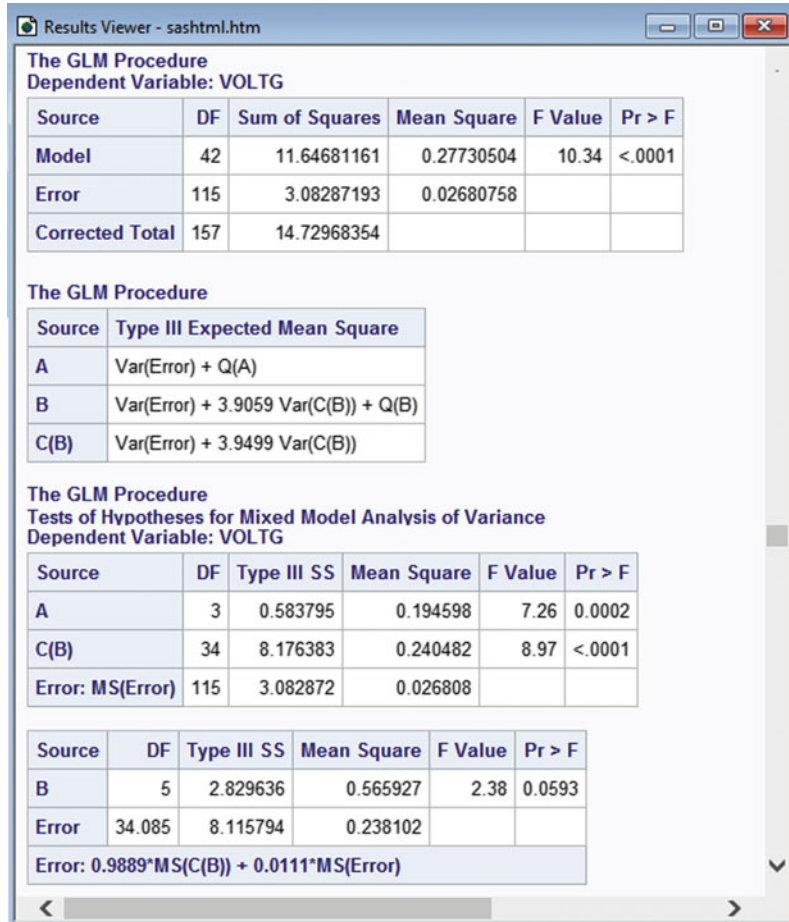
Since $E[MSC(B)] = \sigma^2 + 3.9499\sigma^2_{C(B)}$, the value of $c = $ is 3.9499. So, using $F_{34,115,0.05} \approx 1.52$ and $F_{34,115,0.95} = (F_{115,34,0.05})^{-1} \approx 1.64^{-1} = 0.61$, the confidence interval becomes
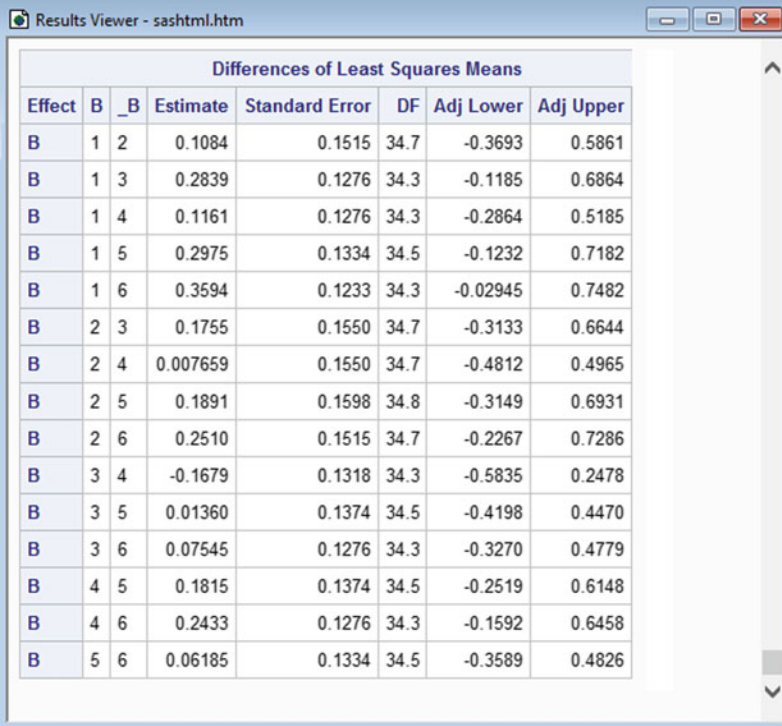
$$1.242 \leq \frac{\sigma^2_{C(B)}}{\sigma^2} \leq 3.471 \,.$$

The general conclusion of the experiment was that the differences between the four testing stations were of little practical importance. However, we note that the residual plots still indicate one or two large residuals, especially from testing station 2, so perhaps testing station 2 should have been examined a little more closely.

Much of the variability in the regulators appeared to be due to the inherent measurement error, and the experimenters concluded that it was not possible to set the regulators within the desired tolerance limits. A quality control scheme to ensure that the current quality did not deteriorate was put in place.

We note in passing that the effect of the outliers on the analysis was actually very small. If the two original outliers had been included in the analysis, the estimates $\hat{\sigma}^2 = 0.0268$ and $\hat{\sigma}^2_{C(B)} = 0.0541$ would have changed to 0.0392 and 0.0461, respectively. There would also be little change in the p-values of the hypothesis tests. There is some benefit in retaining the entire data set, since the coefficient of $\sigma^2_{C(B)}$ in the expected mean squares is then 4.0, as stated by rule 17 on p. 637.

Fig. 18.2 Output from
PROC MIXED for the
voltage experiment

PROC MIXED

The model can also be analyzed using the SAS procedure PROC MIXED and the analysis of variance approach as in Chap. 17. The SAS statements are shown in Table 18.7. The analysis of variance output from PROC MIXED (not shown) would match that generated by PROC GLM, because the option METHOD = TYPE3 implements the same least squares fit and the same analysis based on Type III sums of squares. An advantage of PROC MIXED is that it correctly estimates standard errors for means and contrasts. For example, some of the output generated by the LSMEANS statement for comparing setting stations (B) is shown in Fig. 18.2. Note that the standard error and associated degrees of freedom depend on the levels compared, due to the data imbalance caused by the removal of the two outliers. Composite variance estimates are used, and the degrees of freedom are obtained via Satterthwaite's approximation, due to the option DDFM = SAT in the MODEL statement. The changing number of degrees of freedom from one comparison to another indicates that the corresponding variance estimator also changes. Consequently, the Bonferroni method is used, since Tukey's method requires a common variance estimator, though the latter should also be approximately correct for such nearly balanced data. If the data were more than a little imbalanced, it would be preferable to use restricted maximum likelihood in PROC MIXED for variance components estimation—an approach to be discussed in Chap. 19.

18.6   Using R Software

The analysis of variance approach can be used in R for the analysis of balanced designs involving random and nested effects. The aov function can fit such models. For example, if a model as specified in R includes the terms A and A:B but not B, then A:B represents the effects of B nested within A.

Equivalently, the notation A/B causes inclusion of the terms A and A:B if B is excluded. Random effects are designated by inclusion of a single Error function in the model. For example, if the model includes the terms A and Error(A:B) but excludes the term B, then A:B represents random effects of *B* nested within *A*. The aov function fits models by least squares, the summary function provides the corresponding analysis of variance, including the usual (sometimes approximate) *F* tests for any fixed effects in the model, and the lsmeans function implements multiple comparison procedures. This analysis of variance approach using aov is appropriate and the computations dependable given a balanced design.

For unbalanced designs involving random effects, the data analysis can be accomplished by alternative methods involving estimation of the variance components by *restricted maximum likelihood* (ReML). This approach can be implemented using the lmer function of the lme4 package to fit the model, the anova function to generate tests of fixed effects, and the lsmeans function for multiple comparisons. In our programs, we call the lmerTest package rather than lme4, as the former provides *p*-values for *F*-tests of fixed effects.

We will illustrate the above approaches in Sects. 18.6.2 and 18.6.3, respectively, using the experiment introduced in the following section.

18.6.1 Voltage Experiment

An experiment was described by David Desmond in the 1954 issue of *Applied Statistics* on reducing the variability of voltage regulators fitted to motor cars. The voltage regulator was required to operate within a range of 15.8–16.4 volts. When the experiment took place, records showed that about 18% of regulators required readjustment during inspection, and sometimes this figure rose to 50%. Despite the inspection procedure, some of the regulators reaching customers were still outside the specification limits, and complaints from customers were considered to be excessive.

The experiment was run in order to measure the variability in the regulator setting operation. Measurements were taken on 64 voltage regulators at each of four testing stations. The 64 regulators were selected at random from several different setting stations. In Table 18.6 (p. 688), we have reproduced the data for six of these setting stations, corresponding to 40 voltage regulators. Since the regulators were selected at random, we model their effects as random effects nested within setting station. For purposes of illustration, we consider the four testing stations and six setting stations as the only stations of interest and model them as fixed effects. In the original article, these were modeled as random effects.

The effect of testing station is crossed with the effect of setting station and with regulator. A model to describe the data can be written as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + C_{k(j)} + \epsilon_{ijk}, \tag{18.6.9}$$
$$\epsilon_{ijk} \sim N(0, \sigma^2), \quad C_{k(j)} \sim N(0, \sigma^2_{C(B)}),$$
$$\epsilon_{ijk}\text{'s and } C_{k(j)}\text{'s are all mutually independent}$$
$$i = 1, \ldots, 4; \quad j = 1, \ldots, 6; \quad k = 1, \ldots, r_j,$$

where α_i is the effect of the ith testing station, β_j is the effect of the jth setting station, and $C_{k(j)}$ is the effect of the kth randomly selected regulator from the jth setting station.

There is no reason to suspect that the testing stations would differ in their comparative results for different regulators, so there is no reason to expect a regulator × testing station interaction. Since there

Table 18.8 R program and selected output for analysis of a mixed-effects nested model by the analysis of variance approach

```
> voltage.data = read.table("data/voltage.txt", header=T)
> voltage.data = within(voltage.data, {fSetting = factor(Setting);
+                 fRegul = factor(Regul); fTesting = factor(Testing) })
> head(voltage.data, 3)
  Setting Regul Testing Voltg fTesting fRegul fSetting
1       1     1       1  16.5        1      1        1
2       1     1       2  16.5        2      1        1
3       1     1       3  16.6        3      1        1

> # Least squares ANOVA
> # Set contrast options for correct lsmeans and contrasts
> options(contrasts = c("contr.sum", "contr.poly"))
> model1 = aov(Voltg ~ fSetting + fTesting + Error(fSetting:fRegul),
+              data=voltage.data)
> summary(model1)

Error: fSetting:fRegul
          Df Sum Sq Mean Sq F value Pr(>F)
fSetting   5   2.91   0.582     2.6  0.043
Residuals 34   7.61   0.224

Error: Within
           Df Sum Sq Mean Sq F value Pr(>F)
fTesting    3   0.70  0.2341    5.97 0.0008
Residuals 117   4.59  0.0392

> # Multiple comparisons: Tukey's method
> library(lsmeans)
> lsmTesting1 = lsmeans(model1, ~ fTesting)
> summary(contrast(lsmTesting1, method="pairwise", adjust="tukey"),
+         infer=c(T,T), level=0.98, side="two-sided")

 contrast estimate      SE  df  lower.CL  upper.CL t.ratio p.value
 1 - 2     -0.1550 0.044267 117 -0.285431 -0.024569  -3.502  0.0036
 1 - 3     -0.0825 0.044267 117 -0.212931  0.047931  -1.864  0.2494
 1 - 4     -0.1650 0.044267 117 -0.295431 -0.034569  -3.727  0.0017
 2 - 3      0.0725 0.044267 117 -0.057931  0.202931   1.638  0.3616
 2 - 4     -0.0100 0.044267 117 -0.140431  0.120431  -0.226  0.9959
 3 - 4     -0.0825 0.044267 117 -0.212931  0.047931  -1.864  0.2494

Results are averaged over the levels of: fSetting
Confidence level used: 0.98
Conf-level adjustment: tukey method for comparing a family of 4 estimates
P value adjustment: tukey method for comparing a family of 4 estimates
```

is only one observation per regulator–testing station combination, we would not be able to distinguish such an interaction from experimental error.

18.6.2   Analysis Using Least Squares Estimates and `aov`

Table 18.8 contains an R program for analyzing model (18.6.9) using the `aov` function. The `aov` function, which fits models by ordinary least squares and takes an analysis of variance approach to the data analysis, works fine for models including random effects as long as the design is balanced. The `summary` function generates the appropriate F-tests for each fixed effect, including correct denominators for the tests of $H_0^A : \{\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4\}$ and $H_0^B : \{\beta_1 = \beta_2 = \cdots = \beta_6\}$. If we conduct both tests of fixed effects at level $\alpha^* = 0.02$, we see that there is a significant difference between testing stations ($p = 0.0008$) but not between setting stations ($p = 0.043$). Tukey's method is illustrated for comparing the effects of testing station. Code for comparing the effects of setting station is analogous.

Tests for random effects are not generated by the `aov` and `summary` functions. However, for balanced data, the appropriate tests can be constructed by hand from the mean squares and degrees of freedom generated by the `summary` function, based on the corresponding expected mean squares. Using rule 17 for estimation and hypothesis testing (p. 647), one can show that the expect mean square for regulators nested within setting is $\sigma^2 + 4\,\sigma_{C(B)}^2$. So, to test $H_0^{C(B)} : \{\sigma_{C(B)}^2 = 0\}$, the appropriate test statistic is $F = msC(B)/msE = 0.224/0.0392 = 5.714$ with 34 and 117 degrees of freedom, the mean square and degrees of freedom values being obtained from Table 18.8. The reader may verify that the null hypothesis would be rejected at level $\alpha^* = 0.02$ for example ($p < 0.001$), corresponding to an overall significance level of $\alpha = 0.06$ for the three tests.

Unbiased estimates of σ^2 and $\sigma_{C(B)}^2$ can be obtained as $\hat{\sigma}^2 = msE = 0.0392$ and

$$\hat{\sigma}_{C(B)}^2 \;=\; \frac{msC(B) - msE}{4} \;=\; \frac{0.224 - 0.0392}{4} \;=\; 0.0462\,,$$

respectively. Thus the variability of the regulator strain readings is estimated to be only slightly larger than the experimental error.

A 90% confidence interval for $\sigma_{C(B)}^2/\sigma^2$ can be obtained by adapting the formula (17.3.11) as follows:

$$\frac{1}{c}\left[\frac{msC(B)}{msE\ F_{\nu_1,\nu_2,\alpha/2}} - 1\right] \;\leq\; \frac{\sigma_{C(B)}^2}{\sigma^2} \leq\; \frac{1}{c}\left[\frac{msC(B)}{msE\ F_{v-1,n-v,1-\alpha/2}} - 1\right]$$

$$=\; \frac{1}{c}\left[\frac{0.224}{0.0462\ F_{34,117,0.05}} - 1\right] \;\leq\; \frac{\sigma_{C(B)}^2}{\sigma^2} \leq\; \frac{1}{c}\left[\frac{0.224}{0.0462\ F_{34,117,0.95}} - 1\right]\,.$$

Since $E[MSC(B)] = \sigma^2 + 4\sigma_{C(B)}^2$, the value of $c =$ is 4. So, using $F_{34,117,0.05} \approx 1.533$ and $F_{34,117,0.95} = (F_{117,34,0.05})^{-1} \approx 1.635^{-1} = 0.6116$, the confidence interval becomes

$$0.541 \leq \frac{\sigma_{C(B)}^2}{\sigma^2} \leq 1.732\,.$$

The general conclusion of the experiment was that the differences between the four testing stations were of little practical importance. However, we note that the residual plots still indicate one or two large residuals, especially from testing station 2, so perhaps testing station 2 should have been examined a little more closely.

Much of the variability in the regulators appeared to be due to the inherent measurement error, and the experimenters concluded that it was not possible to set the regulators within the desired tolerance limits. A quality control scheme to ensure that the current quality did not deteriorate was put in place.

One advantage to the `aov` function and ordinary least squares is that residuals are available for checking model assumptions. A plot of the standardized residuals (not shown) highlights two rather large outliers. The two outlying observations are those highlighted in italics in Table 18.6, and we notice

that they are from different regulators and different testing stations. If these outliers were removed, one would need to use different methods for the data analysis, as illustrated in the next section.

18.6.3   Analysis Using Restricted Maximum Likelihood Estimation

Whether or not the design is balanced, model (18.6.9) can also be analyzed using restricted maximum likelihood (ReML) estimation. In particular, the variance components are estimated by restricted maximum likelihood, providing estimates which make the observed data most likely, subject to the restriction that the variance component estimates be non-negative. Given these variance component estimates, estimated generalized least squares estimates are computed for the fixed effects—generalized to take into account the unequal variances of observations as well as their correlations, and estimated since this variance-covariance structure is estimated. This approach will provide the same results as the analysis of variance approach if the design is balanced and all variance component estimates are positive, as is true for the voltage experiment. For further information about this approach, see Sect. 19.8.3.

Table 18.9 contains an R program and selected output for analyzing model (18.6.9) via this ReML-based approach, but excluding the two outliers. After the program reads all of the voltage data into the data set `voltage.data`, a new data set `voltage2.data` is created from it by taking the subset of `voltage.data` that satisfies two conditions that exclude the two outliers. For example, the first condition

```
!(fSetting == 2   &   fRegul == 2   &   fTesting == 1)
```

means not (!) to include observations with `fSetting` value 2 and (&) `fRegul` value 2 and `fTesting` value 1, thereby excluding the first outlier. The second outlier is similarly excluded by the second condition.

The `lmer` function fits model (18.6.9), estimating the variance components by restricted maximum likelihood estimation and the fixed effects by estimated generalized least squares estimation. In the model, specified as

```
Voltg ~ fSetting + fTesting + (1|fSetting:fRegul),
```

the term `(1|fSetting:fRegul)` causes inclusion of the random effects $C_{k(j)}$—one parameter for each combination of `fSetting` and `fRegul`. The `anova` function generates type 3 F-tests of the fixed effects—namely, for the effects of `fSetting` and `fTesting`. Finally, the `lsmeans` command applies Tukey's method for each fixed-effects factor using a 98% confidence level, though the results are only displayed for testing stations.

We note in passing that the effect of the outliers on the analysis was actually very small. Comparing the results in Table 18.8 with the outliers to those in Table 18.9 without the outliers, there is little change in the p-values of the hypothesis tests, and Tukey's method yields the same significant comparisons. There is some benefit in retaining the entire voltage data set, since the analysis of variance approach can be used in R for models involving random effects if the design is balanced, in which case the analysis of variance approach is statistically efficient and provides a more complete data analysis.

Exercises

1. **Viscosity experiment**

   An experiment was described by Johnson and Leone (1977, p. 744) to determine the viscosity of a polymeric material. The material was divided into two samples. The two samples were each divided

Table 18.9 R program and selected output for analysis of a mixed-effects nested model using restricted maximum likelihood estimation

```
> voltage.data = read.table("data/voltage.txt", header=T)
> voltage.data = within(voltage.data, {fSetting = factor(Setting);
+                 fRegul = factor(Regul); fTesting = factor(Testing) })

> # Drop two outliers, then reanalyze the data
> voltage2.data = subset(voltage.data,
+                     !(fSetting == 2 & fRegul == 2 & fTesting == 1)
+                   & !(fSetting == 5 & fRegul == 3 & fTesting == 2) )

> # REML
> # install.packages("lmerTest")
> library(lmerTest) # Attaches/masks lmer and lsmeans,
>                   # adding p-values to anova()
> model2 = lmer(Voltg ~ fSetting + fTesting + (1|fSetting:fRegul),
+               data=voltage2.data)
> anova(model2) # F-tests for fixed effects

Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
         Sum Sq Mean Sq NumDF DenDF F.value  Pr(>F)
fSetting  0.312  0.0623     5    34    2.32 0.06419
fTesting  0.590  0.1967     3   115    7.33 0.00015

> # Multiple comparisons
> library(lsmeans)
> lsmTesting2 = lsmeans(model2, ~ fTesting)
> summary(contrast(lsmTesting2, method="pairwise", adjust="tukey"),
+         infer=c(T,T), level=0.98, side="two-sided")

 contrast  estimate       SE     df   lower.CL  upper.CL t.ratio p.value
 1 - 2    -0.149846 0.037228 115.22 -0.259570 -0.040121  -4.025  0.0006
 1 - 3    -0.055856 0.036922 115.10 -0.164681  0.052968  -1.513  0.4332
 1 - 4    -0.138356 0.036922 115.10 -0.247181 -0.029532  -3.747  0.0016
 2 - 3     0.093989 0.036921 115.12 -0.014833  0.202811   2.546  0.0583
 2 - 4     0.011489 0.036921 115.12 -0.097333  0.120311   0.311  0.9895
 3 - 4    -0.082500 0.036618 115.00 -0.190429  0.025429  -2.253  0.1154

Results are averaged over the levels of: fSetting
Confidence level used: 0.98
Conf-level adjustment: tukey method for comparing a family of 4 estimates
P value adjustment: tukey method for comparing a family of 4 estimates

> lsmSetting2 = lsmeans(model2, ~ fSetting)
> summary(contrast(lsmSetting2, method="pairwise", adjust="tukey"),
+         infer=c(T,T), level=0.98, side="two-sided")
```

into ten "aliquots." After preparation of these aliquots, they were divided into two subaliquots and a further step in the preparation made. Finally, each subaliquot was divided into two parts and the final step of the preparation made. The viscosity determinations are listed in Table 18.10.

Table 18.10 Viscosity determinations for the viscosity experiment

| Sample | Aliquot | Subaliquot 1 | | Subaliquot 2 | |
		Part 1	Part 2	Part 1	Part 2
1	1	59.8	59.4	58.2	63.5
	2	66.6	63.9	61.8	62.0
	3	64.9	68.8	66.3	63.5
	4	62.7	62.2	62.9	62.8
	5	59.5	61.0	54.6	61.5
	6	69.0	69.0	60.6	61.8
	7	64.5	66.8	60.2	57.4
	8	61.6	56.6	64.5	62.3
	9	64.5	61.3	72.7	72.4
	10	65.2	63.9	60.8	61.2
2	1	59.8	61.2	60.0	65.0
	2	65.0	65.8	64.5	64.5
	3	65.0	65.2	65.5	63.5
	4	62.5	61.9	60.9	61.5
	5	59.8	60.9	56.0	57.2
	6	68.8	69.0	62.5	62.0
	7	65.2	65.6	61.0	59.3
	8	59.6	58.5	62.3	61.5
	9	61.0	64.0	73.0	71.7
	10	65.0	64.0	62.0	63.0

Source Johnson and Leone (1977). Copyright © 1977 Johnson and Leone. Reprinted with permission

(a) Write down a model for the viscosity determinations allowing for variability in the samples, aliquots, subaliquots and parts.
(b) Examine the error assumptions on your model.
(c) Estimate the variances of all the random effects in the model.
(d) Give a set of confidence intervals for the variances of all the random effects in the model at overall significance level 90%. At which step of the preparation is most of the variability introduced?

2. **Sleep experiment**

Sleeping patterns can be classified according to periods of "deep sleep" and of "REM sleep" (rapid eye movement). An experiment is done to see how sleeping tablets and amount of daily activity affect the proportion of REM sleep. Three types of sleeping tablets are to be tested, coded 1, 2, 3 (where type 3 is a placebo).

Twelve subjects are selected at random from a large population and are assigned at random to the levels of A, four to each level. Each subject is assigned an activity level for the day, and the proportion of REM sleep is monitored during that night. The four activity levels are:

B1 = read quietly all day,   B2 = walk 10 miles during the day,
B3 = spend the day shopping, B4 = play video games all day.

The experiment continues for four days, so that each subject is observed at each activity level in a random order. The model is assumed to be

$$Y_{hijt} = \mu + S_{h(i)} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{hijt} \,,$$

where α_i is the effect of the ith sleeping tablet, β_j is the effect of the jth activity level, $(\alpha\beta)_{ij}$ is the effect of their interaction, $S_{h(i)}$ is the effect of the hth random subject assigned to the ith sleeping tablet, and $S_{h(i)} \sim N(0, \sigma^2_{S(A)})$ and $\epsilon_{hijt} \sim N(0, \sigma^2)$ and all $S_{h(i)}$ and ϵ_{hijt} are independent.

(a) Write down the degrees of freedom and expected mean squares for the analysis of variance table.
(b) Explain how to test the null hypothesis $H_0^A : \{\alpha_1 = \alpha_2 = \alpha_3\}$ against the alternative hypothesis that at least two of the α_i differ.
(c) Explain how to test the null hypothesis $H_0^{S(A)} : \{\sigma^2_{S(A)} = 0\}$ against the alternative hypothesis $H_A^{S(A)} : \{\sigma^2_{S(A)} > 0\}$.
(d) Suppose that the null hypothesis

$$H_0^{AB} : \{(\alpha\beta)_{ij} - \overline{(\alpha\beta)}_{i.} - \overline{(\alpha\beta)}_{.j} + \overline{(\alpha\beta)}_{..}, \ \ \text{for all } i, \, j\}$$

appears to be correct. Which contrasts would be of particular interest to the experimenter? Why? Give formulas that would provide an overall 95% set of confidence intervals for your chosen contrasts. Give reasons for your choice of formula(s).
(e) If the experimenter thought that a day effect would be important, how would you modify the design of the experiment and the model?

3. Consider the model

$$Y_{ijkl} = \mu + \alpha_i + B_{j(i)} + C_{k(ji)} + \delta_l + (\alpha\delta)_{il} + (B\delta)_{lj(i)} + \epsilon_{ijkl} \,,$$

(a) Calculate the expected mean squares for all effects in the model.
(b) Which ratio would you use to test $H_0 : \{\delta_l + \overline{(\alpha\delta)}_{.l} \ \text{all equal}\}$?
(c) Which ratio would you use to test $H_0 : \sigma^2_A = 0$?

4. **Titanium alloy experiment**

An experiment described by Johnson and Leone (1977, p. 758) was performed by a company to investigate the effects of various factors on the "yield strength" of a particular titanium alloy. The factors investigated were:

A: vendors (4 fixed levels representing suppliers of raw material).
C: bar size (2 fixed levels representing standard sizes of bars of raw material).
B: batch (3 randomly selected levels nested within each combination of levels of A and C).
D: product type (2 fixed levels representing different types of finished product—forgedown and finished-forge blades).

Three observations were taken on each treatment combination. A reasonable model was thought to be

$$Y_{ijklt} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + B_{k(ij)} + \delta_l$$
$$+ (\alpha\delta)_{il} + (\gamma\delta)_{jl} + (B\delta)_{kl(ij)} + \epsilon_{ijklt} \,,$$
$$\epsilon_{ijklt} \sim N(0, \sigma^2) \,, \ \ B_{k(ij)} \sim N(0, \sigma^2_{B(AC)}) \,, \ \ (B\delta)_{kl(ij)} \sim N(0, \sigma^2_{BD(AC)}) \,,$$
$$i = 1, 2, 3, 4; \ \ j = 1, 2; \ \ k = 1, 2, 3; \ \ l = 1, 2; \ \ t = 1, 2, 3,$$

where α_i, γ_j, and δ_l represent the effects of the ith vendor, jth bar size, and lth product type, respectively, and $B_{k(ij)}$ represents the effect of the kth randomly selected batch of the jth bar size made with bar stock from the ith vendor, and random variables on the right-hand side of the model are assumed to be mutually independent.

(a) Write down the degrees of freedom and expected mean squares column of the analysis of variance table.

(b) Give a formula for an approximate 95% confidence interval for $\sigma^2_{B(AC)}$.

(c) How would you test the hypothesis

$$H_0 : \{\text{no differences in yield strength of the titanium alloy}$$
$$\text{can be attributed to the four vendors}\}$$

against the alternative hypothesis $H_A : \{H_0 \text{ is false}\}$?

5. **Titanium alloy experiment, continued**

Suppose that factors C and D are to be investigated further in a followup experiment. Suppose that two new factors P and Q ("heat setting during processing" and "cooling method") are also to be investigated at two levels each. A followup experiment is required with the four factors C, D, P, and Q at two levels each (a 2^4 experiment). Only sixteen observations will be taken, four for each vendor. It is known that the interactions CP, CQ, PQ, CPQ, and $CDPQ$ are likely to be negligible. Also, there was information gained from the previous parts to Exercise 4 to suggest that all interactions of treatment factors with vendor can be assumed negligible.

(a) Divide the 16 treatment combinations into four blocks of size four (one block for each vendor). Show your design explicitly, and indicate what should be randomized.

(b) Write down a suitable model and the degrees of freedom column for the analysis of variance table for your design in part (a).

(c) Before your design in part (a) is run, the management announces that in future, only one vendor will be used by the company. Also, your budget is cut, so that you can take only 8 observations. Thus, you need to design a $\frac{1}{2}$–fraction of a 2^4 experiment. In reviewing the list of negligible interactions above, you discover that two have been omitted. Interactions DP and CDQ are also known to be negligible. Choose a design and list the treatment combinations explicitly. (Hint: Try $I = CPQ$.) State the aliasing scheme and a suitable model. Will there be any problems in interpreting the results of this experiment?

6. **Operator experiment**

An experiment to identify the causes of variability in readings of a spectrometer was described in Exercise 10 of Chap. 7, p. 241. The same authors (Inman et al., *Journal of Quality Technology*, 1992) also described a study to determine how much of the variation in measured manganese concentration in steel was due to operator variation.

Ten steel samples were sliced from a steel billet. Each operator was asked to measure the manganese content of each sample twice. The measurements taken by any one operator were done in a random order on a single day. There were four operators, who were regarded as representative of a large population of potential operators.

(a) Write down a model for this experiment. Indicate clearly which effects are fixed, random, crossed, and nested.

Table 18.11 Manganese concentrations (percentages) for the operator experiment

Sample	Operator							
	1		2		3		4	
1	0.63	0.60	0.62	0.62	0.60	0.60	0.59	0.61
2	0.64	0.63	0.63	0.64	0.67	0.65	0.62	0.64
3	0.60	0.58	0.60	0.61	0.60	0.60	0.58	0.60
4	0.75	0.74	0.74	0.74	0.74	0.73	0.73	0.76
5	0.71	0.68	0.69	0.70	0.69	0.67	0.68	0.71
6	0.65	0.63	0.62	0.65	0.63	0.64	0.62	0.64
7	0.67	0.64	0.66	0.67	0.65	0.65	0.64	0.66
8	0.65	0.63	0.65	0.64	0.62	0.62	0.60	0.62
9	0.68	0.66	0.67	0.68	0.67	0.67	0.65	0.68
10	0.67	0.64	0.66	0.66	0.65	0.64	0.64	0.66

Source Inman, Ledolter, Lenth, and Niemi (1992). Reprinted with Permission from Journal of Quality Technology © 1992 ASQ, www.asq.org

(b) Write down the degrees of freedom, the sums of squares, and the expected mean squares for each of the sources of variation in your model.

(c) The authors analyzed this experiment using a gamma distribution to model the distribution of the error terms. Using the data in Table 18.11, investigate whether or not the normal distribution could be used (it may be necessary to take a transformation).

(d) If the normal distribution can be used as a reasonable approximation to the error distribution, then analyze the experiment. In particular, obtain estimates of the variances of the random effects and identify the major sources of variation.

7. For the two-way nested fixed-effects model (18.2.1) on p. 672, show that the least squares estimator of $\mu + \alpha_i + \beta_{j(i)}$ is given by $\overline{Y}_{ij.}$.

[Hint: Differentiate the sum of squared errors with respect to μ, α_i ($i = 1, \ldots, a$), and $\beta_{j(i)}$ ($j = 1, \ldots, b$; $i = 1, \ldots, a$), in turn. Set the resulting three sets of normal equations equal to zero. Show that the third set of equations adds to the first equation, and that the ith portion of the third set of equations adds to the ith equation in the second set. Thus, the first and second sets of equations are redundant, and $a + 1$ extra equations must be added to the set.]

8. **Red blood cell experiment**

The trout experiment reported by Gutsell (*Biometrics*, 1951) was described in Exercise 15 of Chap. 3. As part of the same experiment, the red blood cell counts in the blood of brown trout were measured. Fish were put at random into eight troughs of water. Two troughs were assigned to each of the four levels of the treatment factor "sulfamerazine" (0, 5, 10, 15 grams per 100 pounds of fish added to the diet per day). After 42 days, five fish were selected at random from each trough and the red blood cell count from the blood of each fish was measured in two different counting chambers, giving two measurements per fish. The observations reported in Table 18.12, when multiplied by 5000, give the number of red blood cells per cubic millimeter of blood.

A possible model for these data is

$$Y_{ijkt} = \mu + \alpha_i + B_{j(i)} + C_{k(ij)} + \epsilon_{ijkt},$$
$$\epsilon_{ijkt} \sim N(0, \sigma^2), \quad B_{j(i)} \sim N(0, \sigma_{B(A)}^2), \quad C_{k(ij)} \sim N(0, \sigma_{C(AB)}^2),$$
$$i = 1, 2, 3, 4; \quad j = 1, 2; \quad k = 1, \ldots, 5; \quad t = 1, 2;$$

Table 18.12 Red blood cell counts from brown trout for the red blood cell experiment

Fish	0 gm sulf.				5 gm sulf.			
	Trough 1		Trough 2		Trough 1		Trough 2	
1	213	230	166	157	296	319	310	309
2	253	231	206	185	278	258	241	270
3	195	164	245	250	345	307	272	311
4	193	203	213	181	322	372	254	237
5	191	195	198	169	248	274	266	275
Fish	10 gm sulf.				15 gm sulf.			
	Trough 1		Trough 2		Trough 1		Trough 2	
1	339	322	196	232	278	212	287	280
2	282	285	205	186	275	311	221	243
3	236	262	252	274	186	158	331	309
4	252	209	245	216	301	281	231	244
5	263	296	249	260	223	246	292	295

Source Gutsell (1951). Copyright © 1951 International Biometric Society. Reprinted with permission

where α_i is the effect of the ith level of sulfamerazine in the diet, $B_{j(i)}$ is the effect of the jth randomly selected trough assigned to the ith level of sulfamerazine, and C_k is the effect of the kth randomly selected fish from the (ij)th trough, and random variables on the right hand-side of the model are assumed to be mutually independent.

(a) What are the experimental units and observational units in this experiment?
(b) Since the data are counts, examine the assumptions of normally distributed errors and equal error variances by treatment. If the assumptions are not approximately satisfied, is there a transformation that can be used to correct the problem?
(c) Write out the degrees of freedom and the expected mean squares for each term in the model.
(d) Test the hypothesis that sulfamerazine has no effect on the red blood cell counts. Examine the linear and quadratic trends.
(e) If the test in part (d) is rejected, calculate simultaneous 95% confidence intervals for pairwise comparisons in the effects of the sulfamerazine levels.

Split-Plot Designs

<div style="text-align: right">**19**</div>

19.1 Introduction

Split-plot designs are needed when the levels of some treatment factors are more difficult to change during the experiment than those of others. The designs have a nested blocking structure. In a block design, the experimental units are nested within the blocks, and a separate random assignment of units to treatments is made within each block. In a split-plot design, the experimental units are called *split plots*, and are nested within *whole plots*, which themselves may or may not be nested within blocks.

The split plots within each whole plot are assigned at random to the levels of one or more of the treatment factors. The levels of other treatment factors are assigned to whole plots and remain constant for all split plots within a whole plot. Typically, these will be the factors whose levels are difficult to change, and the effects of their levels will be less precisely compared than those assigned to the split plots.

In Sect. 19.2 we show an example of an experiment designed as a split-plot design, together with a typical model for this type of design. The analysis of split-plot designs is discussed in Sect. 19.3 and illustrated via a second experiment. Designs with an extra level of nesting (split-split-plot designs) are briefly described in Sect. 19.4, and the issue of confounding treatment contrasts is introduced in Sect. 19.5. Section 19.6 introduces an experiment using a split-plot design without blocking. Section 19.7 introduces an experiment planned as a split-plot design with blocking, but that can also be viewed as a split-split-plot design. The use of SAS and R for analysis of split-plot designs is illustrated in Sects. 19.8 and 19.9, respectively.

19.2 Designs and Models

When a factorial experiment is run as a completely randomized design or a randomized complete block design, the levels of all the factors generally have to be changed frequently during the course of the experiment. For example, in Block I of the design in Table 13.13, p. 452, we see that as the experiment progressed on day 1, the level of the first factor had to be changed from 1 to 0 to 1 to 0 to 1, and the level of the second factor had to be changed from 0 to 1 to 0 to 1 to 0. The levels of the third and fourth factors also had to be changed four times. In most experiments this is no particular problem, but sometimes the level of one of the factors is *not* particularly easy to change.

An experiment is described by Munro in his 1986 University of Southampton dissertation on the effect of lighting conditions (factor *A*) and the speed of a rotating drum (factor *B*) on a subject's ability

Table 19.1 Part of a split-plot design for the rotating drum experiment

Block (Subject)	Whole plot 1 (Session 1)		Whole plot 2 (Session 2)	
	Level of A (Lighting)	Levels of B (Speed)	Level of A (Lighting)	Levels of B (Speed)
I	0	0 3 1 2	1	1 0 2 3
II	0	1 0 2 3	1	2 1 3 0
III	1	2 1 3 0	0	3 2 0 1
⋮	⋮	⋮ ⋮ ⋮ ⋮	⋮	⋮ ⋮ ⋮ ⋮
s	0	0 1 3 2	1	2 3 1 0

to focus on the center of the drum. In this experiment, it was easy to change the speed of rotation by the turn of a dial. The lighting conditions, however, took time to set up, and Munro wished to change these as seldom as possible. He therefore asked each subject to view all the rotation speeds (in a randomized order) under one set of lighting conditions during one session and return for a second session with different lighting conditions at a later date. Part of a possible design is shown in Table 19.1.

A whole plot is defined by a session for a particular subject. A split plot is defined by a time slot nested in a particular session for a particular subject. The two whole plots (sessions) within each block are assigned at random to the levels of one factor (A), and the four split plots (time slots) within each whole plot are assigned at random to the levels of the other factor (B).

If we look at the design in Table 19.1 and ignore factor B, we see that the levels of A are assigned according to a randomized complete block design, where the s subjects play the role of blocks, the 2 whole plots per block play the role of experimental units, and the 2 levels of A are assigned at random to the 2 whole plots within each block. Assuming no block $\times A$ interaction, the difference in the two levels of A could be analyzed like any randomized complete block design, using the whole-plot totals as the observations.

If we now look at the levels of B, they have also been assigned according to a randomized complete block design, but this time, the whole plots play the role of the blocks, and the four split plots nested within each whole plot are assigned to the four levels of B.

The analysis of the split-plot design is divided into two parts, reflecting this nested blocking system, each part with its own error. Analysis of the main effect of A involves comparisons of responses from split plots in different whole plots, whereas analysis of the main effect of B and the AB interaction involve comparisons of responses from split plots within the same whole plots.

In general, split plots within a whole plot will be more similar than split plots in different whole plots. Consequently, within-whole-plot comparisons will generally be more precise than between-whole-plot comparisons. So, in the rotating drum experiment, the main effect of B and the AB interaction will very likely be more precisely estimated than the main effect of A.

Ignoring the effects of the treatment factors for the moment, the response could be modeled as

$$Y_{hpq} = \mu + \theta_h + \epsilon_{p(h)}^{W} + \epsilon_{q(hp)}^{S},$$

where θ_h is the effect of the hth block, $\epsilon_{p(h)}^{W}$ is the effect of the pth whole plot nested within the hth block, and $\epsilon_{q(hp)}^{S}$ is the effect of the qth split plot nested within the pth whole plot in the hth block. We model the whole-plot and split-plot effects, and possibly the block effects, as random effects that are independent and normally distributed with mean 0 and variances σ_W^2, σ_S^2, and σ_θ^2, respectively.

Now, suppose that the levels of factor A are assigned to the whole plots, and in the hth block, the pth whole plot receives the uth assignment of the ith level of A ($i = 1, \ldots, a$; $u = 1, \ldots, \ell$). Also,

suppose that the levels of factor B are assigned to the split plots, and in the (hp)th whole plot, the qth split plot receives the tth assignment of the jth level of B ($j = 1, \ldots, b$; $t = 1, \ldots, m$). Then the model includes the effects of A, B and AB, and p is replaced by iu and q is replaced by jt, as follows:

$$Y_{hiujt} = \mu + \theta_h + \alpha_i + \epsilon^W_{iu(h)}$$
$$+ \beta_j + (\alpha\beta)_{ij} + \epsilon^S_{jt(hiu)}, \tag{19.2.1}$$
$$\epsilon^W_{iu(h)} \sim N(0, \sigma^2_W), \quad \epsilon^S_{jt(hiu)} \sim N(0, \sigma^2_S),$$
$$\epsilon^W_{iu(h)}\text{'s and } \epsilon^S_{jt(hiu)}\text{'s are all mutually independent,}$$
$$h = 1, \ldots, s; \quad i = 1, \ldots, a; \quad u = 1, \ldots, \ell; \quad j = 1, \ldots, b; \quad t = 1, \ldots, m,$$

where θ_h is the effect of the hth block, α_i is the effect of the ith level of factor A measured on the whole plots, the random variables $\epsilon^W_{iu(h)}$ represent the random effects of the whole plots, β_j is the effect of the jth level of B measured on the split plots, $(\alpha\beta)_{ij}$ is the interaction effect of A at level i and B at level j, and the random variables $\epsilon^S_{jt(hiu)}$ represent the random effects of the split plots.

The model (19.2.1) has been written on two lines to emphasize the two different parts of the design. In the design of Table 19.1, each level of A appears exactly once per block and each level of B appears exactly once per whole plot, so we may drop the subscripts u and t, and the model for this design becomes

$$Y_{hij} = \mu + \theta_h + \alpha_i + \epsilon^W_{i(h)}$$
$$+ \beta_j + (\alpha\beta)_{ij} + \epsilon^S_{j(hi)}, \tag{19.2.2}$$
$$\epsilon^W_{i(h)} \sim N(0, \sigma^2_W), \quad \epsilon^S_{j(hi)} \sim N(0, \sigma^2_S),$$
$$\epsilon^W_{i(h)}\text{'s and } \epsilon^S_{j(hi)}\text{'s are all mutually independent,}$$
$$h = 1, \ldots, s; \quad i = 1, \ldots, a; \quad j = 1, \ldots, b.$$

In some experiments the whole plot is the largest unit, which is equivalent to there being only one block ($s = 1$). In this case, model (19.2.1) becomes simpler, since the block effect θ_h and all subscripts h are omitted:

$$Y_{iujt} = \mu + \alpha_i + \epsilon^W_{iu}$$
$$+ \beta_j + (\alpha\beta)_{ij} + \epsilon^S_{jt(iu)}, \tag{19.2.3}$$
$$\epsilon^W_{iu} \sim N(0, \sigma^2_W), \quad \epsilon^S_{jt(iu)} \sim N(0, \sigma^2_S),$$
$$\epsilon^W_{iu}\text{'s and } \epsilon^S_{jt(iu)}\text{'s are all mutually independent,}$$
$$i = 1, \ldots, a; \quad u = 1, \ldots, \ell; \quad j = 1, \ldots, b; \quad t = 1, \ldots, m.$$

For unequal sample sizes, the ranges of the subscripts would be modified in models (19.2.1) and (19.2.3).

19.3   Analysis of a Split-Plot Design with Complete Blocks

In this section we consider only the case of equal sample sizes and randomized complete block designs for each of the treatment factors. There are then s blocks, each of which is divided into a whole plots, and each of these is subdivided into b split plots, giving a total of sab observations. Model (19.2.2) is used, and the degrees of freedom and sums of squares are calculated according to the rules in Chap. 18,

Table 19.2 Outline analysis of variance table for the rotating drum split-plot design

Source of variation	Degrees of freedom	Sum of squares	Mean square	Ratio
Block (Subjects)	$s - 1$	$ss\theta$	–	–
A (Lighting)	$a - 1$	ssA	msA	msA/msE_W
Whole-plot error	$(s-1)(a-1)$	ssE_W	msE_W	
Whole-plot total	$sa - 1$	ssW	–	–
B (Speed)	$b - 1$	ssB	msB	msB/msE_S
AB	$(a-1)(b-1)$	$ss(AB)$	$ms(AB)$	$ms(AB)/msE_S$
Split-plot error	$a(b-1)(s-1)$	ssE_S	msE_S	
Total	$abs - 1$	$sstot$		

Computational formulae

$$ss\theta = ab\Sigma_h \bar{y}_{h..}^2 - sab\bar{y}_{...}^2 \qquad ssW = b\Sigma_h \Sigma_i \bar{y}_{hi.}^2 - sab\bar{y}_{...}^2$$

$$ssA = sb\Sigma_i \bar{y}_{.i.}^2 - sab\bar{y}_{...}^2 \qquad ssB = sa\Sigma_j \bar{y}_{..j}^2 - sab\bar{y}_{...}^2$$

$$ssE_W = ssW - ss\theta - ssA \qquad ss(AB) = s\Sigma_i \Sigma_j \bar{y}_{.ij}^2 - sb\Sigma_i \bar{y}_{.i.}^2$$

$$sstot = \Sigma_h \Sigma_i \Sigma_j y_{hij}^2 - sab\bar{y}_{...}^2 \qquad\qquad - sa\Sigma_j \bar{y}_{..j}^2 + sab\bar{y}_{...}^2$$

$$ssE_S = sstot - ssW - ssB - ss(AB)$$

as shown in the following two subsections. The analysis of variance is outlined in Table 19.2 and, for this setting, is appropriate whether the blocks are fixed or random. The analysis of a split plot design with incomplete blocks is illustrated in Sects. 19.8.4 and 19.9.4, using the SAS and R software, respectively.

19.3.1　Split-Plot Analysis

Consider first the *split-plot analysis*, which is that part of the analysis (shown in the bottom half of the analysis of variance Table 19.2) that is based on the observations arising from the split plots within whole plots. There are $sab - 1$ total degrees of freedom, and the total sum of squares is

$$sstot = \sum_h \sum_i \sum_j y_{hij}^2 - sab\bar{y}_{...}^2 . \tag{19.3.4}$$

The b levels of factor B are assigned to the split plots within each whole plot according to a randomized complete block design. The sa whole plots are playing the role of sa blocks, so there are $sa - 1$ whole-plot degrees of freedom, and the whole-plot-total sum of squares is

$$ssW = b \sum_h \sum_i \bar{y}_{hi.}^2 - sab\bar{y}_{...}^2 . \tag{19.3.5}$$

Due to the fact that all levels of B are observed in every whole plot as in a randomized complete block design, the sum of squares for B needs no adjustment for whole plots, and is given by

$$ssB = sa \sum_j \bar{y}_{..j}^2 - sab\bar{y}_{...}^2 \tag{19.3.6}$$

corresponding to $b - 1$ degrees of freedom. The interaction between the factors A and B is also calculated as part of the split-plot analysis. Again, due to the complete block structure of both the

whole-plot design and the split-plot design, the interaction sum of squares needs no adjustment for blocks. The number of interaction degrees of freedom is $(a-1)(b-1) = ab - a - b + 1$, and the sum of squares is

$$ss(AB) = s \sum_i \sum_j \bar{y}_{.ij}^2 - sb \sum_i \bar{y}_{.i.}^2 - sa \sum_j \bar{y}_{..j}^2 + sab\bar{y}_{...}^2 . \qquad (19.3.7)$$

Since there are b split plots nested within the sa whole plots, there are, in total, $sa(b-1)$ split-plot degrees of freedom. Of these, $b-1$ are used to measure the main effect of B, and $(a-1)(b-1)$ are used to measure the AB interaction, leaving

$$sa(b-1) - (b-1) - (a-1)(b-1) = a(s-1)(b-1)$$

degrees of freedom for error. Equivalently, this can be obtained by subtraction of the whole-plot, B, and AB degrees of freedom from the total

$$(sab-1) - (sa-1) - (b-1) - (a-1)(b-1) = a(s-1)(b-1) .$$

The split-plot error sum of squares can also be calculated by subtraction:

$$ssE_S = sstot - ssW - ssB - ss(AB). \qquad (19.3.8)$$

The split-plot error mean square $msE_S = ssE_S/[a(s-1)(b-1)]$ is used as the error estimate in testing hypotheses and calculating confidence intervals for contrasts in B and AB. Notice that we cannot compare the levels of factor A on the split plots, since within each whole plot the level of A is held constant. The A contrasts are, in fact, confounded with whole plots.

The sums of squares (19.3.4)–(19.3.8) and their associated degrees of freedom are summarized in the bottom half of the analysis of variance table shown in Table 19.2.

19.3.2   Whole-Plot Analysis

We now move on to the *whole-plot analysis*, which is the part of the analysis based on comparisons of whole-plot totals. The levels of A are assigned to the whole plots within blocks according to a randomized complete block design, and so the sum of squares for A needs no block adjustment. There are $a-1$ degrees of freedom for A, so the sum of squares is given by

$$ssA = sb \sum_i \bar{y}_{.i.}^2 - sab\bar{y}_{...}^2 . \qquad (19.3.9)$$

There are $s-1$ degrees of freedom for blocks, giving a block sum of squares of

$$ss\theta = ab \sum_h \bar{y}_{h..}^2 - sab\bar{y}_{...}^2 . \qquad (19.3.10)$$

There are a whole plots nested within each of the s blocks, so there are, in total, $s(a-1)$ whole-plot degrees of freedom. Of these, $a-1$ are used to measure the effects of A leaving $(s-1)(a-1)$ degrees of freedom for whole-plot error. Equivalently, this can be obtained by the subtraction of the block and A degrees of freedom from the whole-plot total degrees of freedom

$$(sa - 1) - (s - 1) - (a - 1) = (s - 1)(a - 1).$$

Similarly, the whole-plot error sum of squares, which is used in testing hypotheses and calculating confidence intervals for contrasts in factor A, is obtained by subtraction:

$$ssE_W = ssW - ss\theta - ssA. \tag{19.3.11}$$

The sums of squares (19.3.9)–(19.3.11) and their corresponding degrees of freedom are summarized in the top half of Table 19.2 (p. 706).

   If the whole plot is the largest unit—namely, if there is no blocking of the whole plots—then the sum of squares for blocks in Table 19.2 effectively gets pooled into the whole-plot error sum of squares. Then the latter, still used in testing hypotheses and calculating confidence intervals for contrasts in factor A, is given by

$$ssE_W = ssW - ssA, \tag{19.3.12}$$

with $(sa - 1) - (a - 1) = a(s - 1)$ degrees of freedom.

19.3.3   Contrasts Within and Between Whole Plots

The formulae for the least squares estimates of the main effect and interaction treatment contrasts are similar to those given by rule 10 of Sect. 7.3 for fixed effects, since no block adjustments are needed. Thus

$$\sum_i c_i \hat{\alpha}_i^* = \sum_i c_i \overline{y}_{.i.}, \tag{19.3.13}$$

$$\sum_j d_j \hat{\beta}_j^* = \sum_j d_j \overline{y}_{..j},$$

$$\sum_i \sum_j k_{ij} \widehat{(\alpha\beta)}_{ij} = \sum_i \sum_j k_{ij} \overline{y}_{.ij},$$

where $\sum_i c_i = 0$, $\sum_j d_j = 0$, and $\sum_i k_{ij} = \sum_j k_{ij} = 0$.

   Consider the consequences of confounding factor A with whole plots. Using model (19.2.2), the corresponding main-effect-of-A contrast estimator can be expressed as

$$\sum_i c_i \hat{\alpha}_i^* = \sum_i c_i \overline{Y}_{.i.} = \sum_i c_i (\alpha_i^* + \overline{\epsilon}_{i(.)}^W + \overline{\epsilon}_{.(.i)}^S).$$

Consequently, this estimator has mean $\sum_i c_i \alpha_i^*$ and variance $\sum_i (c_i^2/sb)(b\sigma_W^2 + \sigma_S^2)$, where $(b\sigma_W^2 + \sigma_S^2)$ replaces σ^2 in rule 11 of Sect. 7.3 for fixed effects models. So, although main effects of A are confounded with whole plots, because the whole plot effects are random with mean zero, main effects of A are estimable but with larger variance reflecting whole plot variability in addition to split plot variability.

   For the estimates in Eq. (19.3.13), the corresponding estimated variances reflect whether the contrasts are measured in terms of whole-plot differences (as for contrasts in the levels of A) or split-plot (within whole-plot) differences (as for contrasts in B or AB). The former use the whole-plot error mean square, and the latter use the split-plot error mean square as follows.

$$\widehat{\text{Var}}\left(\sum_i c_i \hat{\alpha}_i^*\right) = \sum_i \frac{c_i^2}{sb}\, msE_W\,, \tag{19.3.14}$$

$$\widehat{\text{Var}}\left(\sum_j d_j \hat{\beta}_j^*\right) = \sum_j \frac{d_j^2}{sa}\, msE_S\,,$$

$$\widehat{\text{Var}}\left(\sum_i \sum_j k_{ij}\widehat{(\alpha\beta)}_{ij}\right) = \sum_i \sum_j \frac{k_{ij}^2}{s}\, msE_S\,.$$

For main effect and interaction contrasts, the methods of multiple comparison of Bonferroni, Scheffé, Tukey, and Dunnett can be used as usual (incorporating either the whole-plot or split-plot error mean square as above). Inferences for other contrasts in the treatment effects $\tau_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij}$, such as all pairwise comparisons, are more complicated and are discussed later in this chapter.

19.3.4   A Real Experiment—Oats Experiment

An experiment on the yield of three varieties of oats (factor A) and four different levels of manure (factor B) was described by F. Yates in his 1935 paper *Complex Experiments*. The experimental area was divided into $s = 6$ blocks. Each of these was then subdivided into $a = 3$ whole plots. The varieties of oat were sown on the whole plots according to a randomized complete block design (so that every variety appeared in every block exactly once). Each whole plot was then divided into $b = 4$ split plots, and the levels of manure were applied to the split plots according to a randomized complete block design (so that every level of B appeared in every whole plot exactly once). The design, after randomization, is shown in Table 19.3, together with the yields in quarter pounds. Model (19.2.2) was used.

Analysis of Variance—Oats Experiment

Using the formulae (19.3.4)–(19.3.11), we obtain the sums of squares shown in Table 19.4. To test, at significance level $\alpha = 0.01$, the hypothesis H_0^{AB} that the interaction between variety and manure level is negligible against the alternative hypothesis that the interaction is not negligible, we reject H_0^{AB} if

$$\frac{ms(AB)}{msE_S} = \frac{53.63}{177.08} = 0.30 > F_{6,45,0.01}\,.$$

Since $F_{6,45,0.01} \approx 3.2$, we do not reject H_0^{AB}, and we conclude that the interaction is negligible.

The hypothesis H_0^B of no difference in yield due to the different levels of manure (averaged over variety) is also tested using the split-plot error mean square as the denominator. We reject H_0^B in favor of the alternative hypothesis, that the manure levels do affect yield of oats, if

$$\frac{msB}{msE_S} = \frac{6673.50}{177.08} = 37.69 > F_{3,45,0.01}\,.$$

Since $F_{3,45,0.01} \approx 4.3$, we conclude that these four manure levels have different effects on the yield of the oat varieties tested.

Factor A is measured on the whole plots, so the whole-plot error is used as the denominator of the test statistic. We reject the hypothesis H_0^A of no difference in the average yields of the different varieties averaged over the manure levels if

Table 19.3 Split-plot design and yields (in quarter lb) for the oats experiment

Block	Level of A	Level of B (yield)		Block	Level of A	Level of B (yield)	
I	2	3 (156) 2 (118) 1 (140) 0 (105)		II	2	2 (109) 3 (99) 0 (63) 1 (70)	
	0	0 (111) 1 (130) 3 (174) 2 (157)			1	0 (80) 2 (94) 3 (126) 1 (82)	
	1	0 (117) 1 (114) 2 (161) 3 (141)			0	1 (90) 2 (100) 3 (116) 0 (62)	
III	2	2 (104) 0 (70) 1 (89) 3 (117)		IV	1	3 (96) 0 (60) 2 (89) 1 (102)	
	0	3 (122) 0 (74) 1 (89) 2 (81)			0	2 (112) 3 (86) 0 (68) 1 (64)	
	1	1 (103) 0 (64) 2 (132) 3 (133)			2	2 (132) 3 (124) 1 (129) 0 (89)	
V	1	1 (108) 2 (126) 3 (149) 0 (70)		VI	0	2 (118) 0 (53) 3 (113) 1 (74)	
	2	3 (144) 1 (124) 2 (121) 0 (96)			1	3 (104) 2 (86) 0 (89) 1 (82)	
	0	0 (61) 3 (100) 1 (91) 2 (97)			2	0 (97) 1 (99) 2 (119) 3 (121)	

Source Yates (1935). Copyright © 1935 Blackwell Publishers. Reprinted with permission

$$\frac{msA}{msE_W} = \frac{893.18}{601.33} = 1.49 > F_{2,10,0.01}\,.$$

Since $F_{2,10,0.01} = 7.56$, there is no evidence to conclude a difference in average yields of the three varieties of oats.

The same conclusions could be reached from the p-values in Table 19.4.

Multiple Comparisons—Oats Experiment

Suppose that level 0 of A was the currently used variety and that level 0 of B was the usual level of manure, and suppose that two-sided treatment-versus-control intervals had been required for both A and B at an overall level of 98% (that is, 99% for each set of Dunnett intervals).

Since both the split-plot and whole-plot designs are randomized complete block designs, the least squares estimates of the treatment contrasts are given by the formula in (19.3.13), so

$$\begin{array}{ll}
\hat{\alpha}_0^* - \hat{\alpha}_1^* = \overline{y}_{.0.} - \overline{y}_{.1.} = -6.875\,, & \hat{\beta}_0^* - \hat{\beta}_1^* = \overline{y}_{..0} - \overline{y}_{..1} = -19.500\,, \\
\hat{\alpha}_0^* - \hat{\alpha}_2^* = \overline{y}_{.0.} - \overline{y}_{.2.} = -12.167\,, & \hat{\beta}_0^* - \hat{\beta}_2^* = \overline{y}_{..0} - \overline{y}_{..2} = -34.833\,, \\
& \hat{\beta}_0^* - \hat{\beta}_3^* = \overline{y}_{..0} - \overline{y}_{..3} = -44.000\,,
\end{array}$$

Table 19.4 Analysis of variance for the oats split-plot experiment

Source of variation	Degrees of freedom	Sum of square	Mean square	Ratio	p-value
Blocks	5	15875.28	3175.06	–	
A (variety)	2	1786.36	893.18	1.49	0.2724
Whole-plot error	10	6013.31	601.33		
Whole-plot total	17	23674.94	1392.64		
B (manure)	3	20020.50	6673.50	37.69	0.0001
AB	6	321.75	53.63	0.30	0.9322
Split-plot error	45	7968.75	177.08		
Total	71	51985.94			

where $\overline{y}_{.i.}$ is an average over the $b = 4$ split plots within the $s = 6$ whole plots (one per block) on which level i of A is measured. Similarly, $\overline{y}_{..j}$ is an average over the $sa = 18$ split plots (one per whole plot) on which level j of B is measured. The confidence intervals are obtained from (19.3.13) and (19.3.14) as follows:

$$\alpha_0 - \alpha_1 \in \left((\overline{y}_{.0.} - \overline{y}_{.1.}) \pm t^{(0.5)}_{2,10,0.01} \sqrt{\frac{2}{24} msE_W} \right)$$

$$= \left(-6.875 \pm 3.53 \sqrt{\frac{2}{24}(601.33)} \right)$$

$$= (-6.875 \pm 24.99) = (-31.87, 18.12),$$

$$\alpha_0 - \alpha_2 = (-37.16, 18.12),$$

$$\beta_0 - \beta_1 \in \left((\overline{y}_{..0} - \overline{y}_{..1}) \pm t^{(0.5)}_{3,45,0.01} \sqrt{\frac{2}{18} msE_S} \right)$$

$$= \left(-19.5 \pm 3.09 \sqrt{\frac{2}{18}(177.03)} \right)$$

$$= (-19.5 \pm 13.70) = (-33.20, -5.79),$$

$$\beta_0 - \beta_2 \in (-48.54, -21.13),$$

$$\beta_0 - \beta_3 \in (-57.70, -30.30).$$

It is clear that the treatment-versus-control comparisons for the factor B manure levels are made more precisely ($msd = 13.70$) than those for the factor A oat varieties ($msd = 24.99$). This is primarily due to the much smaller error variance estimate, $msE_S < msE_W$, which reflects the fact that split plots within a whole plot are generally more similar than split plots in different whole plots. There are also more degrees of freedom associated with the split-plot error than with the whole-plot error, which also helps to reduce the minimum significant difference. Comparisons for factor B are more precise, despite the fact that the means $\overline{y}_{.i.}$ for factor A involve more observations.

19.4 Split-Split-Plot Designs

In the split-plot designs illustrated in Sect. 19.2, the factor A contrasts were confounded with the whole-plot contrasts, so that the main effect of A was assessed against the whole-plot variability, while the main effect of B and the AB interaction were assessed against the split-plot variability. It is possible

Table 19.5 Part of a split-split-plot design for the rotating drum experiment

| Block (Subject) | Whole-plot (Session) | Level of A (Light) | Split-plot 1 First half session | | Split-plot 2 Second half session | |
			Level of C (Direction)	Levels of B (Speed)	Level of C (Direction)	Levels of B (Speed)
I	1	0	1	0 3 1 2	0	2 1 3 0
	2	1	0	1 0 2 3	1	3 2 0 1
II	1	1	1	1 0 2 3	0	0 3 1 2
	2	0	0	2 1 3 0	1	3 2 0 1
:	:	:	:	: : : :	:	: : : :

Table 19.6 Analysis of variance for a split-split-plot design

Source of variation	Degrees of freedom	Mean square	Ratio
Blocks (subjects)	$s - 1$	$ms\theta$	
A (lighting)	$a - 1$	msA	msA/msE_W
Whole-plot error	$(s - 1)(a - 1)$	msE_W	
Whole-plot total	$sa - 1$	msW	
C (direction)	$c - 1$	msC	msC/msE_S
AC	$(a - 1)(c - 1)$	$ms(AC)$	$ms(AC)/msE_S$
Split-plot error	$a(s - 1)(c - 1)$	msE_S	
Split-plot total	$sac - 1$	msE_S	
B	$b - 1$	msB	msB/msE_{SS}
AB	$(a - 1)(b - 1)$	$ms(AB)$	$ms(AB)/msE_{SS}$
CB	$(c - 1)(b - 1)$	$ms(CB)$	$ms(CB)/msE_{SS}$
ACB	$(a - 1)(c - 1)(b - 1)$	$ms(ACB)$	$ms(ACB)/msE_{SS}$
Split-split-plot error	$ac(s - 1)(b - 1)$	msE_{SS}	
Total	$sacb - 1$	$mstot$	

to extend this idea, and to divide the split plots into split split plots on which are assigned the levels of a third factor.

For example, in the drum rotation experiment described in Sect. 19.2, the experimenter used a third factor C, the direction of rotation of the drum. A possible design for the experiment would be to ask each subject to be present at two sessions (whole plots) with a different lighting condition (A) at each session. In the first half of a session (split plot), set the direction of rotation (C), and run through each speed (B) in a random order (split split plots), changing the direction of rotation in the second half of the session. The design would then appear as in Table 19.5.

The model and analysis of variance table would have three parts, one for the whole plots nested within blocks together with the factor A effect, one for the split plots nested within whole plots together with the factor C effect and the AC interaction, and one for the split split plots nested within split plots together with the factor B effect and the other interactions, as shown in model (19.4.15):

$$\begin{aligned}
Y_{hijk} = \mu + \theta_h + \alpha_i + \epsilon_{i(h)}^W & \quad (19.4.15) \\
+ \gamma_j + (\alpha\gamma)_{ij} + \epsilon_{j(hi)}^S & \\
+ \beta_k + (\alpha\beta)_{ik} + (\gamma\beta)_{jk} + (\alpha\gamma\beta)_{ijk} + \epsilon_{k(hij)}^{SS} \, . &
\end{aligned}$$

The analysis of variance table, shown in Table 19.6, has three sections, reflecting the three parts of the model, and is illustrated through a real experiment in Sect. 19.7.3.

19.5   Split-Plot Confounding

If there are a number of different factors involved in a split-plot design, the size of the whole plots may not be large enough to allow a randomized complete block design to be used for the split-plot factors. We can obtain smaller blocks by confounding one or more interaction contrasts as we did for the single-replicate designs in Chap. 13. For example, suppose a two-replicate 2^4 experiment is to be conducted for treatment factors A, B, C, and D, and for practical reasons, the observations are to be divided into eight whole plots of size four. Suppose that the levels of A are sufficiently difficult to change that it is decided to change the level only after each whole plot of four observations is taken. The A contrasts are confounded with the whole plots, since each whole plot is assigned only one level of A.

Now, only four of the eight combinations of factors B, C, and D can be taken in any one whole plot. Thus, ignoring factor A, a design with $b = 8$ whole plots of size $k = 4$ and $v = 8$ treatment labels is required. An incomplete block design, such as a cyclic design, would be a possible choice. However, since a split-plot design is a complex design to analyze, it is better to select a repeated single-replicate design, so that we know exactly what is confounded with the whole plots. If we choose to confound BCD with the whole plots, as well as A, then $ABCD$ will also be confounded. The single-replicate design that confounds A, BCD, and $ABCD$ is obtained from the equations

$$
\begin{aligned}
a_1 &= 0 \text{ or } 1 \bmod 2\,, \\
a_2 + a_3 + a_4 &= 0 \text{ or } 1 \bmod 2\,.
\end{aligned}
$$

If we repeat this single-replicate design twice, we obtain the split-plot plan shown in Table 19.7. Before this plan can be used, the eight whole plots would need to be randomly ordered, and the four split plots within each whole plot would need to be randomly ordered. An outline analysis of variance table is shown in Table 19.8.

19.6   A Real Experiment—UAV Experiment

Sriram Mahadevan (2009) conducted three experiments at Wright State University to evaluate the performance of a semi-automated computer display system designed to support a human operator's ability to monitor and control the complex dynamic operation of multiple unmanned aerial vehicles (UAVs) when the UAVs are involved in multiple combat-related tasks. One of his experiments was conducted to examine the effects of different visual tools on the task performance efficiency and situation awareness of a single operator at a computer display to monitor and control dual-task scenarios involving UAVs.

The experiment involved 16 subjects (factor W) and $a = 2$ cue conditions (factor A), with eight of the 16 participants utilizing a baseline user interface with basic visual tools ($A = 1$), and with

Table 19.7   A split-plot confounded 2^4 experiment in 8 whole plots of size 4

Whole plot	A	Levels of B, C, D on the split plots				Whole plot	A	Levels of B, C, D on the split plots			
I	0	000	011	101	110	II	1	001	010	100	111
III	0	001	010	100	111	IV	1	000	011	101	110
V	0	000	011	101	110	VI	1	001	010	100	111
VII	0	001	010	100	111	VIII	1	000	011	101	110

Table 19.8 Outline analysis of variance table for a split-plot confounded 2^4 experiment in 8 whole plots of size 4

Source of variation	Degrees of freedom	Mean square	Ratio
A	1	msA	msA/msE_W
BCD	1	$ms(BCD)$	$ms(BCD)/msE_W$
$ABCD$	1	$ms(ABCD)$	$ms(ABCD)/msE_W$
Whole-plot error	4	msE_W	
Whole-plot total	7	msW	
B	1	msB	msB/msE_S
C	1	msC	msC/msE_S
D	1	msD	msD/msE_S
BC	1	$ms(BC)$	$ms(BC)/msE_S$
BD	1	$ms(BD)$	$ms(BD)/msE_S$
CD	1	$ms(CD)$	$ms(CD)/msE_S$
AB	1	$ms(AB)$	$ms(AB)/msE_S$
AC	1	$ms(AC)$	$ms(AC)/msE_S$
AD	1	$ms(AD)$	$ms(AD)/msE_S$
ABC	1	$ms(ABC)$	$msABC/msE_S$
ABD	1	$ms(ABD)$	$ms(ABD)/msE_S$
ACD	1	$ms(ACD)$	$ms(ACD)/msE_S$
Split-plot error	12	msE_S	
Total	31		

the other eight participants utilizing an advanced user interface involving more advanced visual tools ($A = 2$). Each subject ran through eight trials—one at each combination of task similarity (factor B, with $b = 2$ levels) and task complexity (factor C, with $c = 4$ levels). Task similarity depended on whether the primary and secondary tasks (which were done concurrently) were similar (coded 1), each task being a suppression-of-enemy-air-defenses (SEAD) mission, or dissimilar (coded 2), the primary and secondary tasks being SEAD and reconnaissance missions, respectively. Task complexity corresponds to the number of UAVs in the scenario, the levels being: simple-simple if both the primary and secondary tasks each involves two UAVs; simple-complex if the primary and secondary tasks involve two and four UAVs, respectively; complex-simple if the primary and secondary tasks involve four and two UAVs, respectively; and complex-complex if both the primary and secondary tasks each involves four UAVs. These levels were coded 1, 2, 3, 4, respectively.

This may be viewed as a split-plot design, with subjects serving as the whole plots, and with the eight trials per subject serving as the split plots. Main effects of cue are comparisons between subjects (i.e. between whole plots), since subjects are nested within cue levels, whereas main effects of B and C and all interactions are comparisons within whole plots.

One of the response variables measured was the time taken to perform situation awareness perception tasks in the primary task, yielding the data shown in Table 19.9. The model, including all treatment effects, is as follows.

Table 19.9 Split-plot design and times (in seconds) for the UAV experiment: situation awareness perception in the primary task

B (Similarity)		1				2			
C (Complexity)		1	2	3	4	1	2	3	4
A (Cue)	W (Subject)								
1	1	29	28	49	46	36	35	42	48
	2	26	26	53	42	35	40	46	44
	3	33	25	45	56	34	36	40	40
	4	27	28	44	47	36	32	50	54
	5	31	26	47	48	33	36	44	60
	6	28	27	51	41	31	34	38	43
	7	34	28	57	44	34	32	49	35
	8	25	35	50	54	35	50	45	48
2	1	15	13	16	18	21	18	20	21
	2	12	11	19	22	18	19	25	22
	3	14	13	21	19	25	23	21	23
	4	18	16	17	20	24	27	25	24
	5	15	15	18	21	19	20	21	21
	6	13	16	19	17	21	22	23	25
	7	17	18	20	16	22	23	27	24
	8	14	14	16	19	20	21	22	22

Source Mahadevan (2009). Copyright © 2009 Sriram Mahadevan. Reprinted with permission

$$
\begin{aligned}
Y_{iujk} = \mu + \alpha_i + \epsilon_{iu}^{W} & \\
+ \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{jk(iu)}^{S} \, , & \quad (19.6.16)\\
\epsilon_{iu}^{W} \sim N(0, \sigma_W^2) \, , \quad \epsilon_{jk(iu)}^{S} \sim N(0, \sigma_S^2) \, , & \\
\epsilon_{iu}^{W}\text{'s and } \epsilon_{jk(iu)}^{S}\text{'s are all mutually independent,} & \\
i = 1, 2; \ \ u = 1, \ldots, 8; \ \ j = 1, 2; \ \ k = 1, 2, 3, 4 \, , &
\end{aligned}
$$

This is analogous to model (19.2.1), except: (i) this model excludes block effects ($s = 1$); (ii) it includes the factorial effects γ_k, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$, and $(\alpha\beta\gamma)_{ijk}$ associated with factor C, all estimated as split-plot differences; and (iii) there is no replication of treatments within whole plots ($m = 1$).

19.6.1   Analysis of Variance

The analysis of variance table for the UAV experiment, given in Table 19.10, is similar to the one in Table 19.4 for the oats experiment, except for the following. There are no block effects, so the whole-plot error is obtained by subtraction as

$$
ssE_W = ssW - ssA \, ,
$$

with the corresponding degrees of freedom adjusted accordingly. Also, more treatment effects are involved in the split-plot analysis so the error sum of squares, obtained by subtraction of other effect sums of squares from the total sum of squares, is

Table 19.10   Analysis of variance for UAV experiment

Source of variation	Degrees of freedom	Sum of square	Mean square	Ratio	p-value
A (cue)	1	12880.13	12880.13	693.30	0.0001
Whole-plot error	14	260.09	18.58		
Whole-plot total	15	13140.22	876.01		
B (similarity)	1	457.53	457.53	34.57	0.0001
C (complexity)	3	2470.03	823.34	62.22	0.0001
AB	1	98.00	98.00	7.41	0.0077
AC	3	1177.94	392.65	29.67	0.0001
BC	3	351.28	117.09	8.85	0.0001
ABC	3	153.31	51.10	3.86	0.0117
Split-plot error	98	1296.91	13.23		
Total	127	19145.22			

$$ssE_S = sstot - ssW - ssB - ssC - ss(AB) - ss(AC) - ss(BC) - ss(ABC).$$

Factor A is tested relative to whole-plot error, while the other treatment effects are tested relative to the split-plot error. As one can see from the p-values in Table 19.10, all treatment main and interaction effects are significant at the individual $\alpha = .01$ levels, except for the ABC interaction which has a p-value slightly larger than .01.

19.6.2   Multiple Comparisons

The experimenter was interested in the main effect of cue, averaged over the other factors; that is,

$$\alpha_1^* - \alpha_2^* = [\alpha_1 + \overline{(\alpha\beta)}_{1.} + \overline{(\alpha\gamma)}_{1.} + \overline{(\alpha\beta\gamma)}_{1..}] - [\alpha_2 + \overline{(\alpha\beta)}_{2.} + \overline{(\alpha\gamma)}_{2.} + \overline{(\alpha\beta\gamma)}_{2..}],$$

comparing the effects of the two cue conditions, averaging over combinations of levels of similarity and complexity. Paralleling the discussion in Sect. 19.3: the least squares estimate of the main effect of cue is

$$\hat{\alpha}_1^* - \hat{\alpha}_2^* = \overline{y}_{1...} - \overline{y}_{2...} = 39.4531 - 19.3906 = 20.0625 \text{ seconds};$$

also, $\mathrm{Var}(\overline{Y}_{1...} - \overline{Y}_{2...}) = \mathrm{Var}[(\overline{\epsilon}_{1.}^W + \overline{\epsilon}_{..(1.)}^S) - (\overline{\epsilon}_{2.}^W + \overline{\epsilon}_{..(2.)}^S)] = 2(\sigma_W^2/8 + \sigma_S^2/64) = \sigma_W^2/4 + \sigma_S^2/32$, and this is estimated by $msE_W/32$. So, with 99% confidence,

$$\alpha_1^* - \alpha_2^* \in \left((\overline{y}_{.1..} - \overline{y}_{.2..}) \pm t_{14, 0.005} \sqrt{msE_W/32} \right)$$
$$= \left(20.0625 \pm 2.977 \sqrt{18.58/32} \right)$$
$$= (20.0625 \pm 2.2684) = (17.7941, \ 22.3309).$$

Hence, with 99% confidence, the main effect of cue is between 17.7941 and 22.3309 seconds so, on average, operators using the enhanced user interface spend between 17.7941 and 22.3309 seconds less time to perform the requested situation awareness perception tasks in the primary task than do operators using the baseline user interface.

Given the significant main effect of cue, coupled with the p-value of 0.0117 for the ABC interaction, the experimenter decided to also investigate the simple pairwise differences comparing the effect of cue at each of the eight combinations of the other two factors. For each such combination jk of BC, the estimator of the simple pairwise difference

$$[\alpha_1 + (\alpha\beta)_{1j} + (\alpha\gamma)_{1k} + (\alpha\beta\gamma)_{2jk}] - [\alpha_2 + (\alpha\beta)_{2j} + (\alpha\gamma)_{2k} + (\alpha\beta\gamma)_{2jk}] \qquad (19.6.17)$$

is $\overline{Y}_{1.jk} - \overline{Y}_{2.jk}$, with variance

$$\text{Var}[(\overline{\epsilon}_{1.}^W + \overline{\epsilon}_{jk(1.)}^S) - (\overline{\epsilon}_{2.}^W + \overline{\epsilon}_{jk(2.)}^S)] = 2(\sigma_W^2/8 + \sigma_S^2/8) = (\sigma_W^2 + \sigma_S^2)/4.$$

These simple pairwise differences are neither within- nor between-whole-plot comparisons, so a composite variance estimator is needed. Now, $E[MSE_S] = \sigma_S^2$. Also, one can show that $E[MSE_W] = 8\sigma_W^2 + \sigma_S^2$, (by rule 17 of Sect. 17.8.1, there being 8 observations on each whole plot). So, the variance estimate is

$$(msE_W + 7msE_S)/32 = (18.578 + 7(13.234))/32 \approx 3.4755,$$

with estimated standard error 1.8643 obtained as the square root. The degrees of freedom associated with this variance estimate is computed using Satterthwaite's approximation as follows:

$$df = \frac{(msE_W + 7msE_S)^2}{(msE_W)^2/14 + (7msE_S)^2/98} = \frac{12,369}{24.6533 + 87.5659} \approx 110 \,.$$

So for example, an individual 99% confidence interval for (19.6.17) has minimum significant difference $msd = (t_{110,0.005})(1.8643) = (2.621)(1.8643) \approx 4.8863$. The corresponding simple pairwise difference estimates are as follows,

jk	11	12	13	14	21	22	23	24
$\overline{y}_{.1jk} - \overline{y}_{.2jk}$	14.375	13.375	31.250	28.250	13.000	15.250	21.250	23.750

where levels 1 and 2 of B are similar and dissimilar, respectively, and levels 1–4 of C are simple-simple, simple-complex, complex-simple, and complex-complex, respectively. Since each of these simple pairwise difference estimates exceeds the minimum significant difference 4.8863 and is positive, the enhanced user interface provides significant mean time reductions in performing the requested situation awareness perception tasks in the primary task for each combination of task similarity and task complexity. The overall significance level for testing these eight differences would have been at most 0.08, had these comparisons been preplanned. While they perhaps were not, one can also show that each of these comparisons is significantly nonzero with a p-value of less than 0.0001 when tested individually, and so one can feel rather confident that the effects are real.

19.7   A Real Experiment—Mobile Computing Field Study

Mary Mc. Wesler (2001) conducted a field experiment to study the effective use of mobile computing devices for real-time navigation and situation awareness, with applications in the military domain. One goal of the experiment was to study the effects of display type and visual presentation format on navigational performance, taking information garnered from laboratory studies and putting it to the test in more realistic field studies. Two display types (factor A) were studied: a handheld, headdown

display; and a helmet mounted display, facilitating headup usage. Three visual presentation formats (factor B) were studied: second person egocentric; birds' eye view egocentric; and birds' eye view exocentric. Each experimental run consisted of a subject attempting to navigate, quickly and accurately, a prescribed route or path through unfamiliar, moderate, lightly wooded terrain in the dark of night. Response variables of interest included: elapsed time, distance traveled, and root mean square error of distance off of the prescribed path. The latter two response variables were based on GPS data.

The experiment involved 12 subjects and 24 days, with each subject participating for two days. There were various constraints on the experiment. For example, on any given day it was feasible to make three runs with one subject, with subjects and days both considered possible sources of variation. Involving each subject for two days provided a single replicate of the 2×3 combinations of A and B with each subject, allowing all treatment effects to be within-subject comparisons. However, with only three runs per day, it was decided to fix display type but vary visual presentation format each day, so visual presentation format comparisons are within days, but display type comparisons are between days. Correspondingly, each subject only needed to be trained on one display type each day. In short, the experimenter used a split-plot design—the 12 subjects constitute blocks, the 24 days constitute whole plots with two whole plots per block, and the three runs per day constitute split plots.

The experiment also involved paths and run order—additional nuisance sources of variation the experimenter chose to control in the experiment and adjust for in the analysis. Each subject navigated each of three paths frontward and backward for six paths in total, each path being roughly square in shape with north, south, east, and west legs.

Twelve subjects was enough to design the experiment so treatment effects were not confounded with subject, path, or run order, though display type was confounded with days (i.e. whole plots). The design construction is discussed in Sect. 19.7.4. The resulting root mean square error (*RMSE*) data are in Table 19.11. Prior to analysis of the data, it was discovered that one of the subjects traversed one of the paths in the reverse direction. This was deemed to adversely affect some response variables, making the original data inappropriate to use in the evaluation. As such, one of the 72 planned observations is listed as missing.

The model used by the experimenter for the data analysis is as follows.

$$
\begin{aligned}
Y_{dhijqr} = \mu &+ \theta_h + \alpha_i + (\phi_2)_{q_2} + (\rho_2)_{r_2} + \epsilon_{d(h)}^W \\
&+ \beta_j + (\alpha\beta)_{ij} + (\phi_3)_{q_3} + (\phi_2\phi_3)_{q_2q_3} \\
&+ (\rho_3)_{r_3} + (\rho_2\rho_3)_{r_2r_3} + \epsilon_{ijq_3r_3(dhq_2r_2)}^S ,
\end{aligned}
\tag{19.7.18}
$$

$$
\theta_h \sim N(0, \sigma_\theta^2) , \quad \epsilon_{d(h)}^W \sim N(0, \sigma_W^2) , \quad \epsilon_{ijq_3r_3(dhq_2r_2)}^S \sim N(0, \sigma_S^2) ,
$$

$$
\theta_h\text{'s}, \ \epsilon_{d(h)}^W\text{'s and } \epsilon_{ijq_3r_3(dhq_2r_2)}^S\text{'s are all mutually independent,}
$$

$$
d = 1, 2; \quad h = 1, \ldots, 12; \quad i = 1, 2; \quad j = 1, 2, 3;
$$

$$
q = 3(q_3 - 1) + 2(q_2 - 1) + 1, \ \text{for } q_2 = 1, 2, \ \ q_3 = 1, 2, 3;
$$

$$
r = 3(r_3 - 1) + 2(r_2 - 1) + 1, \ \text{for } r_2 = 1, 2, \ \ r_3 = 1, 2, 3.
$$

In the model, θ_h denotes subject (i.e. block) effects, α_i and β_j denote the display type and visual presentation format effects, respectively, $\epsilon_{d(h)}^W$ denotes day (i.e. whole plot) effects, and $\epsilon_{ijq_3r_3(dhq_2r_2)}^S$ denotes the split-plot effects.

Table 19.11 Mobile computing field study: root mean square error (*RMSE*) data for each subject (Subj), day, run order, and path-treatment combination (*PAB*), with one observation missing

| Subj | Day 1 | | | | | | Day 2 | | | | | |
| | Run 1 | | Run 2 | | Run 3 | | Run 4 | | Run 5 | | Run 6 | |
	PAB	RMSE	PAB	RMSE	PAB	RMSE	PAB	RMSE	PAB	RMSE	PAB	RMSE
1	111	49.321	212	24.386	313	37.680	421	34.291	522	32.053	623	43.121
2	213	45.469	311	24.224	112	29.063	523	32.020	621	33.478	422	36.942
3	312	20.378	113	47.680	211	39.962	622	19.876	423	37.471	521	27.410
4	121	50.088	222	19.010	323	28.749	411	41.306	512	21.924	613	41.469
5	223	27.452	321	46.204	122	42.188	513	28.919	611	24.599	412	28.361
6	322	73.913	123	59.856	221	73.709	612	32.315	413	80.809	511	36.911
7	421	45.506	522	40.380	623	37.135	111	54.271	212	26.685	313	26.076
8	523	29.368	621	34.627	422	27.021	213	33.799	311	43.408	112	40.724
9	622	19.946	423	36.251	521	29.726	312	26.041	113	53.162	211	27.546
10	411	77.787	512	66.956	613	51.114	121	67.176	222	38.774	323	48.602
11	513	26.584	611	41.038	412	47.183	223	37.914	321	30.026	122	44.577
12	612	31.994	413	35.671	511	30.146	322	.	123	58.481	221	62.659

Source Wesler (2001). Copyright © 2001 Mary Mc. Wesler. Research was performed under U.S. Army Research Laboratory, Federated Laboratory Research Consortium (DAAL01-96-0003) directed by Mr. Bernie Corona. Reprinted with permission

Furthermore, construction of the design involved various pseudofactors. For example, the factor path (P) has been represented by pseudofactors: path direction P_2 at two levels, corresponding to using a path frontwards ($P_2 = 0$, for paths 1–3) or backwards ($P_2 = 1$, for paths 4–6); and P_3 at three levels, corresponding to the three paths used. The corresponding effects of P_2, P_3 and $P_2 P_3$ are modeled by the parameters $(\phi_2)_{q_2}$, $(\phi_3)_{q_3}$, and $(\phi_2 \phi_3)_{q_2 q_3}$, respectively. Likewise, the factor run order (O) has been represented by pseudofactors: O_2 at two levels, corresponding to runs on day 1 ($O_2 = 0$, for runs 1–3) or day 2 ($O_2 = 1$, for runs 4–6); and O_3 at three levels, corresponding to the three runs on a given day. The effects corresponding to O_2, O_3 and $O_2 O_3$ are modeled by the parameters $(\rho_2)_{r_2}$, $(\rho_3)_{r_3}$, and $(\rho_2 \rho_3)_{r_2 r_3}$, respectively.

Since each subject makes the first three runs on one day and the other three on a second day, one degree of freedom for run order corresponding to O_2 is confounded with days (i.e. whole plots). Also, since paths 1–3 are always used on one day for each subject and paths 4–6 on the other day, the one degree of freedom for path direction corresponding to P_2 is also confounded with days.

Consequently, the between-whole-plot comparisons consist of: main effects of display type A, subject main effects, the path direction effect P_2, and the run order effect O_2. Within-whole-plot comparisons include: main effects of display format B, AB interaction effects, the path effects P_3 and $P_2 P_3$, and the run order effects O_3 and $O_2 O_3$.

19.7.1   Analysis of Variance

Recall, one observation is missing since one of the subjects traversed one of the paths in the wrong direction from what was intended. Consequently, the resulting 71-run experiment deviates modestly from the balanced 72-run experiment that was planned, complicating the data analysis, since the standard formulas for a balanced design are no longer applicable. Direct analysis of the 71 observations collected is an option, and it will be illustrated using the SAS and R software packages in Sects. 19.8.4

Table 19.12 Analysis of variance for the mobile computing field study

Source of variation	Degrees of freedom	Sum of square	Mean square	Ratio	p-value
Subject	11	6230.92	566.45	2.92	0.0594
Order (O_2)	1	30.48	30.48	0.16	0.7012
Path (P_2)	1	379.34	379.34	1.95	0.1957
A (display)	1	47.95	47.95	0.25	0.6311
Whole-plot error (day)	9	1747.52	194.17	2.17	0.0483
Whole-plot total	23	8436.21			
Order (O_3, $O_2 O_3$)	4	56.56	14.14	0.16	0.9581
Path (P_3, $P_2 P_3$)	4	1941.84	485.46	5.42	0.0016
B (VPF)	2	806.44	403.22	4.51	0.0179
AB	2	166.48	83.24	0.93	0.4038
Split-plot error	36	3222.00	89.50		
Total	71	14629.54			

and 19.9.4, respectively. Another option, illustrated here, is the following traditional remedy involving the estimation of missing values.

Given the missing observation, a standard approach is to fit the intended model using the 71 observations collected, estimate the missing value, then analyze the data as a 72-run experiment including the estimated value as if not missing, except making the following accommodation. With one observation missing, there is one fewer degrees of freedom for the analysis. In comparing the results when using software to fit model (19.7.18) with the 71 and 72 observations, respectively, the analysis of variance for the 71-run fit (not shown here) yields one fewer degrees of freedom for split-plot error—namely, 35 rather than 36. So, the same effects are estimable in each case, but the lost observation results in one fewer split-plot error degrees of freedom. Consequently, in analyzing the 72-run data set including the estimated missing value, the number of split-plot error degrees of freedom should be reduced by one. That said, the reduction from 36 to 35 split-plot error degrees of freedom has negligible impact on the analysis, as corresponding critical values are nearly the same with this many error degrees of freedom, so such an adjustment is not always made in the analyses presented here.

Recall, the experimental plan was a split-plot design, with a single response on each independent variable for each of the six treatment combinations per subject. The subjects served as blocks in the design, days as whole plots, and the 72 runs as split plots. The analysis of variance is given in Table 19.12 and is analogous to the one in Table 19.4 (p. 710) for the oats split-plot experiment.

The experimenter planned to test each treatment effect using a 5% significance level. As one can see from the p-values in Table 19.12, the interaction between display type and visual presentation format is not significant ($p = 0.4038$) using $\alpha = 0.05$, nor is the main effect of display type ($p = 0.6311$), but the main effect of visual presentation format is significant ($p = 0.0179$) at the 5% level.

As one can also see from the p-values, there are significant effects of path (i.e. P_3 and $P_2 P_3$ collectively) but not of P_2, so the three paths used had significant differences either themselves or interacting with path direction, but the path direction was not significant. Also, the subject and day effects were modestly significant with $p = 0.0594$ and $p = 0.0483$, respectively, but order effects were not significant.

19.7.2 Multiple Comparisons

For each treatment effect found to be significant in the analysis of variance, the experimenter followed up with multiple comparisons using simultaneous 95% confidence intervals. Since only the main effect of visual presentation format was significant, Tukey's method was used to compare the three formats.

The estimated pairwise comparisons are:

$$\hat{\beta}_1^* - \hat{\beta}_2^* = \overline{y}_{...1..} - \overline{y}_{...2..} = 7.792\,,$$
$$\hat{\beta}_3^* - \hat{\beta}_2^* = \overline{y}_{...3..} - \overline{y}_{...2..} = 6.102\,,$$
$$\hat{\beta}_1^* - \hat{\beta}_3^* = \overline{y}_{...1..} - \overline{y}_{...3..} = 1.690\,.$$

Since the corresponding F-test uses the split-plot mean squared error as the denominator and the 72-run design is balanced, the estimated standard error of each pairwise comparison also utilizes the split-plot mean squared error. Also, since each treatment sample mean is the average of 24 observations, the variance estimate of each pairwise contrast is $2msE_S/24 = 89.5/12 \approx 7.4583$. So, the minimum significant difference for each comparison is $(q_{3,35,0.05}/\sqrt{2})\sqrt{7.4583} \approx 6.69$, for $q_{3,35,0.05} = 3.465$, with the split-plot error degrees of freedom having been reduced by one to 35 because of the missing value. Hence, visual presentation format 1 (second person egocentric) has a significantly larger mean *RMSE* from the intended path than does visual presentation format 2 (birds' eye view egocentric), indicating that format 2 is preferable. The other two pairwise comparisons are not significantly different, though format 3 also does substantially worse than format 2. The reader may verify for example that the minimum significant difference would only be 5.89 for simultaneous 90% confidence intervals, in which case visual presentation format 2 would be significantly better than both of the other formats.

19.7.3 Analysis as a Split-Split-Plot Design

The original experimental plan was a split-plot design, with a single response on each independent variable for each of the six treatment combinations per subject. The subjects served as blocks in the design, days as whole plots, and the 72 runs as split plots. The split-plot analysis was provided in Sects. 19.7.1 and 19.7.2.

It subsequently became of interest to study whether the effectiveness of the display types (factor A) or visual presentation formats (factor B) depended on the direction being navigated. One could consider utilizing the data for this purpose because, as noted previously, each path was roughly square in shape with north, south, east, and west legs, (coded 1–4, respectively). Furthermore, data was available on the independent variables for each of the four legs of each run. It was ultimately decided to analyze the data as a split-split-plot design, including leg direction as a factor, treating each leg of a run as a split-split-plot. In that regard, one should note that leg order may also have an effect and may be confounded with leg direction. Nonetheless, it was decided to include leg direction rather than leg order as a factor, since previous studies indicate that leg direction is a stronger independent variable, noting for example that participants in another study had difficulty cognitively transposing information from a fixed north-up display while traversing a south-bound leg of a test course. For this reason, visual display format by leg direction interaction effects were also included in the model.

Adding leg direction factor D with effects δ_t ($t = 1, \ldots, 4$), visual presentation format by leg direction interaction effects $(\beta\delta)_{jt}$, and split-split-plot errors $\epsilon_{t(dhijqr)}^{SS}$ to the split-plot design model (19.7.18), the split-split-plot design model is as follows.

$$
\begin{aligned}
Y_{dhijqrt} = {} & \mu + \theta_h + \alpha_i + (\phi_2)_{q_2} + (\rho_2)_{r_2} + \epsilon^W_{d(h)} \\
& + \beta_j + (\alpha\beta)_{ij} + (\phi_3)_{q_3} + (\phi_2\phi_3)_{q_2 q_3} \\
& + (\rho_3)_{r_3} + (\rho_2\rho_3)_{r_2 r_3} + \epsilon^S_{ijq_3 r_3(dhq_2 r_2)} \\
& + \delta_t + (\beta\delta)_{jt} + \epsilon^{SS}_{t(dhijqr)},
\end{aligned}
\tag{19.7.19}
$$

$$
\theta_h \sim N(0, \sigma^2_\theta), \quad \epsilon^W_{d(h)} \sim N(0, \sigma^2_W),
$$

$$
\epsilon^S_{ijq_3 r_3(dhq_2 r_2)} \sim N(0, \sigma^2_S), \quad \epsilon^{SS}_{t(dhijqr)} \sim N(0, \sigma^2_{SS})
$$

θ_h's, $\epsilon^W_{d(h)}$'s, $\epsilon^S_{ijq_3 r_3(dhq_2 r_2)}$'s, $\epsilon^{SS}_{t(dhijqr)}$'s are all mutually independent,

$$
d = 1, 2; \quad h = 1, \dots, 12; \quad i = 1, 2; \quad j = 1, 2, 3;
$$

$$
q = 3(q_3 - 1) + 2(q_2 - 1) + 1, \text{ for } q_2 = 1, 2, \quad q_3 = 1, 2, 3;
$$

$$
r = 3(r_3 - 1) + 2(r_2 - 1) + 1, \text{ for } r_2 = 1, 2, \quad r_3 = 1, 2, 3;
$$

$$
t = 1, 2, 3, 4.
$$

The data for the split-split-plot design consist of the *RMSE* values for each of the four legs of each run. The data, shown in Table 19.13, now include four missing observations, since the one missing run of the split-plot design includes four legs. Estimating these four missing values, the analysis of variance is provided in Table 19.14. As the reader may verify, if the model is fit without the estimated missing values, one loses one degree of freedom for split-plot error and three for split-split-plot error, but adjusting the corresponding degrees of freedom from 36 to 35 and from 207 to 204, respectively, would have minimal impact on any analysis. The first two sections of Table 19.14 duplicate information given in Table 19.12 for the split-plot analysis—the degrees of freedom, ratio, and p-value is the same for each source so conclusions are unchanged, though the sum of squares and mean squares are different since the data are observations on legs of runs, not runs. The visual presentation format by leg direction interaction is not significant, but leg direction is significant with $p = 0.0117 < 0.05$. Hence, the experimenter compared the leg direction effects using simultaneous 95% confidence intervals obtained via Tukey's method.

The estimated pairwise comparisons are:

$$
\begin{array}{ll}
\hat{\delta}^*_2 - \hat{\delta}^*_1 = \overline{y}_{\dots\dots 2} - \overline{y}_{\dots\dots 1} = 2.8645, & \hat{\delta}^*_4 - \hat{\delta}^*_1 = \overline{y}_{\dots\dots 4} - \overline{y}_{\dots\dots 1} = 1.5666, \\
\hat{\delta}^*_2 - \hat{\delta}^*_3 = \overline{y}_{\dots\dots 2} - \overline{y}_{\dots\dots 3} = 2.0217, & \hat{\delta}^*_4 - \hat{\delta}^*_3 = \overline{y}_{\dots\dots 4} - \overline{y}_{\dots\dots 3} = 0.7239, \\
\hat{\delta}^*_2 - \hat{\delta}^*_4 = \overline{y}_{\dots\dots 2} - \overline{y}_{\dots\dots 4} = 1.2979, & \hat{\delta}^*_3 - \hat{\delta}^*_1 = \overline{y}_{\dots\dots 3} - \overline{y}_{\dots\dots 1} = 0.8427.
\end{array}
$$

Since the corresponding F-test uses the (split-split-plot) mean square error as the denominator and the 288-leg design is balanced, the estimated standard error of each pairwise comparison also utilizes the mean squared error. Also, since each treatment sample mean is the average of 72 of the 288 observations, the variance estimate of each pairwise contrast is $2(msE)/72 = 2(28.19)/72 \approx 0.7831$. So, the minimum significant difference for each comparison is $(q_{4,204,0.05}/\sqrt{2})\sqrt{0.7831} \approx 2.27$, for $q_{3,204,0.05} \approx q_{3,\infty,0.05} = 3.63$, with the split-plot error degrees of freedom having been reduced by three to 204 because of the missing values. Hence, mean *RMSE* is significantly greater for leg direction 2 (south) than for leg direction 1 (north). This provides some credence to the observation in another study that participants had difficulty cognitively transposing information from a fixed north-up display while traversing a south-bound leg of a test course.

Table 19.13   Mobile computing field study split-split-plot design: root mean square error (*RMSE*) data for each subject (Subj), leg, day, run order, and path-treatment combination (*PAB*), with four observations missing

Subj	Leg	Day 1						Day 2					
		Run 1		Run 2		Run 3		Run 4		Run 5		Run 6	
		PAB	RMSE	PAB	RMSE	PAB	RMSE	PAB	RMSE	PAB	RMSE	PAB	RMSE
1	1	111	6.313	212	3.653	313	6.762	421	6.223	522	6.888	623	7.506
	2	111	25.032	212	4.259	313	4.269	421	5.621	522	9.165	623	12.035
	3	111	7.600	212	7.011	313	11.870	421	13.733	522	8.129	623	17.760
	4	111	10.376	212	9.464	313	14.780	421	8.713	522	7.871	623	5.820
2	1	213	8.574	311	6.332	112	5.313	523	5.832	621	3.601	422	5.813
	2	213	7.244	311	5.257	112	12.927	523	7.441	621	6.756	422	7.498
	3	213	19.585	311	10.759	112	3.666	523	8.659	621	4.837	422	10.979
	4	213	10.066	311	1.877	112	7.157	523	10.088	621	18.283	422	12.651
3	1	312	8.700	113	10.037	211	11.989	622	3.607	423	10.548	521	9.446
	2	312	5.436	113	15.766	211	6.458	622	5.892	423	7.323	521	4.318
	3	312	4.101	113	9.985	211	11.384	622	3.577	423	6.705	521	5.304
	4	312	2.142	113	11.892	211	10.130	622	6.800	423	12.895	521	8.343
4	1	121	3.700	222	4.378	323	4.348	411	18.534	512	3.393	613	4.472
	2	121	26.995	222	4.065	323	10.347	411	4.280	512	4.250	613	13.711
	3	121	6.061	222	5.764	323	2.645	411	9.041	512	7.794	613	9.561
	4	121	13.332	222	4.804	323	11.409	411	9.450	512	6.487	613	13.724
5	1	223	7.924	321	8.801	122	6.337	513	3.435	611	5.669	412	7.954
	2	223	2.563	321	16.128	122	21.415	513	4.446	611	8.044	412	6.689
	3	223	6.068	321	4.969	122	4.171	513	4.716	611	5.759	412	5.960
	4	223	10.897	321	16.305	122	10.265	513	16.321	611	5.127	412	7.758
6	1	322	22.194	123	11.316	221	11.540	612	3.791	413	15.781	511	5.236
	2	322	13.209	123	24.914	221	18.154	612	9.139	413	20.223	511	3.422
	3	322	13.197	123	9.051	221	34.058	612	7.538	413	28.214	511	4.137
	4	322	25.314	123	14.575	221	9.957	612	11.847	413	16.591	511	24.117
7	1	421	15.819	522	17.896	623	6.257	111	2.320	212	6.519	313	4.058
	2	421	6.846	522	7.257	623	8.781	111	26.503	212	6.613	313	5.936
	3	421	14.657	522	7.117	623	13.149	111	21.167	212	8.820	313	11.339
	4	421	8.184	522	8.110	623	8.948	111	4.280	212	4.733	313	4.743
8	1	523	11.039	621	3.718	422	7.356	213	3.837	311	5.856	112	6.410
	2	523	7.973	621	18.952	422	9.132	213	3.759	311	16.997	112	26.575
	3	523	6.242	621	5.573	422	4.332	213	9.298	311	8.671	112	2.342
	4	523	4.115	621	6.384	422	6.200	213	16.905	311	11.883	112	5.397
9	1	622	4.371	423	8.725	521	7.110	312	4.352	113	7.832	211	7.286
	2	622	4.846	423	11.287	521	6.264	312	6.641	113	31.707	211	6.662
	3	622	6.666	423	10.870	521	9.541	312	6.450	113	1.774	211	9.589
	4	622	4.063	423	5.369	521	6.810	312	8.597	113	11.848	211	4.008
10	1	411	22.287	512	22.921	613	15.016	121	26.405	222	7.143	323	10.467
	2	411	19.107	512	12.482	613	5.827	121	13.756	222	11.160	323	21.939
	3	411	15.018	512	15.430	613	18.707	121	11.025	222	12.443	323	5.404
	4	411	21.375	512	16.123	613	11.564	121	15.990	222	8.029	323	10.791
11	1	513	5.883	611	6.531	412	10.834	223	9.284	321	6.103	122	3.570
	2	513	7.288	611	15.533	412	8.584	223	10.828	321	12.479	122	24.502
	3	513	3.260	611	8.494	412	18.250	223	9.102	321	6.008	122	5.115
	4	513	10.154	611	10.479	412	9.515	223	8.701	321	5.436	122	11.389
12	1	612	5.826	413	12.041	511	8.747	322	.	123	9.794	221	16.560
	2	612	10.782	413	8.086	511	8.952	322	.	123	23.941	221	17.369
	3	612	7.168	413	5.918	511	1.080	322	.	123	20.566	221	11.995
	4	612	8.218	413	9.626	511	11.367	322	.	123	4.180	221	16.735

Source Wesler (2001). Copyright © 2001 Mary Mc. Wesler. Research was performed under U.S. Army Research Laboratory, Federated Laboratory Research Consortium (DAAL01-96-0003) directed by Mr. Bernie Corona. Reprinted with permission

Table 19.14 Analysis of variance for the mobile computing field study

Source of variation	Degrees offreedom	Sum ofsquare	Meansquare	Ratio	*p*-value
Subject	11	1557.78	141.62	2.92	0.0594
O2	1	7.62	7.62	0.16	0.7011
P2	1	94.84	94.84	1.95	0.1957
Display	1	11.99	11.99	0.25	0.6311
Whole-plot error (day)	9	436.86	48.54	2.17	0.0483
Whole-plot total	23	2109.09			
Order	4	14.15	3.54	0.16	0.9581
Path	4	485.43	121.36	5.42	0.0016
VPF	2	201.60	100.80	4.51	0.0179
VPF*Display	2	41.62	20.81	0.93	0.4038
Split-plot error	36	805.47	22.37	0.79	0.7934
Split-plot total	71	3657.36			
LegDir	3	317.98	105.99	3.76	0.0117
VPF*LegDir	6	34.85	5.81	0.21	0.9746
Split-split-plot error	207	5835.85	28.19		
Total	287	9846.03			

19.7.4   Design Construction

In this section we discuss construction of the split-plot design used in the mobile computing field study. The design used (Table 19.11, p. 719) is a fractional factorial design with some day effects confounded with blocks (subjects). The design construction illustrated here utilizes the techniques of Chaps. 13 and 14, including the use of pseudofactors introduced in Sect. 14.3. The first step in the design construction is to determine an appropriate fraction of the non-blocking factors, then some effects (and their aliases)—some day effects in this case—are chosen to be confounded with blocks. Before constructing the fraction, here are some preliminary considerations.

Recall, the basic experiment was 2×3 in the treatment factors A (display type) and B (visual presentation format), the purpose being accurate comparison of the treatment effects. Multiple subjects were to be recruited for the study based on appropriate protocols. It was pragmatically reasonable to have each subject make three runs per day and to involve each subject for two days. This would allow each subject to be assigned each of the six treatment combinations once, providing a balanced design with treatment comparisons as within-subject comparisons, good for precision. Since day-to-day variation was anticipated to be relatively small and simplicity of the experimental design is desirable, it was decided to confound main effects of A with days. Viewing subjects as blocks, days as whole plots, and runs as split plots, levels of A would be assigned to whole plots and levels of B to split plots.

There remains the question of how many days $2h$ to include in the experiment, (or equivalently, how many subjects, h). Consider this question from the viewpoint of design construction rather than as a sample size question, to see how small an appropriate design can be given the circumstances. For a 2×3 experiment, standard confounding techniques can be applied if the number of subjects is a power of two times a power of three, $h = 2^m 3^n$ say, so represent the subjects factor S by pseudofactors S_{2i} ($i = 1, \ldots, m$) each at two levels and pseudofactors S_{3j} ($j = 1, \ldots, n$) each at three levels, where the number of each remains to be determined. The other nuisance sources of variation are days D, with two days for each subject, and run order O and path P each at six levels. Let O_2, O_3, P_2 and P_3 be pseudofactors for run order and path, with the subscript indicating the number of levels. Likewise,

denote pseudofactors for days by D_{2i} ($i = 0, 1, \ldots, m$) each at two levels and D_{3j} ($j = 1, 2, \ldots, n$) each at three levels. So, factors at two levels are A_2, the D_{2i}, O_2, P_2, and the S_{2i}, each with levels 0 and 1, say. Similarly, factors at three levels are B_3, the D_{3j}, O_3, P_3, and the S_{3j}, each with levels 0, 1 and 2, say.

In the resulting design, take the treatment factor levels to be: $A = A_2 + 1$ and $B = B_3 + 1$. Furthermore, let order $O = 3 \times O_2 + O_3 + 1$, so the levels of O_2 distinguish a subject's three runs on their first day (i.e. $O = 1, 2, 3$) from the subject's three runs on their second day (i.e. $O = 4, 5, 6$), and the levels of O_3 distinguish the three runs on each day. Similarly, let path $P = 3 \times P_2 + P_3 + 1$, with the understanding that the levels of P_2 distinguish whether a path is run frontwards ($P_2 = 0$, say) or backwards ($P_2 = 1$, say), and the levels of P_3 correspond to the three physical paths. Similar relationships between factors and pseudofactors for days and subjects will be provided later.

The design is a fractional factorial block design. Following the approach introduced in Sect. 15.5, the first step is to determine an appropriate fractional factorial design in the non-blocking factors, then determine what effects to confound with blocks. That said, subjects will not participate on the same day, so comparisons between subjects will necessarily be comparisons between days, so the corresponding between-days comparisons will be confounded with subjects. Let the levels of D_{20} represent the two days each subject participates, so D_{20} will not be confounded with subjects, but anticipate that the remaining days pseudofactors will be confounded with subjects.

The experiment includes one observation for each combination of day and O_3. Thus, one can view the day pseudofactors D_{2i} ($i = 0, \ldots, m$) and D_{3j} ($j = 1, \ldots, n$) and the order pseudofactor O_3 as *free variables* in constructing the fractional factorial design, with all other pseudofactors defined in relation to them using defining relations.

Consider first the remaining 2-level pseudofactors—A_2, O_2, and P_2—in relation to free variables D_{2i}. Run order is necessarily related to days, since each subject makes their first three runs on their first day and their last three runs on their second day. This corresponds to the constraint $O_2 = D_{20}$, or equivalently $I_2 = D_{20}O_2$.

To avoid aliasing the effects of A_2 with either order or path, the experimenter chose to let $A_2 = D_{21} + D_{22} + O_2$ (mod 2) and $P_2 = A_2 + D_{22}$ (mod 2), so the defining relation for the 2-level pseudofactors is

$$
\begin{aligned}
I_2 \quad &= \underline{D_{20}O_2} \quad &&= \underline{A_2 D_{21} D_{22} O_2} = A_2 D_{20} D_{21} D_{22} \\
&= \underline{A_2 D_{22} P_2} = A_2 D_{20} D_{22} O_2 P_2 &&= D_{21} O_2 P_2 \quad = D_{20} D_{21} P_2 ,
\end{aligned}
$$

with defining relation generators underlined, the other terms being generalized interactions.

With these choices used by the experimenter, using only three 2-level day pseudofactors D_{20}, D_{21} and D_{22}, the fraction of the 2-level pseudofactors is a 1/8 fraction with eight runs involving eight days. Hence, as will be seen, the number of days in the full experiment is a multiple of eight. Technically, the fraction for the 2-level pseudofactors is not of resolution III, because day effects D_{2i} are aliased with the effects of three factors: A, since $A_2 = D_{20} D_{21} D_{22}$; order, since $O_2 = D_{20}$; and path, since $P_2 = D_{20} D_{21}$. However, the day effects are random, modeling variability but with mean zero. So, the effects of A, O_2 and P_2, being "aliased" with the random effects of days, are comparisons between days (i.e. between whole plots), so these effects are estimable, but estimation and testing of these effects will involve the whole-plot mean squared error.

Now consider the remaining pseudofactors at three levels—namely, B_3 and P_3—to be defined in relation to the free variables O_3 and the D_{3j}. Four 3-level factors are necessary and sufficient to obtain a resolution III fraction, so one 3-level days pseudofactor D_{31} will suffice. The experimenter chose to let $B_3 = 2D_{31} + O_3$ (mod 3) and $P_3 = D_{31} + O_3$ (mod 3). Thus, the defining relation for the resulting 1/9 fraction for the 3-level pseudofactors is

$$I_3 \quad = \underline{B_3 D_{31} O_3^2} = B_3^2 D_{31}^2 O_3$$
$$= \underline{D_{31} O_3 P_3^2} = B_3 D_{31}^2 P_3^2 = B_3^2 O_3^2 P_3^2$$
$$= D_{31}^2 O_3^2 P_3 = B_3 O_3 P_3 \quad = B_3^2 D_{31} P_3,$$

involving the contrasts $(B; B^2), (B_3 D_{31} O_3^2; B_3^2 D_{31}^2 O_3), (D_{31} O_3 P_3^2; D_{31}^2 O_3^2 P_3), (B_3 D_{31}^2 P_3^2; B_3^2 D_{31} P_3)$, and $(B_3 O_3 P_3; B_3^2 O_3^2 P_3^2)$.

Since the effects of B_3, O_3 and P_3 are not aliased with other main effects, and in particular not with day effects, they are estimable within-day comparisons under a main-effects model, free of day-to-day (i.e. whole-plot-to-whole-plot) variation, as desired.

With these choices used by the experimenter, using only one 3-level day pseudofactor D_{31}, the fraction of the 3-level pseudofactors is a 1/9 resolution III fraction involving nine runs, including three days and three runs per day (corresponding to the nine combinations of the free variables D_{31} and O_3).

The 2-level and 3-level fractions are combined by pairing each of the eight 2-level pseudofactor combinations in the 2-level fraction with each of the nine 3-level pseudofactor combinations in the 3-level fraction, providing a *composite fractional factorial design* with 72 pseudofactor combinations or runs, 24 days, six run orders, and six paths. The corresponding *composite defining relation* includes 72 terms, including each product of one of the eight terms from the defining relation of the 2-level fraction with one of the nine terms from the defining relation of the 3-level fraction. In short, letting $I_6 = I_2 \times I_3$, and treating I_2 and I_3 as multiplicative identities, the composite defining relation is as follows.

$\times$	I_2	$= D_{20} O_2$	$= \ldots = D_{20} D_{21} P_2$
I_3	I_6	$= D_{20} O_2$	$= \ldots = D_{20} D_{21} P_2$
$= B_3 D_{31} O_3^2$	$= B_3 D_{31} O_3^2$	$= D_{20} O_2 B_3 D_{31} O_3^2$	$= \ldots = D_{20} D_{21} P_2 B_3 D_{31} O_3^2$
$= B_3^2 D_{31}^2 O_3$	$= B_3^2 D_{31}^2 O_3$	$= D_{20} O_2 B_3^2 D_{31}^2 O_3$	$= \ldots = D_{20} D_{21} P_2 B_3^2 D_{31}^2 O_3$
$= \vdots$	$= \vdots$	$= \vdots$	$= \vdots = \vdots$
$= B_3^2 D_{31} P_3$	$= B_3^2 D_{31} P_3$	$= D_{20} O_2 B_3^2 D_{31} P_3$	$= \ldots = D_{20} D_{21} P_2 B_3^2 D_{31} P_3$

This composite defining relation includes each of the terms from the 2-level and 3-level defining relations, indicating for example that days are still aliased with the effects of A, O_2 and P_2 in the composite fraction, as was the case in the 2-level fraction, so A and one degree of freedom for each of order and path remain between-day comparisons in the composite fraction. However, the remaining terms in the composite defining relation are the generalized interactions that include terms other than I_2 and I_3 from each of the 2-level and 3-level defining relations, so they tend to be of higher order. Consequently, one may observe for example that AB effects are *not* aliased with day effects, so they are within-day (i.e. within-whole-plot) comparisons. The same is true of the order and path effects except O_2 and P_2, so order and path each have four degrees of freedom that are within-day comparisons, in addition to each having one degree of freedom that is a between-day (i.e. between-whole-plot) comparison.

Finally, the block design is obtained by confounding the day effects D_{21}, D_{22} and D_{31} with subjects, which also confounds the corresponding generalized interactions $D_{21} D_{22}$, $D_{21} D_{31}$, $D_{22} D_{31}$ and $D_{21} D_{22} D_{31}$, confounding 11 degrees of freedom for days with subjects. The *composite design* involves 12 subjects and 24 days, with three runs per day, so 72 runs in total. The resulting design is shown in Table 19.11, with factor levels in the table obtain as follows: $D = D_{20} + 1$ and Subj $= 6 S_{21} + 3 S_{22} + S_{31} + 1$.

In summary, 12 subjects is enough to design the experiment so treatment effects are not aliased with subject, path, or run order, though effects of A (display type) are aliased with days (whole plots), as is one degree of freedom each for order and path. One might say that the design is of resolution III in the fixed effects but of resolution II when random effects are considered.

The experimenter determined that this was a large enough experiment, with each pairwise comparison of visual presentation formats being estimable with variance $\sigma_S^2/12$, and with the pairwise comparison of display types being estimable with variance $(3\sigma_W^2 + \sigma_S^2)/18$.

If a larger design was needed, one could be obtain by inclusion of additional free variables D_{2i} ($i > 2$) or D_{3j} ($j > 1$) and corresponding additional subject pseudofactors, coupled with an appropriate choice of defining relations to determine all non-free variables in terms of the free variables. For example, to have two replicates of the design obtained above, simply include the additional free variable D_{23} to double the number of days, include the pseudofactor S_{23} to double the number of subjects, impose the same constraints as above, and confound D_{23} with subjects. Alternatively, one could seek a better fraction—namely, either one of higher resolution or, short of that, a fraction with less aliasing together of lower order effects of interest—by inclusion of additional free variables D_{2i} or D_{3j} and corresponding subject pseudofactors, coupled with some better choice of defining relations that presumably would involve these additional free variables.

The approach used above to obtain a blocked fractional factorial experiment involved two steps: first obtaining a fraction of the non-blocking factors, then confounding some effects with blocks. Another equivalent approach is simply to obtain an appropriate fraction of all of the factor combinations, including the blocking factors as non-free variables. For example, one could obtain the same blocked fractional factorial design in one step by imposing the same defining relations as above plus the relations $S_{21} = D_{21}$, $S_{22} = D_{22}$, and $S_{31} = D_{31}$.

19.8   Using SAS Software

In this section, we illustrate use of the SAS software for analyzing split-plot designs. The analysis of variance approach can be used for balanced data using `PROC GLM` or `PROC MIXED`, as illustrated for analysis of the oats experiment in Sect. 19.8.1. In Sect. 19.8.2, the UAV experiment is briefly revisited to illustrate the analysis of simple contrasts using the `SLICE` statement in `PROC MIXED`. In Sect. 19.8.3, we introduce a restricted maximum likelihood approach to analysis of split plot designs using `PROC MIXED`, comparing it to the analysis of variance approach and `PROC GLM` in the context of the UAV switch experiment—a new example involving a negative variance component estimate. In Sect. 19.8.4, using a subset of the oats data yielding a balanced incomplete block design for the whole-plots factor, `PROC MIXED` is used to illustrate the recovery of inter-block information. Finally, in Sect. 19.8.5, analysis of unbalanced data is illustrated using `PROC MIXED` and the restricted maximum likelihood approach for the mobile computing field study, excluding the missing observation from analysis of the split plot design.

19.8.1   The Analysis of Variance Approach

The *analysis of variance approach* to the analysis of mixed models involves: (i) fitting the model by ordinary least squares, treating the random effects as fixed; and (ii) using the expected mean squares to determine unbiased estimates of the variance components. In analyzing a balanced split-plot design by the analysis of variance approach, care needs to be taken with `PROC GLM` or `PROC MIXED` in order to obtain the two separate parts of the analysis of variance table, corresponding to

Table 19.15 SAS program illustrating type I and III analyses—the oats split-plot experiment

```
DATA OAT;
  INPUT BLOCK WP A B  Y;
  LINES;
  1   1  2 3 156
  1   1  2 2 118
  1   1  2 1 140
  1   1  2 0 105
  1   2  0 0 111
  :   :  : :  :
  6   3  2 3 121
;
*** Analysis of variance: type I;
PROC GLM;
  CLASS BLOCK A B WP;
  MODEL Y = BLOCK A WP(BLOCK)  B A*B / E1;
  RANDOM BLOCK WP(BLOCK) / TEST;
  MEANS A / DUNNETT('0') CLDIFF ALPHA = 0.01;
  * The following statement fails to generate output due to
    nonestimability;
  LSMEANS A / ADJUST = DUNNETT ALPHA = 0.01 CL PDIFF = CONTROL('0')
              E = WP(BLOCK);
  LSMEANS B / ADJUST = DUNNETT ALPHA = 0.01 CL PDIFF = CONTROL('0');
PROC MIXED METHOD = TYPE1;
  CLASS BLOCK A B WP;
  MODEL Y = A B A*B / DDFM = SAT;
  RANDOM BLOCK WP(BLOCK);
  LSMEANS A B / PDIFF = CONTROL('0', '0')
                CL DIFF ADJUST = DUNNETT ALPHA = 0.01;
*** Analysis of variance: type III;
PROC GLM;
  CLASS BLOCK A B;
  MODEL  Y = BLOCK A BLOCK*A  B A*B;
  RANDOM BLOCK A*BLOCK / TEST;
  LSMEANS A / ADJUST = DUNNETT ALPHA = 0.01 CL PDIFF = CONTROL('0')
              E = BLOCK*A;
  LSMEANS B / ADJUST = DUNNETT ALPHA = 0.01 CL PDIFF = CONTROL('0');
PROC MIXED METHOD = TYPE3;
  CLASS BLOCK A B;
  MODEL Y = A B A*B / DDFM = SAT;
  RANDOM BLOCK A*BLOCK;
  LSMEANS A B / PDIFF = CONTROL('0', '0')
                CL DIFF ADJUST = DUNNETT ALPHA = 0.01;
```

between-whole-plots and within-whole-plots comparisons. We illustrate two different methods of producing an analysis of variance table similar to the one in Table 19.4 (p. 710) presented for the oats experiment in Sect. 19.3.4. Table 19.15 contains SAS code illustrating the two methods—namely, the type I and type III analyses—using the data of the oats experiment.

Type I Analysis

One approach to analysis of a split plot design is a *type I analysis*, utilizing the type I (sequential) sums of squares and corresponding type I expected mean squares. In the first call of PROC GLM in Table 19.15, the option E1 in the model statement requests a type I analysis. Correspondingly,

the sources of variation need to be entered into the model in the same order as in model (19.2.2). The first three terms yield the whole-plot analysis. The whole-plot error term $\epsilon_{i(h)}^{W}$ is represented by WP(BLOCK)—the nested effect of whole plots within blocks—since it accounts for any whole plot comparisons not yet modeled. The remaining terms provide the split-plot analysis. No term is needed to represent the split-plot error term $\epsilon_{j(hi)}^{S}$, since this plays the role of the usual error variable, and the corresponding error sum of squares is automatically calculated by the SAS software. The type I analysis is necessary because, with inclusion of WP(BLOCK) in the model, the type III analysis would yield $SSA = 0$ with zero degrees of freedom for A. Otherwise, the type I and type III sums of squares would be the same. Inclusion of the statement

```
RANDOM BLOCK WP(BLOCK) / TEST;
```

ensures that the correct denominators are used for all of the hypothesis tests.

The output is shown in Fig. 19.1. The type I sums of squares are listed, but the F-statistics and p-values are incorrect for testing the effects of blocks and A, since SAS software uses the split-plot error mean square *MSE* throughout. The expected mean squares are listed, and the reader can verify that the error estimate for A differs from those of B and AB. The correct hypothesis tests are listed in the bottom half of the output.

One would like to use the LSMEANS statement to obtain standard multiple comparisons procedures. For example, Dunnett's procedure for comparing each level of A with control level 0 and each level of B with control level 0 is requested via the LSMEANS statements shown in Table 19.15. Unfortunately, this generates no output for A, because PROC GLM treats all effects as fixed, including the random effects, so concludes that main effects of A are not estimable. While we have generally avoided use of the MEANS statement for multiple comparisons, since it uses means rather than least squares means, the two are equivalent in this case, since both the whole-plot design and the split-plot design are randomized complete block designs. As such, the MEANS statement can be used to obtain standard multiple comparison procedures for A, while the LSMEANS statement cannot. For example, Dunnett's procedure for comparing each level of A with control level 0 and each level of B with control level 0 are obtained via the MEANS statement and the second LSMEANS statement, respectively, shown in Table 19.15. For the MEANS statement, SAS software will use the split-plot term *MSE* to estimate standard errors unless told to do otherwise, which would be correct for the B contrasts but not for the A contrasts. To use the whole-plot error mean square for the A contrasts, we include the option E = WP(BLOCK). Partial output is shown in Fig. 19.2. For both factors, note that the SAS software has relabeled the levels starting at 1 rather than 0.

The first call of PROC MIXED in Table 19.15 would generate the same correct information, including only the correct F-tests for the analysis of variance, and automatically computing correct standard errors for contrasts. The option METHOD = TYPE1 calls for a type I analysis. In PROC MIXED, all fixed effects are entered into the model before the random effects, but that is not problematic in this case.

Type III Analysis

The second approach to analysis of a split plot design is a *type III analysis*, utilizing the type III sums of squares and corresponding expected mean squares. This approach makes use of the fact that the whole-plot error sum of squares uses the same collective degrees of freedom as any block by whole-plot-treatment interactions deemed negligible. In the oats experiment, with only one whole-plot factor, A, and with the block $\times A$ interaction deemed negligible, the block $\times A$ effect represents whole-plot error in the model. In the second call of PROC GLM in Table 19.15, the same sources of variation are entered into the SAS model as in model (19.2.2), but the whole plot error $\epsilon_{i(h)}^{W}$ is replaced by BLOCK*A—the negligible block $\times$ whole-plot-treatment interaction. For the type III analysis, the

The GLM Procedure
Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	26	44017.19444	1692.96902	9.56	<.0001
Error	45	7968.75000	177.08333		
Corrected Total	71	51985.94444			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BLOCK	5	15875.27778	3175.05556	17.93	<.0001
A	2	1786.36111	893.18056	5.04	0.0106
WP(BLOCK)	10	6013.30556	601.33056	3.40	0.0023
B	3	20020.50000	6673.50000	37.69	<.0001
A*B	6	321.75000	53.62500	0.30	0.9322

Source	Type I Expected Mean Square
BLOCK	Var(Error) + 4 Var(WP(BLOCK)) + 12 Var(BLOCK)
A	Var(Error) + 4 Var(WP(BLOCK)) + Q(A,A*B)
WP(BLOCK)	Var(Error) + 4 Var(WP(BLOCK))
B	Var(Error) + Q(B,A*B)
A*B	Var(Error) + Q(A*B)

Tests of Hypotheses for Mixed Model Analysis of Variance

	Source	DF	Type I SS	Mean Square	F Value	Pr > F
	BLOCK	5	15875	3175.055556	5.28	0.0124
*	A	2	1786.361111	893.180556	1.49	0.2724
	Error: MS(WP(BLOCK))	10	6013.305556	601.330556		

* This test assumes one or more other fixed effects are zero.

	Source	DF	Type I SS	Mean Square	F Value	Pr > F
	WP(BLOCK)	10	6013.305556	601.330556	3.40	0.0023
*	B	3	20021	6673.500000	37.69	<.0001
	A*B	6	321.750000	53.625000	0.30	0.9322
	Error: MS(Error)	45	7968.750000	177.083333		

* This test assumes one or more other fixed effects are zero.

Fig. 19.1 Type I analysis of variance—the oats split-plot experiment

Least Squares Means
Adjustment for Multiple Comparisons: Dunnett

Standard Errors and Probabilities Calculated Using the Type III MS for BLOCK*A as an Error Term

A	Y LSMEAN	99% Confidence Limits	
0	97.625000	81.761076	113.488924
1	104.500000	88.636076	120.363924
2	109.791667	93.927743	125.655591

		Least Squares Means for Effect A	
i	**j**	**Difference Between Means**	**Simultaneous 99% Confidence Limits for LSMean(i)-LSMean(j)**
2	1	6.875000	-18.122835	31.872835
3	1	12.166667	-12.831168	37.164502

B	Y LSMEAN	99% Confidence Limits	
0	79.388889	70.952864	87.824914
1	98.888889	90.452864	107.324914
2	114.222222	105.786197	122.658247
3	123.388889	114.952864	131.824914

		Least Squares Means for Effect B	
i	**j**	**Difference Between Means**	**Simultaneous 99% Confidence Limits for LSMean(i)-LSMean(j)**
2	1	19.500000	5.878974	33.121026
3	1	34.833333	21.212307	48.454360
4	1	44.000000	30.378974	57.621026

Fig. 19.2   Multiple comparisons—the oats split-plot experiment

order of entry of terms into the SAS model does not matter. The SAS output is not shown, but is identical to that in Fig. 19.1 with WP(BLOCK) replaced by BLOCK*A, and Type I Expected Mean Square replaced by Type III Expected Mean Square. The second call of PROC MIXED in Table 19.15 would generate the same information, but with only the correct F-tests for the analysis of variance, the option METHOD = TYPE3 calling for a type III analysis.

More generally, suppose the experiment included two whole-plot factors A and B and one split-plot factor C, with the AB-combinations assigned to whole plots within blocks constituting a randomized complete block design, and with the levels of C assigned to split plots within whole plots constituting a randomized complete block design. Then, assuming all block by whole-plot-treatment interactions are negligible, the following SAS statements would provide the analogous type III analysis.

Table 19.16 SAS program illustrating analysis of simple contrasts—UAV experiment

```
DATA UAV;
 INPUT A W B C TIME;
 LINES;
 1 1 1 1 29
 1 1 1 2 28
 1 1 1 3 49
 1 1 1 4 46
 1 1 2 1 36
 : : : : ::
 2 8 2 4 22
;
PROC MIXED METHOD = TYPE3;
 CLASS A W B C;
 MODEL TIME = A | B | C / DDFM = SAT;
 RANDOM W(A);
 LSMEANS A / CL DIFF ALPHA = 0.01;
 * Compares levels of A at each B*C combination;
 SLICE A*B*C / SLICEBY = B*C CL DIFF ALPHA = 0.01;
;
PROC GLM;
 CLASS A W B C;
 MODEL TIME = A | B | C W(A);
 RANDOM W(A) / TEST;
 MEANS A / T CLDIFF E = W(A) ALPHA = 0.01;
 * need PDIFF, else E = mse erroneously;
 LSMEANS A / CL PDIFF E = W(A) ALPHA = 0.01;
```

```
PROC GLM;
  CLASS BLOCK A B C;
  MODEL  Y = BLOCK A B A*B BLOCK*A*B  C A*C B*C A*B*C;
  RANDOM BLOCK BLOCK*A*B / TEST;
  LSMEANS A B / ADJUST = DUNNETT ALPHA = 0.01 CL
               PDIFF = CONTROL('0', '0') E = BLOCK*A*B;
  LSMEANS C / ADJUST = DUNNETT ALPHA = 0.01 CL PDIFF = CONTROL('0');
```

In this case, the whole-plot error sum of squares uses the collective degrees of freedom of the block $\times A$, block $\times B$, and block $\times A \times B$ interactions. These are collected into the single term BLOCK*A*B by including it in the model but leaving out the terms BLOCK*A and BLOCK*B.

19.8.2   Simple Contrasts

In this section, we briefly revisit the UAV experiment of Sect. 19.6 to illustrate the analysis of simple contrasts using the SLICE statement in PROC MIXED. The SAS program in Table 19.16 shows how either PROC MIXED or PROC GLM can be used to generate the analyses provided in Sect. 19.6. For example, both procedures generate the analysis of variance in Table 19.10, as well as the 99% confidence interval for the main effect of cue (A).

PROC MIXED generally provides more options and functionality for multiple comparisons. For example, the SLICE statement illustrated in the SAS program generates the output in Table 19.17, which includes the information on simple contrasts for cue that was provided in Sect. 19.6.2 (p. 717), comparing the effects of the two levels of cue (A) at each of the eight BC combinations. The standard

Table 19.17 SAS output for simple contrasts—UAV experiment

```
                              The SAS System
                            The Mixed Procedure

                 Simple Differences of A*B*C Least Squares Means

                              Standard              t
     Slice      A _A Estimate   Error    DF    Value  Pr>|t| Alpha    Lower    Upper
     B 1 C 1    1 2   14.3750  1.8643 110.2     7.71  <.0001  0.01   9.4885  19.2615
     B 1 C 2    1 2   13.3750  1.8643 110.2     7.17  <.0001  0.01   8.4885  18.2615
     B 1 C 3    1 2   31.2500  1.8643 110.2    16.76  <.0001  0.01  26.3635  36.1365
     B 1 C 4    1 2   28.2500  1.8643 110.2    15.15  <.0001  0.01  23.3635  33.1365
     B 2 C 1    1 2   13.0000  1.8643 110.2     6.97  <.0001  0.01   8.1135  17.8865
     B 2 C 2    1 2   15.2500  1.8643 110.2     8.18  <.0001  0.01  10.3635  20.1365
     B 2 C 3    1 2   21.2500  1.8643 110.2    11.40  <.0001  0.01  16.3635  26.1365
     B 2 C 4    1 2   23.7500  1.8643 110.2    12.74  <.0001  0.01  18.8635  28.6365
```

errors involve composite variance estimates, and Satterthwaite's approximation is used as a consequence of the denominator degrees of freedom option DDFM = SAT. The SLICE statement uses much of the same syntax as the LSMEANS statement in PROC MIXED. Analogous results do not seem to be available from PROC GLM.

19.8.3   The Restricted Maximum Likelihood Approach

In this section, we introduce a restricted maximum likelihood (ReML) approach to analysis of split plot designs using PROC MIXED, comparing it to the analysis of variance approach in the context of a new example—UAV switch experiment. The two approaches yield the same results if (i) the design is balanced and (ii) all variance component estimates are either positive or allowed to be negative. The restricted maximum likelihood approach is generally preferable for (i) the estimation and testing of variance components and (ii) the analysis of fixed effects given sufficiently unbalanced data. The UAV switch experiment provides an interesting comparison of the approaches, because the design is balanced but a variance component estimate is negative if unconstrained.

The analysis of variance approach was illustrated in Sects. 17.10.2, 18.5, 19.8.1 and 19.8.2 for balanced data using both PROC MIXED and PROC GLM. Indeed, the analysis of variance approach can be reasonable and appropriate for the analysis of balanced data, as the fixed effect estimates are then best linear unbiased estimates, and the variance component estimates are then minimum variance unbiased estimates under normality. That said, variance component estimates can be negative even for balanced data. This may be reasonable for analysis of fixed effects, but it seems problematic if one is interested in inferences on variance components. Moreover, for unbalanced data, the fixed effects estimates can be inefficient, and the properties of the ANOVA-based variance component estimates are not well understood.

As an alternative to the analysis of variance approach, consider the use of maximum likelihood estimates, which generally have good large-sample properties. Under appropriate regularity conditions, the maximum likelihood estimator $\hat{\theta}$ of an estimable parameter θ is asymptotically $N(\theta, \text{CRLB})$, where CRLB is the Cramer–Rao lower bound for the variance of an unbiased estimator. Mind you, for the regularity conditions to hold, the estimate must be a solution to the likelihood equations—not obtained

as a boundary condition—so these asymptotic properties do not apply for example to a variance component estimator when the estimate is constrained to be zero because the solution to the likelihood equations is negative. Likewise, given a factor with random effects but with few levels observed in an experiment, one should be skeptical of the asymptotic properties of the corresponding variance component estimate. That said, the asymptotic properties of MLEs should usually be reasonably applicable for the analysis of fixed effects unless an experiment is quite small. We will utilize restricted maximum likelihood estimation—a special case of maximum likelihood estimation.

The *restricted maximum likelihood approach* to fitting a mixed model involves two steps: (i) estimating the variance components by restricted maximum likelihood; then (ii) estimating the fixed effects by *estimated generalized least squares*—namely, treating the variance component estimates as true values, then computing the maximum likelihood estimates of the fixed effects, or equivalently, the *generalized least squares estimates*. To compute restricted maximum likelihood (ReML) estimates of the variance components, the original data is essentially replaced by a maximal set of linearly independent contrasts in the data each with mean zero (i.e. data contrasts the distributions of which do not involve the fixed effects), then the likelihood function of the contrasts is maximized. (Equivalently, one can estimate the fixed effects by ordinary least squares, then maximize the likelihood function of the residuals, the joint distribution of which does not involve the fixed effects. Hence, the restricted maximum likelihood approach is also known as *residual maximum likelihood*.) The ReML estimates of the variance components may be preferable to the usual maximum likelihood estimates, because the ReML estimates are the same as the ANOVA-based estimates if the data are balanced and all variance component estimates are positive. Also, restricted maximum likelihood adjusts in some sense for fixed effects, so often provides unbiased estimates of variance components. A new experiment will be introduced to illustrate the restricted maximum likelihood approach to analysis of split plot designs.

UAV Switch Experiment

Mahadevan (2009) conducted three experiments to evaluate the performance of a semi-automated computer display system designed to support a human operator's ability to monitor and control the complex dynamic operation of multiple unmanned aerial vehicles (UAVs) when the UAVs are involved in multiple combat-related tasks. His third experiment involved: 16 subjects (factor W); two alert techniques—a visual alert ($A = 1$) and an audio-visual alert ($A = 2$); and two levels of task complexity—a simple primary task coupled with a simple secondary task ($B = 1$), and a complex primary task coupled with a simple secondary task ($B = 2$). The experiment was a 2×2 split-plot design, with subjects as whole plots and two trials per subject as split-plots, with each level of A assigned to half of the subjects, and with both levels of B observed once on each subject. Hence, subject is nested within alert type. For each trial, each subject, while working on the primary task, was warned of the secondary task using one of the two alert techniques. One of the response variables measured was the time taken to switch to the secondary task following the alert, and the experimenter was interested in the effect of the nature of the alert (A) on the mean time to switch from the primary to the secondary task. The data are shown in Table 19.18. The model used by the experimenter is as follows.

$$
\begin{aligned}
Y_{iuj} = \mu + \alpha_i + \epsilon_{iu}^W & \\
+ \beta_j + (\alpha\beta)_{ij} + \epsilon_{j(iu)}^S \,, & \\
\epsilon_{iu}^W \sim N(0, \sigma_W^2)\,, \quad \epsilon_{j(iu)}^S \sim N(0, \sigma_S^2)\,, & \\
\epsilon_{iu}^W \text{'s and } \epsilon_{j(iu)}^S \text{'s are all mutually independent,} & \\
i = 1, 2; \quad u = 1, \dots, 8; \quad j = 1, 2. &
\end{aligned}
\tag{19.8.20}
$$

Table 19.18 Time taken (in seconds) to switch to secondary task for the UAV switch experiment

A (Alert Type)	B (Complexity)	W (Subject)							
		1	2	3	4	5	6	7	8
1	1	6	5	5	5	5	7	5	6
	2	16	22	16	20	12	18	16	14
2	1	7	6	5	6	5	6	4	4
	2	6	7	6	8	6	6	7	6

Source Mahadevan (2009). Copyright © 2009 Sriram Mahadevan. Reprinted with permission

The SAS program for the UAV switch experiment is shown in Table 19.19. The first calls of PROC GLM and PROC MIXED each fit model (19.8.20), and some of the corresponding output is shown in Fig. 19.3. The random effects term W(A) in each SAS model represents the whole plot errors ϵ_{iu}^W in model (19.8.20), so $\sigma_{W(A)}^2$ in the SAS output corresponds to σ_W^2 in our model. Also, msE in SAS output corresponds to msE_S in our formulae. PROC MIXED is by default fitting the model by ReML but, because of the NOBOUND option, the variance component estimates are not constrained to be positive. In fact, the whole-plots variance component estimate, $\hat{\sigma}_{W(A)}^2 = -0.03571$, is negative, matching the estimate one could compute by hand from the PROC GLM output—namely, $\hat{\sigma}_{W(A)}^2 = [ms(W(A)) - msE]/2 \approx -0.036$. Because the data are balanced and the ReML estimates are allowed to be negative, the two procedures provide the same estimates of fixed effects and variance components, the same test results for fixed effects, and the same results for multiple comparisons (not shown).

Consider in comparison the output generated by the second calls of these procedures, shown in Fig. 19.4. PROC MIXED is still fitting model (19.8.20), but the restricted maximum likelihood estimate of $\sigma_{W(A)}^2$, now bounded, is zero. Correspondingly, the whole-plots term is effectively removed from the model, with the 14 degrees of freedom associated with it essentially pooled with the 14 degrees of freedom for error. Hence, there are now 28 degrees of freedom for error in the analysis, the (residual) error variance estimate of 3.1205 being the average of ms(W(A)) and msE generated by the first call of PROC GLM. These results match those at the bottom of Fig. 19.4, generated by the second call of PROC GLM, for which the whole-plots term has been removed from the model for sake of illustration. While the multiple comparisons output generated by these procedure calls is not shown, the reader may confirm that the results would also match, as they do for the F-tests, with both procedures using 28 error degrees of freedom for all contrast standard errors.

The above information begs the question, "What is the correct analysis?" The analyses provided by the first calls of PROC GLM and PROC MIXED are arguably correct, except one should use the multiple comparisons results from PROC MIXED in any instances where PROC GLM uses incorrect standard errors. Otherwise, the SAS procedures correctly implement the planned statistical analyses using the proposed model. If the model is reasonable and appropriate, then the statistical results are valid, even if a simpler model with $\sigma_{W(A)}^2 = 0$ suggested by the data would also be correct.

It is an open problem whether the analyses generated by the second calls of PROC MIXED or PROC GLM are correct—namely, whether they control error rates for any preplanned analysis, assuming the originally posed model (19.8.20) is correct. It seems *improper* for example to fit the first model using PROC GLM, see that the estimate $\hat{\sigma}_{W(A)}^2$ is negative, so change the model by removing the whole-plots term from the model, then fit the reduced model as in the second call of PROC GLM and use these results for the analysis. Control of error rates is an open problem when one uses the data to determine the model then uses the model to analyze the same data. Still, it is interesting that the second call of PROC MIXED fits and conducts the analysis under the assumption that model (19.8.20) is correct and, while it so happens that $\hat{\sigma}_{W(A)}^2 = 0$, the originally postulated model is used for the analysis. If this approach could be shown to control error rates, then this would be the preferred analysis, since there are

Table 19.19 SAS program for the UAV switch experiment

```
DATA UAV3;
 INPUT A W B TIME;
 LINES;
 1 1 1  6
 1 1 2 16
 1 2 1  5
 1 2 2 22
 1 3 1  5
 : : :  :
 2 8 2  6
;
PROC GLM;
 TITLE 'Proc GLM: full model';
 CLASS A W B;
 MODEL TIME = A | B W(A);
 RANDOM W(A) / TEST;
 LSMEANS A / CL PDIFF E = W(A); * need PDIFF, else E = mse;
 LSMEANS A*B; * SAS would erroneously use MSE for CLs;
PROC MIXED NOBOUND;
 TITLE 'Proc Mixed NoBound';
 CLASS A W B;
 MODEL TIME = A | B / DDFM = SAT;
 RANDOM W(A);
 LSMEANS A A*B / CL DIFF;
PROC MIXED;
 TITLE 'Proc Mixed (Bounded)';
 CLASS A W B;
 MODEL TIME = A | B / DDFM = SAT;
 RANDOM W(A);
 LSMEANS A A*B / CL DIFF;
* Dropping W(A) from the Proc GLM model;
PROC GLM;
 TITLE 'Proc GLM: W(A) dropped from model';
 CLASS A B;
 MODEL TIME = A | B;
 LSMEANS A A*B / CL PDIFF;
```

benefits associated with essentially pooling $ms(W(A))$ and msE, including for example increased power for some tests, as well as the availability of a common variance estimator for multiple comparisons of *AB*-treatment combinations.

Given any uncertainty regarding whether the latter approach (corresponding to the second calls) controls error rates, our formal (rather than exploratory) approach to data analysis is to advocate that the experimenter use the analyses generated by the first calls of the procedures, since this approach is known to control error rates for any preplanned analyses if the proposed model (19.8.20) is correct.

19.8.4   Recovery of Inter-block Information

The oats experiment, introduced in Sect. 19.3.4, involves a split-plot design with complete blocks at both the whole- and split-plots levels—namely, the levels of *A* assigned to whole plots within blocks

Fig. 19.3 Output for the UAV switch experiment: Procs GLM and Mixed, First Calls

Proc GLM: full model
The GLM Procedure
Dependent Variable: TIME

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	769.7812500	45.2812500	14.35	<.0001
Error	14	44.1875000	3.1562500		
Corrected Total	31	813.9687500	.		

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	215.2812500	215.2812500	(see test below)	
B	1	306.2812500	306.2812500	97.04	<.0001
A*B	1	205.0312500	205.0312500	64.96	<.0001
W(A)	14	43.1875000	3.0848214	0.98	0.5168

Source	Type III Expected Mean Square
A	Var(Error) + 2 Var(W(A)) + Q(A,A*B)
B	Var(Error) + Q(B,A*B)
A*B	Var(Error) + Q(A*B)
W(A)	Var(Error) + 2 Var(W(A))

Tests of Hypotheses for Mixed Model Analysis of Variance

	Source	DF	Type III SS	Mean Square	F Value	Pr > F
*	A	1	215.281250	215.281250	69.79	<.0001
	Error: MS(W(A))	14	43.187500	3.084821		

* This test assumes one or more other fixed effects are zero.

Proc Mixed NoBound
The Mixed Procedure

Covariance Parameter Estimates	
Cov Parm	**Estimate**
W(A)	-0.03571
Residual	3.1563

Type 3 Tests of Fixed Effects			
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
A	1	14	69.79	<.0001
B	1	14	97.04	<.0001
A*B	1	14	64.96	<.0001

constitute a randomized complete block design, and the levels of *B* assigned to split plots within whole plots constitute a randomized complete block design. With this dual complete block structure, the data analysis is the same whether the block effects are modeled as fixed (as in Sect. 19.3.4) or random (as in Sect. 19.8.4). Such would not be the case if for example the split-plot design involves incomplete blocks at the whole-plots level, as illustrated in this section.

Consider again the oats experiment and corresponding data in Table 19.3 (p. 710), but suppose one only has the data for levels 1 and 2 of *A* in blocks 3 and 5, for levels 0 and 2 of *A* in blocks 1 and 4, and for levels 0 and 1 of *A* in blocks 2 and 6. For this subset of the data, the levels of *A* assigned to whole plots in blocks constitute a balanced incomplete block design with blocks of size 2. The SAS program in Table 19.20 provides several approaches to the analysis of these data.

Fig. 19.4 Output for the
UAV switch experiment:
Procs GLM and Mixed,
Second Calls

In the first call of PROC MIXED, the ANOVA approach is used to conduct a type 3 analysis, and block effects are modeled as fixed. Figure 19.5 contains the corresponding output. Because the block effects are modeled as fixed, the estimates of main-effect-of-A contrasts are *intra-block estimates*—namely, each is composed (by summing over blocks) of within-block contrasts of observations—as is necessary for the fixed block effects to cancel out to yield unbiased estimates. Analogously, for testing for main effects of A, the numerator of the F-statistic is the mean square for A adjusted for block effects, which can be computed from the sum of squares of an appropriate set of such intra-block estimates of main-effect-of-A contrasts. The corresponding data analysis is called the *intra-block analysis*. The second call of PROC MIXED uses the restricted maximum likelihood approach but provides the same intra-block analysis, since the block effects are again modeled as fixed. The output (not shown) would include the same Type 3 Tests of Fixed Effects and Differences of Least Squares Means as in Fig. 19.5. For sake of comparison with subsequent analyses, note that the pairwise comparisons with the control for A each have estimated standard error 11.2883 with 4 degrees of freedom.

Table 19.20 SAS program illustrating the intra-block analysis and also the recovery of inter-block information—the oats split-plot experiment

```
DATA OATBIBD;
 INPUT BLOCK WP A B  Y;
 IF A = 0 AND (BLOCK = 3 OR BLOCK = 5) THEN DELETE;
 IF A = 1 AND (BLOCK = 1 OR BLOCK = 4) THEN DELETE;
 IF A = 2 AND (BLOCK = 2 OR BLOCK = 6) THEN DELETE;
 LINES;
  1 1  2 3  156
  1 1  2 2  118
  1 1  2 1  140
  1 1  2 0  105
  1 2  0 0  111
  : :  : :   :
  6 3  2 3  121
;
*** Intra-block analyses;
PROC MIXED METHOD = TYPE3;
 TITLE 'ANOVA Approach, Type 3, Fixed Block Effects';
 CLASS BLOCK A B;
 MODEL Y = BLOCK A B A*B / DDFM = SAT;
 RANDOM A*BLOCK; * Random whole-plot effects;
 LSMEANS A / PDIFF = CONTROL('0') CL DIFF ADJUST = DUNNETT ALPHA = 0.01;
PROC MIXED METHOD = ReML NOBOUND;
 TITLE 'ReML Approach, Fixed Block Effects';
 CLASS BLOCK A B WP;
 MODEL Y = BLOCK A B A*B / DDFM = SAT;
 RANDOM WP(BLOCK);
 LSMEANS A / PDIFF = CONTROL('0') CL DIFF ADJUST = DUNNETT ALPHA = 0.01;
*** Analyses recovering inter-block information;
PROC MIXED METHOD = TYPE3;
 TITLE 'ANOVA Approach, Type 3, Random Block Effects';
 CLASS BLOCK A B;
 MODEL Y = A B A*B / DDFM = SAT;
 RANDOM BLOCK A*BLOCK;
 LSMEANS A / PDIFF = CONTROL('0') CL DIFF ADJUST = DUNNETT ALPHA = 0.01;
PROC MIXED METHOD = ReML NOBOUND;
 TITLE 'ReML Approach, Random Block Effects';
 CLASS BLOCK A B WP;
 MODEL Y = A B A*B / DDFM = SAT;
 RANDOM BLOCK WP(BLOCK);
 LSMEANS A / PDIFF = CONTROL('0') CL DIFF ADJUST = DUNNETT ALPHA = 0.01;
```

In the third call of PROC MIXED, the ANOVA approach is used to conduct a type 3 analysis, but the block effects are modeled as independent random effects with mean zero and variance σ_θ^2. With random block effects, one can obtain unbiased ordinary least squares estimates of main-effect-of-A contrasts using contrasts of block totals (the sum total of the observations in each block), and these so-called *inter-block estimates* are in addition to and independent of the intra-block estimates. So, for any main-effect-of-A contrast, any fixed weighted average of the corresponding intra- and inter-block estimates provides an unbiased estimate of the contrast, and the best (minimum variance) estimate would be obtained when each weight is inversely proportional to the variance of the corresponding estimate. Unfortunately, this best weighting is unknown, because the variance of the intra-block estimate is a

ANOVA Approach, Type 3, Fixed Block Effects
The Mixed Procedure

				Type 3 Analysis of Variance
Source	DF	Sum of Squares	Mean Square	Expected Mean Square
BLOCK	5	10237	2047.308333	Var(Residual) + 4 Var(BLOCK*A) + Q(BLOCK)
A	2	376.541667	188.270833	Var(Residual) + 4 Var(BLOCK*A) + Q(A,A*B)
B	3	15966	5321.916667	Var(Residual) + Q(B,A*B)
A*B	6	1361.500000	226.916667	Var(Residual) + Q(A*B)
BLOCK*A	4	3058.208333	764.552083	Var(Residual) + 4 Var(BLOCK*A)
Residual	27	4197.750000	155.472222	Var(Residual)

Source	Error Term	Error DF	F Value	Pr > F
BLOCK	MS(BLOCK*A)	4	2.68	0.1806
A	MS(BLOCK*A)	4	0.25	0.7928
B	MS(Residual)	27	34.23	<.0001
A*B	MS(Residual)	27	1.46	0.2294
BLOCK*A	MS(Residual)	27	4.92	0.0041
Residual	.			

	Type 3 Tests of Fixed Effects			
Effect	Num DF	Den DF	F Value	Pr > F
BLOCK	5	4	2.68	0.1806
A	2	4	0.25	0.7928
B	3	27	34.23	<.0001
A*B	6	27	1.46	0.2294

			Differences of Least Squares Means				
Effect	A	_A	Estimate	Standard Error	DF	Adjustment	Adj Lower	Adj Upper
A	1	0	6.0417	11.2883	4	Dunnett-Hsu	-54.5061	66.5894
A	2	0	7.4583	11.2883	4	Dunnett-Hsu	-53.0894	68.0061

Fig. 19.5 SAS output illustrating the intra-block analysis—the oats split-plot experiment

multiple of $\sigma^2 + 4\sigma_W^2$, the variance of the inter-block estimate is a multiple of $\sigma^2 + 4\sigma_W^2 + 8\sigma_\theta^2$, and these variances are unknown. A common approach is to use the estimated variances to determine the weights, but random weights would make the resulting analyses approximate. If the weights are well chosen, the resulting combined estimate is better (has smaller variance) than the intra-block estimate, due to recovery of inter-block information.

The output corresponding to the third call of PROC MIXED is in Fig. 19.6. Because a type 3 analysis is requested, the model is initially fit by ordinary least squares, providing a corresponding Type 3 Analysis of Variance identical to that generated by the first call of PROC MIXED, again with effects of A adjusted for block effects. However, this ANOVA information is then used to estimate the variance components, then the variance component estimates are in turn used to re-estimate the fixed

Fig. 19.6 SAS output illustrating the recovery of inter-block information—the oats split-plot experiment; ANOVA approach

ANOVA Approach, Type 3, Random Block Effects
The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate
BLOCK	178.16
BLOCK*A	152.27
Residual	155.47

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
A	2	5	0.48	0.6422
B	3	27	34.23	<.0001
A*B	6	27	1.46	0.2294

Differences of Least Squares Means

Effect	A	_A	Estimate	Standard Error	DF	Adjustment	Adj Lower	Adj Upper
A	1	0	3.5224	10.6837	5	Dunnett-Hsu	-45.8986	52.9434
A	2	0	10.3425	10.6837	5	Dunnett-Hsu	-39.0785	59.7635

effects by estimated generalized least squares. Using this revised information, SAS software provides updated `Type 3 Tests of Fixed Effects`, including an approximate F test for A with 5 rather than 4 denominator degrees of freedom, and revised estimates of the pairwise comparisons with the control for A, each now with estimated standard error 10.6837 with 5 degrees of freedom. The smaller standard error and increased number of degrees of freedom are both helpful consequences of the recovered inter-block information, increasing test power and decreasing confidence interval width.

In the fourth call of `PROC MIXED`, the restricted maximum likelihood approach is used for the analysis, with the block effects again modeled as independent random effects with mean zero and variance σ_θ^2. The output corresponding to the third call of `PROC MIXED` is in Fig. 19.7. Under this approach, ReML estimates of the variance components are computed, then these are used to estimate the fixed effects by generalized least squares. Based on these estimates, SAS software provides `Type 3 Tests of Fixed Effects`, including an approximate F test for A with 4.95 denominator degrees of freedom, and estimates of the pairwise comparisons with the control for A, each with estimated standard error 10.7095 with 4.95 degrees of freedom. These values are slightly worse than the results of the third call of `PROC MIXED`, but still much better than for the intra-block analysis generated by the first two calls. Clearly there is a helpful recovery of inter-block information.

Turning attention to the split-plot comparisons, note that the F-tests for B, and likewise for AB, are the same in each analysis. This is the case because a randomized complete block design was used for the assignment of levels of B to split plots in whole plots (serving as blocks). Consequently, for each main effect of B and AB-interaction contrast, the corresponding contrast in treatment means is a within-whole-plot contrast, free of whole-plot or block effects, so no adjustment is needed.

Fig. 19.7 SAS output
illustrating the recovery of
inter-block information—
the oats split-plot
experiment; ReML
approach

ReML Approach, Random Block Effects
The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate
BLOCK	178.31
WP(BLOCK)	153.25
Residual	155.47

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
A	2	4.95	0.48	0.6432
B	3	27	34.23	<.0001
A*B	6	27	1.46	0.2294

Differences of Least Squares Means

Effect	A	_A	Estimate	Standard Error	DF	Adjustment	Adj Lower	Adj Upper
A	1	0	3.5161	10.7095	4.95	Dunnett-Hsu	-46.2998	53.3319
A	2	0	10.3497	10.7095	4.95	Dunnett-Hsu	-39.4661	60.1656

In the above example, recovery of the inter-block information provided better information for analysis of the whole-plots factor A. For example, for each treatment-versus-control contrast estimate for A, recovery of inter-block information provided a tighter confidence interval, due to both a smaller standard error and more error degrees of freedom. In planning such an experiment, the experimenter should determine in advance how the data will be analyzed. While one could plan on conducting the intra-block analysis, recovery of the inter-block information should usually be beneficial.

19.8.5   ReML and Unbalanced Data

In this section, we illustrate the restricted maximum likelihood approach using `PROC MIXED` for analysis of an unbalance split-plot design, revisiting the mobile computing field study experiment introduced in Sect. 19.7. Recall, while the planned experiment was nicely balanced, it was discovered prior to analysis of the data that one of the subjects traversed one of the paths in the reverse direction, making the corresponding observation inappropriate to use in the analysis. As such, one of the 72 planned observations (given in Table 19.11, p. 719) is listed as missing. In Sects. 19.7.1 and 19.7.2, an approximate analysis of the split-plot design was provided by estimating the missing value and treating it as observed, using standard formulas for analysis of the balanced design via the analysis of variance approach, but adjusting degrees of freedom appropriately. With one observation lost, the data are no longer balanced, making the data analysis conceptually and computationally more difficult. However, SAS `PROC MIXED` handles this situation nicely, as we shall illustrate.

Table 19.21 SAS program for the mobile computing field study experiment

```
DATA MCFS71;
 INPUT SUBJ DAY ORDER O2 O3 PATH P2 P3 A B Y;
 LINES;
  1 1 1 0 0 1 0 0 1 1 49.321
  1 1 2 0 1 2 0 1 1 2 24.386
  1 1 3 0 2 3 0 2 1 3 37.680
  : : : : : : : : : :   :
 12 1 3 0 2 5 1 1 1 1 30.146
 12 2 4 1 0 3 0 2 2 2   .
 12 2 5 1 1 1 0 0 2 3 58.481
 12 2 6 1 2 2 0 1 2 1 62.659
 ;
 PROC MIXED NOBOUND;
  TITLE 'Analysis of RMSE without missing value';
  CLASS SUBJ DAY O2 O3 P2 P3 A B;
  MODEL Y = O2 P2 A  O3 O2*O3 P3 P2*P3 B A*B / DDFM = SAT;
  RANDOM SUBJ DAY(SUBJ);
  LSMEANS A B / CL DIFF ALPHA = 0.05;
 * To get consolidated tests for order and path;
 PROC MIXED NOBOUND;
  CLASS SUBJ DAY ORDER PATH A B;
  MODEL Y = ORDER PATH A B A*B / DDFM = SAT;
  RANDOM SUBJ DAY(SUBJ);
```

The SAS program for analysis of the 71-observation split-plot design via the restricted maximum likelihood approach is shown in Table 19.21, with the corresponding output in Fig. 19.8. The mixed model for the analysis was provided in Eq. (19.7.18), p. 718. Using PROC MIXED, the three variance components are (by default) estimated by restricted maximum likelihood. In this case, the three estimates are all positive. Given these variance component estimates, the fixed effects are then estimated by generalized least squares.

PROC MIXED provides Type III F-tests for each fixed effect—Type III in the sense that the tests are based on estimates of effects and corresponding variability obtained by fitting the full model. The option DDFM = SAT in the MODEL statement causes use of a general Satterthwaite approximation for the denominator degrees of freedom, equivalent to Satterthwaite's approximation for a balanced design. Readers are referred to SAS documentation for details. Regarding the two treatment factors, only the effects of visual presentation format (B) are significant, with $p = 0.0166$. Concerning the nuisance factors path and order, these involved both within- and between-whole-plot comparisons before the lost observation caused design imbalance. For each of these factors, the first call of PROC MIXED provides tests of each of three pseudofactor components, including for example tests for P_2, P_3 and $P_2 P_3$ for path. The second call of PROC MIXED provides a consolidated test for each of these nuisance factors. It is not surprising that path has significant effects, and these are attributable to the three distinct paths used as distinguished by P_3.

The LSMEANS statement generated the pairwise Differences of Least Squares Means output for each of factors A and B, including individual t-tests and individual 95% confidence intervals for each pairwise comparison. These results are similar to those obtained in Sect. 19.7.2, where the approach used was to estimate the missing value and analyze the balanced design, though there Tukey's method was used. Based on the individual 95% confidence intervals provided here, level 2 of B—the

Analysis of RMSE without missing value
The Mixed Procedure

Covariance Parameter Estimates	
Cov Parm	**Estimate**
SUBJ	63.9909
DAY(SUBJ)	28.1379
Residual	92.9959

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
O2	1	8.5	0.25	0.6316
P2	1	8.5	1.87	0.2064
A	1	8.5	0.19	0.6768
O3	2	34.6	0.04	0.9581
O2*O3	2	34.6	0.37	0.6927
P3	2	34.6	10.62	0.0003
P2*P3	2	34.6	0.14	0.8718
B	2	34.6	4.62	0.0166
A*B	2	34.6	0.79	0.4634

Differences of Least Squares Means											
Effect	**A**	**B**	**_A**	**_B**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**
A	1		2		-1.3630	3.1587	8.5	-0.43	0.6768	0.05	-8.5734	5.8474
B		1		2	8.1958	2.8325	34.8	2.89	0.0065	0.05	2.4442	13.9475
B		1		3	1.6903	2.7838	34.4	0.61	0.5477	0.05	-3.9648	7.3454
B		2		3	-6.5055	2.8325	34.8	-2.30	0.0278	0.05	-12.2572	-0.7539

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
ORDER	5	31.1	0.21	0.9555
PATH	5	31.1	4.70	0.0026

Fig. 19.8 Output for the mobile computing field study experiment

birds' eye view egocentric visual presentation format—has a significantly smaller *RMSE* from the intended path than do the other two visual presentation formats. It is interesting that the denominator degrees of freedom is not constant for the factor *B* pairwise comparisons, due to the imbalance created by the lost observation. This indicates that the pairwise comparisons do not all utilize the same variance estimator. As a consequence, the Bonferroni method should be used if multiple comparisons for *B* are of interest. These comparisons could be generated by adding the option ADJUST = BON to the LSMEANS statement. Replacing the keyword BON by DUNNETT, SCHEFFE or TUKEY would yield the corresponding multiple comparisons method, though these latter methods assume availability of a common variance estimator so are not applicable here.

One could also generate pairwise comparisons for the *AB* combinations by including A*B in the list of effects in the LSMEANS statement. However, only the Bonferroni method of multiple comparisons would be applicable, since there would not be a common variance estimator for all of these pairwise

comparisons. Also, the corresponding variance estimators would generally be composite variance estimators, in which case Satterthwaite's approximation would be used.

19.9   Using R Software

In this section, we illustrate use of the R software for analyzing split-plot designs. The analysis of variance approach can be used for balanced data using `aov`, as illustrated for analysis of the oats experiment in Sect. 19.9.1. In Sect. 19.9.2, the UAV experiment is briefly revisited to illustrate the analysis of simple contrasts by conditioning on other factor level combinations, given a model fit using the `aov` function. In Sect. 19.9.3, we introduce a restricted maximum likelihood approach to analysis of split plot designs using the `lmer` function of the `lme4` package, comparing it to the analysis of variance approach using the `aov` function in the context of UAV switch experiment—a new example involving a negative variance component estimate. In Sect. 19.9.4, the recovery of inter-block information is illustrated, using a subset of the oats data yielding a balanced incomplete block design for the whole-plots factor, and using the `lmer` function to fit the mixed model by restricted maximum likelihood. Finally, in Sect. 19.9.5, analysis of unbalanced data is illustrated using the `lmer` function and restricted maximum likelihood estimation for the mobile computing field study, excluding the missing observation from analysis of the split plot design.

19.9.1   The Analysis of Variance Approach

The *analysis of variance approach* to the analysis of mixed models involves: (i) fitting the model by ordinary least squares, treating the random effects as fixed; and (ii) using the expected mean squares to determine unbiased estimates of the variance components. In analyzing a balanced split-plot design by the analysis of variance approach, the `aov` function automatically provides correct F tests of fixed effects, taking into account the two separate parts of the analysis of variance table, corresponding to between-whole-plots and within-whole-plots comparisons. Table 19.22 contains R code and output, providing an analysis of variance table similar to the one in Table 19.4 (p. 710) presented for the oats experiment in Sect. 19.3.4.

The approach used by `aov` in this analysis of a split plot design is a *type III analysis*, utilizing the type III sums of squares and corresponding expected mean squares. In the call of `aov` in Table 19.22, the model explicitly includes all sources of variation as in model (19.2.2) except the split-plot error term $\epsilon^S_{j(hi)}$, which plays the role of the usual error variable. The whole-plot error term $\epsilon^W_{i(h)}$ is represented by `fWP:fBlock`—the random effects of whole plots nested within blocks—entered into the model via the `Error` function used to designate random effects. The term `fBlock` is likewise included in the `Error` function to designate the block effects as random.

Table 19.23 continues the R program and output of Table 19.22, illustrating the use of `lsmeans` to implement standard multiple comparisons procedures. For example, Dunnett's procedure for comparing each level of A with control level 0 and each level of B with control level 0 is requested via the `lsmeans` and `summary(contrast`… statements shown in Table 19.23. The syntax `ref=1` designates the first level of a factor as the reference level or control, 0 being the first level of both A and B. R automatically uses the correct standard error estimates, using the split-plot term *MSE* to estimate standard errors for the B contrasts, while using the whole-plot error mean square represented by `fBlock:fWP` for the A contrasts.

Table 19.22 R program and output illustrating type 3 analysis of variance, including tests for fixed effects—the oats split-plot experiment

```
> oats.data = read.table("data/oats.txt", header=T)
> oats.data = within(oats.data, {fBlock = factor(Block);
+              fWP = factor(WP); fA = factor(A); fB = factor(B) })
> head(oats.data, 3)
  Block WP A B   y fB fA fWP fBlock
1     1  1 2 3 156  3  2   1      1
2     1  1 2 2 118  2  2   1      1
3     1  1 2 1 140  1  2   1      1

> # Least squares ANOVA
> # Set contrast options for correct lsmeans and contrasts
> options(contrasts = c("contr.sum", "contr.poly"))
> model1 = aov(y ~ fA + fB + fA:fB + Error(fBlock + fWP:fBlock),
+              data=oats.data)
> summary(model1)

Error: fBlock
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  5  15875    3175

Error: fBlock:fWP
          Df Sum Sq Mean Sq F value Pr(>F)
fA         2   1786     893    1.49   0.27
Residuals 10   6013     601

Error: Within
          Df Sum Sq Mean Sq F value  Pr(>F)
fB         3  20021    6674    37.7 2.5e-12
fA:fB      6    322      54     0.3    0.93
Residuals 45   7969     177
```

19.9.2   Simple Contrasts

In this section, we briefly revisit the UAV experiment of Sect. 19.6 to illustrate the analysis of simple contrasts by conditioning on other factor level combinations. The R program in Table 19.24 shows how to generate the analyses provided in Sect. 19.6 and includes selected output. The aov and summary functions generate the analysis of variance in Table 19.10 (not shown here). The lsmeans and summary(contrast...) functions are used to generate the 99% confidence interval for the main effect of cue (A). Then they are used a second time to generate information for the simple contrasts for cue, analogous to what was provided in Sect. 19.6.2 (p. 716), comparing the effects of the two levels of cue (A) at each of the eight BC combinations. For the simple contrasts, the syntax "fA | fC:fB" says to compare the levels of A given each combination of levels of B and C. The corresponding standard errors involve composite variance estimates, and Satterthwaite's approximation is used by R by default. Output for three of the eight simple contrasts is in the bottom of Table 19.24.

Table 19.23   Multiple comparisons—the oats split-plot experiment

```
> # Multiple comparisons: Dunnett's method
> library(lsmeans)
> lsmA = lsmeans(model1, ~ fA)
> set.seed(21531957)
> summary(contrast(lsmA, method="trt.vs.ctrl", adjust="mvt", ref=1),
+          infer=c(T,T), level=0.99, side="two-sided")

 contrast estimate      SE df lower.CL upper.CL t.ratio p.value
 1 - 0        6.875 7.0789 10  -18.123   31.873   0.971  0.5423
 2 - 0       12.167 7.0789 10  -12.831   37.164   1.719  0.1973

Results are averaged over the levels of: fB
Confidence level used: 0.99
Conf-level adjustment: mvt method for 2 estimates
P value adjustment: mvt method for 2 tests

> lsmB = lsmeans(model1, ~ fB)
> set.seed(21531957)
> summary(contrast(lsmB, method="trt.vs.ctrl", adjust="mvt", ref=1),
+          infer=c(T,T), level=0.99, side="two-sided")

 contrast estimate      SE df lower.CL upper.CL t.ratio p.value
 1 - 0       19.500 4.4358 45   5.8789   33.121   4.396  0.0002
 2 - 0       34.833 4.4358 45  21.2122   48.454   7.853  <.0001
 3 - 0       44.000 4.4358 45  30.3789   57.621   9.919  <.0001

Results are averaged over the levels of: fA
Confidence level used: 0.99
Conf-level adjustment: mvt method for 3 estimates
P value adjustment: mvt method for 3 tests
```

19.9.3   The Restricted Maximum Likelihood Approach

In this section, we introduce a restricted maximum likelihood (ReML) approach to analysis of split plot designs using the `lmer` function of the `lme4` package, comparing the restricted maximum likelihood approach to the analysis of variance approach in the context of a new example—UAV switch experiment. The two approaches yield the same results if (i) the design is balanced and (ii) all variance component estimates are either positive or allowed to be negative, though the `lmer` function does not allow variance component estimates to be negative. The restricted maximum likelihood approach is generally preferable for (i) the estimation and testing of variance components and (ii) the analysis of fixed effects given sufficiently unbalanced data. The UAV switch experiment provides an interesting comparison of the approaches, because the design is balanced but a variance component estimate is negative if unconstrained.

The analysis of variance approach was illustrated in Sects. 17.11.2, 18.6, 19.9.1 and 19.9.2 for balanced data using the `aov` function. Indeed, the analysis of variance approach can be reasonable and appropriate for the analysis of balanced data, as the fixed effect estimates are then best linear unbiased estimates, and the variance component estimates are then minimum variance unbiased estimates under normality. That said, variance component estimates can be negative even for balanced data. This may be reasonable for analysis of fixed effects, but it seems problematic if one is interested in inferences

Table 19.24 R program and selected output illustrating analysis of simple contrasts—UAV experiment

```
> uav.data = read.table("data/uav.txt", header=T)
> uav.data = within(uav.data, {fA = factor(A); fW = factor(W);
+                               fB = factor(B); fC = factor(C) })
> head(uav.data, 3)
A W B C Time fC fB fW fA
1 1 1 1 1   29  1  1  1  1
2 1 1 1 2   28  2  1  1  1
3 1 1 1 3   49  3  1  1  1

> # Least squares ANOVA
> # Set contrast options for correct lsmeans and contrasts
> options(contrasts = c("contr.sum", "contr.poly"))
> model1 = aov(Time ~ fA*fB*fC + Error(fA:fW), data=uav.data)
> summary(model1)

> # Compare 2 levels of fA pairwise
> library(lsmeans)
> lsmA = lsmeans(model1, ~ fA)
> summary(contrast(lsmA, method="pairwise"), infer=c(T,T))

contrast estimate      SE df lower.CL upper.CL t.ratio p.value
1 - 2      20.062 0.76195 14   18.428   21.697   26.33  <.0001

Results are averaged over the levels of: fB, fC
Confidence level used: 0.95

> # Compare 2 levels of fA pairwise given each fB:fC combination
> library(lsmeans)
> lsmAgBC = lsmeans(model1, ~ fA | fC:fB)
> summary(contrast(lsmAgBC, method="pairwise", adjust="none"),
+         infer=c(T,T), level=0.99)

fC = 1, fB = 1:
contrast estimate    SE     df lower.CL upper.CL t.ratio p.value
1 - 2      14.375 1.8643 110.22   9.4885   19.262   7.711  <.0001

fC = 2, fB = 1:
contrast estimate    SE     df lower.CL upper.CL t.ratio p.value
1 - 2      13.375 1.8643 110.22   8.4885   18.262   7.174  <.0001
...
fC = 4, fB = 2:
contrast estimate    SE     df lower.CL upper.CL t.ratio p.value
1 - 2      23.750 1.8643 110.22  18.8635   28.637  12.740  <.0001

Confidence level used: 0.99
```

on variance components. Moreover, for unbalanced data, the fixed effects estimates can be inefficient, and the properties of the ANOVA-based variance component estimates are not well understood.

As an alternative to the analysis of variance approach, consider the use of maximum likelihood estimates, which generally have good large-sample properties. Under appropriate regularity conditions, the maximum likelihood estimator $\hat{\theta}$ of an estimable parameter θ is asymptotically $N(\theta, \text{CRLB})$, where CRLB is the Cramer–Rao lower bound for the variance of an unbiased estimator. Mind you, for the regularity conditions to hold, the estimate must be a solution to the likelihood equations—not obtained as a boundary condition—so these asymptotic properties do not apply for example to a variance component estimator when the estimate is constrained to be zero because the solution to the likelihood equations is negative. Likewise, given a factor with random effects but with few levels observed in an experiment, one should be skeptical of the asymptotic properties of the corresponding variance component estimate. That said, the asymptotic properties of MLEs should usually be reasonably applicable for the analysis of fixed effects unless an experiment is quite small. We will utilize restricted maximum likelihood estimation—a special case of maximum likelihood estimation.

The *restricted maximum likelihood approach* to fitting a mixed model involves two steps: (i) estimating the variance components by restricted maximum likelihood; then (ii) estimating the fixed effects by *estimated generalized least squares*—namely, treating the variance component estimates as true values, then computing the maximum likelihood estimates of the fixed effects, or equivalently, the *generalized least squares estimates*. To compute restricted maximum likelihood (ReML) estimates of the variance components, the original data is essentially replaced by a maximal set of linearly independent contrasts in the data each with mean zero (i.e. data contrasts the distributions of which do not involve the fixed effects), then the likelihood function of the contrasts is maximized. (Equivalently, one can estimate the fixed effects by ordinary least squares, then maximize the likelihood function of the residuals, the joint distribution of which does not involve the fixed effects. Hence, the restricted maximum likelihood approach is also known as *residual maximum likelihood*.) The ReML estimates of the variance components may be preferable to the usual maximum likelihood estimates, because the ReML estimates are the same as the ANOVA-based estimates if the data are balanced and all variance component estimates are positive. Also, restricted maximum likelihood adjusts in some sense for fixed effects, so often provides unbiased estimates of variance components. A new experiment will be introduced to illustrate the restricted maximum likelihood approach to analysis of split plot designs.

UAV Switch Experiment

Mahadevan (2009) conducted three experiments to evaluate the performance of a semi-automated computer display system designed to support a human operator's ability to monitor and control the complex dynamic operation of multiple unmanned aerial vehicles (UAVs) when the UAVs are involved in multiple combat-related tasks. His third experiment involved: 16 subjects (factor W); two alert techniques—a visual alert ($A = 1$) and an audio-visual alert ($A = 2$); and two levels of task complexity—a simple primary task coupled with a simple secondary task ($B = 1$), and a complex primary task coupled with a simple secondary task ($B = 2$). The experiment was a 2×2 split-plot design, with subjects as whole plots and two trials per subject as split-plots, with each level of A assigned to half of the subjects, and with both levels of B observed once on each subject. Hence, subject is nested within alert type. For each trial, each subject, while working on the primary task, was warned of the secondary task using one of the two alert techniques. One of the response variables measured was the time taken to switch to the secondary task following the alert, and the experimenter was interested in the effect of the nature of the alert (A) on the mean time to switch from the primary to the secondary task. The data are shown in Table 19.25. The model used by the experimenter is as follows.

Table 19.25 Time taken (in seconds) to switch to secondary task for the UAV switch experiment

A (Alert Type)	B (Complexity)	W (Subject)							
		1	2	3	4	5	6	7	8
1	1	6	5	5	5	5	7	5	6
	2	16	22	16	20	12	18	16	14
2	1	7	6	5	6	5	6	4	4
	2	6	7	6	8	6	6	7	6

Source Mahadevan (2009). Copyright © 2009 Sriram Mahadevan. Reprinted with permission

$$
\begin{aligned}
Y_{iuj} = \mu + \alpha_i + \epsilon_{iu}^W & \\
+ \beta_j + (\alpha\beta)_{ij} + \epsilon_{j(iu)}^S \,, &
\end{aligned}
\tag{19.9.21}
$$
$$
\epsilon_{iu}^W \sim N(0, \sigma_W^2)\,, \quad \epsilon_{j(iu)}^S \sim N(0, \sigma_S^2)\,,
$$
$$
\epsilon_{iu}^W \text{'s and } \epsilon_{j(iu)}^S \text{'s are all mutually independent,}
$$
$$
i = 1, 2; \quad u = 1, \dots, 8; \quad j = 1, 2.
$$

The R program and selected output for the UAV switch experiment is shown in Tables 19.26 and 19.27. Table 19.26 illustrates the analysis of variance approach, which is recommended for analysis of fixed effects given a balanced design—the case here. The model term `Error(fA:fW)` in the `aov` function models the whole plot errors ϵ_{iu}^W as random whole plots effects $W(A)$, and the corresponding mean square provides the denominator for testing main effects of A, as appropriate. R code for multiple comparisons is shown, but not the corresponding output.

For sake of comparison, the restricted maximum likelihood approach is illustrated in the top of Table 19.27. The `lmer` function fits a linear mixed effects model by restricted maximum likelihood. This function is part of the `lme4` package, but loading it via the `lmerTest` package adds the p-values to the analysis of variance table. For the model specified in the `lmer` statement, the term `(1|fA:fW)` causes inclusion of random $W(A)$ effects in the model, representing the whole plot errors ϵ_{iu}^W. The corresponding variance component estimate, bounded to be nonnegative, is essentially zero, ($\hat{\sigma}_W^2 = 3.88 \times 10^{-19}$ per output from the `summary` command). Consequently, the restricted maximum likelihood approach effectively removes this term from the model, pooling its degrees of freedom with that for (split plot) error, yielding 28 error degrees of freedom rather than 14 each for $W(A)$ and error as in the analysis of variance approach. This is confirmed by the analysis of variance provided at the bottom of Table 19.27, obtained by removing whole plot effects from the model. These two analyses provide exactly the same F tests for the fixed effects. They also generate the same multiple comparisons results (not shown), with both procedures using 28 error degrees of freedom for all contrast standard errors.

The analysis in Table 19.26 is correct and should be used. It correctly implements the planned statistical analyses using the proposed model. If the model is reasonable and appropriate, then the statistical results are valid, even if a simpler model with $\sigma_W^2 = 0$ suggested by the data would also be correct.

Table 19.26 R program and selected output for the UAV switch experiment: analysis of variance approach

```
> uav3.data = read.table("data/uav3.txt", header=T)
> uav3.data = within(uav3.data,
+                     {fA = factor(A); fW = factor(W); fB = factor(B) })
> head(uav3.data, 3)

  A W B Time fB fW fA
1 1 1 1    6  1  1  1
2 1 1 2   16  2  1  1
3 1 2 1    5  1  2  1

> # ANOVA: full model
> # Set contrast options for correct lsmeans and contrasts
> options(contrasts = c("contr.sum", "contr.poly"))
> model1 = aov(Time ~ fA + fB + fA:fB + Error(fA:fW), data=uav3.data)
> summary(model1)

Error: fA:fW
          Df Sum Sq Mean Sq F value  Pr(>F)
fA         1  215.3   215.3    69.8 8.2e-07
Residuals 14   43.2     3.1

Error: Within
          Df Sum Sq Mean Sq F value  Pr(>F)
fB         1  306.3   306.3      97 1.1e-07
fA:fB      1  205.0   205.0      65 1.3e-06
Residuals 14   44.2     3.2

> # Multiple comparisons
> library(lsmeans)
> lsmA.aov = lsmeans(model1, ~ fA)
> summary(contrast(lsmA.aov, method="pairwise"),
+          infer=c(T,T), side="two-sided")
> lsmAB.aov = lsmeans(model1, ~ fA:fB)
> summary(contrast(lsmAB.aov, method="pairwise", adjust="tukey"),
+          infer=c(T,T), side="two-sided")
```

It is an open problem whether the analyses provided in Table 19.27 are strictly correct—namely, whether they control error rates for any preplanned analysis, assuming the originally posed model (19.9.21) is correct. It seems *improper* for example if one were to fit the original full model by least squares, see that the estimate $\hat{\sigma}_W^2$ is negative, so change the model by removing the whole-plots term from the model, then fit the reduced model and use these results for the analysis. Control of error rates is an open problem when one uses the data to determine the model then uses the model to analyze the same data. Still, it is interesting that the restricted maximum likelihood approach fits and conducts the analysis under the assumption that model (19.9.21) is correct and, while it so happens that $\hat{\sigma}_W^2 = 0$, the originally postulated model is used for the analysis. If this approach could be shown to control error rates, then this would seem to be the preferred analysis, since effectively setting $\hat{\sigma}_W^2 = 0$

Table 19.27 R program and selected output for the UAV switch experiment, continued: restricted maximum likelihood approach, and reduced ANOVA model

```
> # ReML: full model
> library(lmerTest)
> model2 = lmer(Time ~ fA + fB + fA:fB + (1|fA:fW), data=uav3.data)
> anova(model2)

Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
      Sum Sq Mean Sq NumDF DenDF F.value  Pr(>F)
fA       215     215     1    28    69.0 4.9e-09
fB       306     306     1    28    98.2 1.2e-10
fA:fB    205     205     1    28    65.7 8.0e-09

> summary(model2)

...
Random effects:
 Groups   Name        Variance Std.Dev.
 fA:fW    (Intercept) 3.88e-19 6.23e-10
 Residual             3.12e+00 1.77e+00
...

> # Multiple comparisons
> # Detach then reload lsmeans, to avoid issues from its masking by lmerTest
> detach("package:lsmeans", unload=TRUE)
> library(lsmeans)
> lsmA2 = lsmeans(model2, ~ fA)
> summary(contrast(lsmA2, method="pairwise"),
+         infer=c(T,T), side="two-sided")
> lsmAB.reml = lsmeans(model2, ~ fA:fB)
> summary(contrast(lsmAB.reml, method="pairwise", adjust="tukey"),
+         infer=c(T,T), side="two-sided")

> # ANOVA: reduced model--without fW
> model3 = aov(Time ~ fA + fB + fA:fB, data=uav3.data)
> summary(model3)

            Df Sum Sq Mean Sq F value  Pr(>F)
fA           1  215.3   215.3    69.0 4.9e-09
fB           1  306.3   306.3    98.2 1.2e-10
fA:fB        1  205.0   205.0    65.7 8.0e-09
Residuals   28   87.4     3.1

> # Multiple comparisons
> lsmA.aov2 = lsmeans(model3, ~ fA)
> summary(contrast(lsmA.aov2, method="pairwise"),
+         infer=c(T,T), side="two-sided")
> lsmAB.aov2 = lsmeans(model3, ~ fA:fB)
> summary(contrast(lsmAB.aov2, method="pairwise", adjust="tukey"),
+         infer=c(T,T), side="two-sided")
```

causes pooling of $ms(W(A))$ and msE (i.e. msE_W and msE_S) to estimate σ_S^2 with more degrees of freedom. One benefit should be increased power for some tests—especially tests of effects compared between whole plots. It would also provide a common variance estimator for multiple comparisons of AB-treatment combinations, making most if not all methods of multiple comparisons applicable. Moreover, the restricted maximum likelihood approach becomes preferable to the analysis of variance approach for unbalanced designs.

19.9.4   Recovery of Inter-block Information

The oats experiment, introduced in Sect. 19.3.4, involves a split-plot design with complete blocks at both the whole- and split-plots levels—namely, the levels of A assigned to whole plots within blocks constitute a randomized complete block design, and the levels of B assigned to split plots within whole plots constitute a randomized complete block design. With this dual complete block structure, the data analysis is the same whether the block effects are modeled as fixed (as in Sect. 19.3.4) or random (as in Sect. 19.9.4). Such would not be the case if for example the split-plot design involves incomplete blocks at the whole-plots level, as illustrated in this section.

Consider again the oats experiment and corresponding data in Table 19.3 (p. 710), but suppose one only has the data for levels 1 and 2 of A in blocks 3 and 5, for levels 0 and 2 of A in blocks 1 and 4, and for levels 0 and 1 of A in blocks 2 and 6. For this subset of the data, the levels of A assigned to whole plots in blocks constitute a balanced incomplete block design with blocks of size 2. The R program and output in Tables 19.28 and 19.29 provides two approaches to the analysis of these data.

In Table 19.28, the analysis of variance approach is used and block effects are modeled as fixed. This is a Type I analysis. Since block effects are modeled as fixed, unbiased estimates of main-effect-of-A contrasts must be *intra-block estimates*—namely, each is composed (by summing over blocks) of within-block contrasts of observations—as is necessary for the fixed block effects to cancel out to yield unbiased estimates. Analogously, for testing for main effects of A, the numerator of the F-statistic is the mean square for A adjusted for block effects, which can be computed from the sum of squares of an appropriate set of such intra-block estimates of main-effect-of-A contrasts. The corresponding data analysis is called the *intra-block analysis*. For sake of comparison with subsequent analyses, note that the pairwise comparisons with the control for A each have estimated standard error 11.288 with 4 degrees of freedom. The same results would be obtained by the restricted maximum likelihood approach using the following R code, except Type III tests would be provided, so the sum of squares for blocks would be adjusted for A.

```
library(lmerTest)
model2 = lmer(y ~ fBlock + fA + fB + fA:fB + (1|fWP:fBlock),
              data=oats2.data)
anova(model2)
```

The program and output are continued in Table 19.29, where the restricted maximum likelihood approach is used, but this time block effects are modeled as random. With random block effects, one can obtain unbiased estimates of main-effect-of-A contrasts using contrasts of block totals (the sum total of the observations in each block), and these so-called *inter-block estimates* are in addition to and independent of the intra-block estimates used in the first analysis. So, for any main-effect-of-A contrast, any fixed weighted average of the corresponding intra- and inter-block estimates provides an unbiased estimate of the contrast, and the best (minimum variance) estimate would be obtained when each weight is inversely proportional to the variance of the corresponding estimate. Unfortunately, this best weighting is unknown because, for any main-effect-of-A contrast, the variances of the intra- and inter-block estimates are different and depend on the unknown variance components. A common

Table 19.28 R program and output illustrating the intra-block analysis—the oats split-plot experiment

```
> oats.data = read.table("data/oats.txt", header=T)
> oats.data = within(oats.data, {fBlock = factor(Block);
+              fWP = factor(WP); fA = factor(A); fB = factor(B) })
> # Reduce data so level of A constitute a BIBD
> oats2.data = subset(oats.data,
+                    !(fA == 0 & (fBlock == 3 | fBlock == 5))
+                    & !(fA == 1 & (fBlock == 1 | fBlock == 4))
+                    & !(fA == 2 & (fBlock == 2 | fBlock == 6)) )
>
> # Intra-block analysis: ANOVA, fixed block effects
> options(contrasts = c("contr.sum", "contr.poly")) # For correct lsmeans
> model1 = aov(y ~ fBlock + fA + fB + fA:fB + Error(fWP:fBlock),
+              data=oats2.data)
> summary(model1)

Error: fWP:fBlock
          Df Sum Sq Mean Sq F value Pr(>F)
fBlock     5  12064    2413    3.16   0.14
fA         2    377     188    0.25   0.79
Residuals  4   3058     765

Error: Within
          Df Sum Sq Mean Sq F value  Pr(>F)
fB         3  15966    5322   34.23 2.4e-09
fA:fB      6   1361     227    1.46    0.23
Residuals 27   4198     155

> # Dunnett's Method for A: ANOVA approach
> library(lsmeans)
> lsmA1 = lsmeans(model1, ~ fA)
> set.seed(19831957)
> summary(contrast(lsmA1, method="trt.vs.ctrl", adjust="mvt"),
+         infer=c(T,T), level=0.99)

 contrast estimate     SE df lower.CL upper.CL t.ratio p.value
 1 - 0      6.0417 11.288  4  -54.508   66.592   0.535  0.8228
 2 - 0      7.4583 11.288  4  -53.092   68.008   0.661  0.7503

Confidence level used: 0.99
Conf-level adjustment: mvt method for 2 estimates
P value adjustment: mvt method for 2 tests
```

approach is to use the estimated variances to determine the weights, but random weights would make the resulting analyses approximate. If the weights are well chosen, the resulting combined estimate is better (has smaller variance) than the intra-block estimate, due to recovery of inter-block information.

Consider now the output in Table 19.29, obtained by recovering and using this inter-block information. For factor *A*, the *F*-test and corresponding *p* values have changed, as have the results of Dunnett's method comparing levels of *A* to the control level. Recovery of the inter-block information ought to provide contrast estimates that are more accurate, and indeed the standard errors for the treatment-versus-control comparisons are smaller than in the inter-block analysis. Also, the variance estimate for each main-effect-of-*A* contrast is essentially a composite variance estimate, so Satterthwaite's method is used to compute the number of denominator degrees of freedom (i.e. 4.95) for the *F* test of *A*,

Table 19.29 R program and output illustrating the recovery of inter-block information—the oats split-plot experiment

```
> # Recovering inter-block information: ReML, random block effects
> model3 = lmer(y ~ fA + fB + fA:fB + (1|fBlock) + (1|fWP:fBlock),
+               data=oats2.data)
> anova(model3)

Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
      Sum Sq Mean Sq NumDF DenDF F.value  Pr(>F)
fA       150      75     2  4.95     0.5    0.64
fB     15966    5322     3 27.00    34.2 2.4e-09
fA:fB   1362     227     6 27.00     1.5    0.23

> # Dunnett's Method for A
> lsmA3 = lsmeans(model3, ~ fA)
> set.seed(19831957)
> summary(contrast(lsmA3, method="trt.vs.ctrl", adjust="mvt"),
+         infer=c(T,T), level=0.99)

 contrast estimate    SE   df lower.CL upper.CL t.ratio p.value
 1 - 0       3.516 10.71 4.95  -46.042   53.074   0.328  0.9257
 2 - 0      10.350 10.71 4.95  -39.208   59.908   0.966  0.5638
```

which is also equal to the number of degrees of freedom associated with the standard error for each main effect of *A* contrast. For these pairwise comparisons, each standard error is smaller with more degrees of freedom than for the intra-block analysis, so there is a helpful recovery of the inter-block information.

The restricted maximum likelihood approach was used above to recover the inter-block information for main-effect-of-*A* contrasts. One may not be able to accomplish this directly in R using the analysis of variance approach. Consider the following code, for example.

```
model4 = aov(y ~ fA + fB + fA:fB + Error(fBlock + fWP:fBlock),
             data=oats2.data)
summary(model4)
```

The `Error` function, used to enter random effects in the model, causes separate analyses to be conducted for each error strata. In this example, the error strata correspond to: comparisons within whole plots, that involve split-plot error; comparisons between whole plots within blocks, that involve split-plot and whole-plot error; and comparisons between blocks, that involve split-plot error, whole-plot error, and variation due to random block effects. The intra-block and inter-block estimates of main-effect-of-*A* contrasts involve two different error strata, so the above code using the analysis of variance approach generates two *F* tests for *A*—the intra-block test in Table 19.28 and an inter-block test (not shown)—but no composite test.

Turning attention to the split-plot comparisons, note that the *F*-tests for *B*, and likewise for *AB*, are the same in each analysis. This is the case because a randomized complete block design was used for the assignment of levels of *B* to split plots in whole plots (serving as blocks). Consequently, for each main effect of *B* and *AB*-interaction contrast, the corresponding contrast in treatment means is a within-whole-plot contrast, free of whole-plot or block effects, so no adjustment is needed.

In the above example, recovery of the inter-block information provided better information for analysis of the whole-plots factor *A*. For example, for each treatment-versus-control contrast estimate for *A*, recovery of inter-block information provided a tighter confidence interval, due to having both

a smaller standard error and more error degrees of freedom. In planning such an experiment, the experimenter should determine in advance how the data will be analyzed. While one could plan on conducting the intra-block analysis, recovery of the inter-block information should usually be beneficial.

19.9.5 ReML and Unbalanced Data

In this section, we illustrate the restricted maximum likelihood approach using the linear mixed effects function `lmer` for analysis of an unbalance split-plot design, revisiting the mobile computing field study experiment introduced in Sect. 19.7. Recall, while the planned experiment was nicely balanced, it was discovered prior to analysis of the data that one of the subjects traversed one of the paths in the reverse direction, making the corresponding observation inappropriate to use in the analysis. As such, one of the 72 planned observations (given in Table 19.11, p. 719) is listed as missing. In Sects. 19.7.1 and 19.7.2, an approximate analysis of the split-plot design was provided by estimating the missing value and treating it as observed, using standard formulas for analysis of the balanced design via the analysis of variance approach, but adjusting degrees of freedom appropriately. With one observation lost, the data are no longer balanced, making the data analysis conceptually and computationally more difficult. However, the R function `lmer` handles this situation nicely, as we shall illustrate.

The R program and output for analysis of the 71-observation split-plot design via the restricted maximum likelihood approach is shown in Tables 19.30 and 19.31. The mixed model for the analysis, provided in Eq. (19.7.18), p. 718, was fit in Table 19.30 using the `lmer` function of the `lme4` (and `lmerTest`) package. The three variance components are estimated by restricted maximum likelihood. In this case, the three estimates are all positive, as shown in selected output from the `summary` command. Given these variance component estimates, the fixed effects are then estimated by generalized least squares.

The `anova` function then provides Type III F-tests for each fixed effect—Type III in the sense that the tests are based on estimates of effects and corresponding variability obtained by fitting the full model. By default, Satterthwaite's approximation is used to compute the denominator degrees of freedom for each F test, analogous to Satterthwaite's approximation for a balanced design. Regarding the two treatment factors, only the effects of visual presentation format (B) are significant, with $p = 0.01660$. Concerning the nuisance factors path and order, these involved both within- and between-whole-plot comparisons before the lost observation caused design imbalance. For each of these factors, tests are provided for each of three pseudofactor components, including for example tests for P_2, P_3 and $P_2 P_3$ for path. At the top of Table 19.31, a second call of the `lmer` and `anova` functions fits a model using factors rather than pseudo-factors for order and path, providing a consolidated F-test for each of these nuisance factors. It is not surprising that path has significant effects, and these are attributable to the three distinct paths used as distinguished by P_3.

At the bottom of Table 19.31, the `lsmeans` function has generated the pairwise comparisons for each of factors A and B, including individual t-tests and individual 95% confidence intervals for each pairwise comparison. These results are similar to those obtained in Sect. 19.7.2, where the approach used was to estimate the missing value and analyze the balanced design, though there Tukey's method was used. Based on the individual 95% confidence intervals provided here, level 2 of B—the birds' eye view egocentric visual presentation format—has a significantly smaller $RMSE$ from the intended path than do the other two visual presentation formats. It is interesting that the denominator degrees of freedom is not constant for the factor B pairwise comparisons, due to the imbalance created by the lost observation. This indicates that the pairwise comparisons do not all utilize the same variance estimator, though presumably nearly so. As a consequence, the Bonferroni method should perhaps be used if multiple comparisons for B are of interest.

Table 19.30 R program and output for the mobile computing field study experiment

```
> MCFS71.data = read.table("data/MCFS71.txt", header=T)
> head(MCFS71.data, 3)

  Subj Day Order O2 O3 Path P2 P3 A B      y
1   1   1     1  0  0    1  0  0 1 1 49.321
2   1   1     2  0  1    2  0  1 1 2 24.386
3   1   1     3  0  2    3  0  2 1 3 37.680

> MCFS71.data = within(MCFS71.data, {fSubj = factor(Subj);
+         fDay = factor(Day); fOrder = factor(Order); fO2 = factor(O2);
+         fO3 = factor(O3); fPath = factor(Path); fP2 = factor(P2);
+         fP3 = factor(P3); fA = factor(A); fB = factor(B) })

> # ReML
> library(lmerTest)
> model1 = lmer(y ~ fO2 + fP2 + fA + fO3 + fO2:fO3 + fP3 + fP2:fP3
+                  + fB + fA:fB + (1|fSubj) + (1|fSubj:fDay),
+                  data=MCFS71.data)
> summary(model1)

...
Random effects:
 Groups      Name        Variance Std.Dev.
 fSubj:fDay (Intercept) 28.2     5.31
 fSubj      (Intercept) 64.0     8.00
 Residual               93.0     9.64
...

> anova(model1)

Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
        Sum Sq Mean Sq NumDF DenDF F.value  Pr(>F)
fO2         23      23     1   8.5    0.25 0.63172
fP2        174     174     1   8.5    1.87 0.20643
fA          17      17     1   8.5    0.19 0.67688
fO3          8       4     2  34.6    0.04 0.95807
fP3       1976     988     2  34.6   10.63 0.00025
fB         860     430     2  34.6    4.62 0.01660
fO2:fO3     69      35     2  34.6    0.37 0.69270
fP2:fP3     26      13     2  34.6    0.14 0.87176
fA:fB      146      73     2  34.6    0.79 0.46336
```

One could also generate pairwise comparisons for the *AB* combinations by using A:B as the effect in an lsmeans statement, as follows.

Table 19.31 R program and output, continued, for the mobile computing field study experiment

```
> # Consolidating tests for Order and Path
> model2 = lmer(y ~ fOrder + fPath + fA + fB + fA:fB + (1|fSubj)
+                         + (1|fSubj:fDay), data=MCFS71.data)
> anova(model2)

Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
       Sum Sq Mean Sq NumDF DenDF F.value Pr(>F)
fOrder     98      20     5  31.1    0.21 0.9555
fPath    2184     437     5  31.1    4.70 0.0026
fA         17      17     1   8.5    0.19 0.6769
fB        860     430     2  34.6    4.62 0.0166
fA:fB     146      73     2  34.6    0.79 0.4634


> # Multiple comparisons
> library(lsmeans)
> lsmA = lsmeans(model1, ~ fA)
> summary(contrast(lsmA, method="pairwise", adjust="none"),
+         infer=c(T,T), level=0.95)

 contrast estimate     SE   df lower.CL upper.CL t.ratio p.value
 1 - 2     -1.3631 3.1606 8.98  -8.5149   5.7887  -0.431  0.6764

> lsmB = lsmeans(model1, ~ fB)
> summary(contrast(lsmB, method="pairwise", adjust="none"),
+         infer=c(T,T), level=0.95)

 contrast estimate     SE    df lower.CL upper.CL t.ratio p.value
 1 - 2      8.1957 2.8360 35.39   2.4406 13.95082   2.890  0.0065
 1 - 3      1.6903 2.7837 35.02  -3.9608  7.34143   0.607  0.5476
 2 - 3     -6.5054 2.8360 35.39 -12.2605 -0.75034  -2.294  0.0278
```

```
lsmAB = lsmeans(model1, ~ fA:fB)
summary(contrast(lsmAB, method="pairwise", adjust="bon"),
        infer=c(T,T), level=0.95)
```

However, only the Bonferroni method of multiple comparisons would be applicable, since these comparisons involve a mix of with-and between-day comparisons, even before an observation was lost, so there would not be a common variance estimator for all of these pairwise comparisons. Also, the corresponding variance estimators would generally be composite variance estimators, in which case Satterthwaite's approximation would be used.

Exercises

1. **Drug experiment**

   An experiment designed as a split-plot design was described by W.M. Wooding in the *Journal of Quality Technology* in 1973. The experiment concerned the evaluation of eight drugs (factor *A* at $a = 8$ levels) for the treatment of arthritis. A second factor was the dose of the drug (factor *B* at $b = 2$ levels), and the third factor was the length of time (factor *C* at $c = 2$ levels) that a measurement was taken after injection by a substance known to cause an inflammatory reaction. The

Table 19.32 Fluid (milliliters) in pleural cavity for the drug experiment

	Block I									
Whole	Dose	Time				Drug A				
Plot	B	C	1	2	3	4	5	6	7	8
1	1	1	5.7	8.6	6.9	6.6	6.7	7.4	5.7	6.7
2	1	2	8.4	9.6	9.3	11.1	12.5	8.7	9.3	9.5
3	2	1	5.1	7.2	6.8	6.4	6.6	8.7	6.7	7.0
4	2	2	7.3	8.7	7.9	6.9	8.9	9.5	8.3	11.3
	Block II									
Whole	Dose	Time				Drug A				
Plot	B	C	1	2	3	4	5	6	7	8
5	1	1	5.8	6.8	7.0	8.5	7.8	7.3	6.4	8.5
6	1	2	9.1	10.8	6.9	12.2	9.9	10.4	10.6	10.5
7	2	1	5.4	7.9	8.0	6.4	8.4	7.1	6.4	7.2
8	2	2	5.3	10.4	8.2	8.1	10.9	9.8	8.4	14.6

Source Wooding (1973). Reprinted with Permission from Journal of Quality Technology © 1973 ASQ, www.asq.org

experimental units used in the study were $n = 64$ rats. The response was the amount of fluid (in milliliters) measured in the pleural cavity of an animal after having been administered a particular treatment combination.

In many pharmacological studies, time of day has an effect on the response due to changing laboratory conditions, etc. Consequently, the experiment was divided into blocks, whole plots and split plots. The blocks were of size 32, each set of 32 observations being measured on a single day. Each treatment combination was measured once per day. Each day was then subdivided into 4 whole plots of size 8, where the eight measurements within a whole plot were taken fairly close together in time.

Since the effect of the drug (A) was of primary importance, and since the effects of B and C were of interest only in the form of an interaction with A, the main effects of B and C and the BC interaction were confounded with the whole plots. The data are shown in Table 19.32, and the experimenter used the logarithms of the data in his analysis. Notice that the design for A on the split plots is a randomized block design, and the design for the BC combinations on the whole plots is also a randomized block design.

(a) Write out a model for this experiment.
(b) Calculate an analysis of variance table using the logarithms of the data. Distinguish between the effects measured on the whole plots and those measured on the split plots. Identify the whole-plot error and split-plot error.
(c) Test any hypotheses of interest and state your conclusions clearly.
(d) Examine interaction plots of any important interactions. Calculate a set of 95% confidence intervals for the differences between pairs of drugs. State your conclusions.

2. **Fishing line experiment**

The fishing line experiment was run by C. Reynolds, B. Grunden, and K. Taylor in 1996 in order to compare the strengths of two brands of fishing line exposed to two different levels of stress. Two different reels of fishing line were purchased for each of the two brands, and sections of line were cut from each reel. Thus the reels were automatically assigned to the levels of factor A (Brand), and constituted the four whole plots. There were no blocks in this experiment. The split plots constituted sixteen sections of line, four cut from each of the four reels (that is, 16 split plots in total, 4 per whole plot). The split plots were randomly assigned to two different stress levels (stressed, "S";

Table 19.33   Strength of line for the fishing line experiment

Whole plot (Reel)	A (Brand)	Level of B (weight)			
1	1	N (6.70)	S (6.40)	S (7.20)	N (7.00)
2	2	S (8.10)	S (8.90)	N (8.00)	N (6.10)
3	2	S (8.00)	S (8.00)	N (8.75)	N (8.50)
4	1	N (8.50)	S (9.50)	N (9.70)	S (9.40)

nonstressed, "N") so that each stress level was assigned two split plots per whole plot.

The stress was induced by hanging a brick from the assigned section of line for 14 hours. Although this did not precisely mimic the stress induced during fishing, it was still expected to give some information about the strength of the lines under stress.

The strength test was accomplished by hanging a bucket on the end of the line, which was suspended from a beam. The bucket was gradually filled with water through a small hole in the lid until the line broke. The data are the resulting weights of water to the nearest 0.01 lb and are shown in Table 19.33.

(a) Write down a model for this experiment.
(b) Construct an analysis of variance table and test the hypotheses that you think are of interest. State your conclusions.
(c) If you were to repeat this experiment, suggest ways in which you would try to improve it.

3. **Cigarette experiment**

The cigarette experiment was run by J. Edwards, H. Hwang, S. Jamison, J. Kindelberger, and J. Steinbugl in 1996 in order to determine the factors that affect the length of time that a cigarette will burn. There were three factors of interest:

- "Tar" (factor A) at two levels, "regular" and "ultra-light,"
- "Brand" (factor B) at two levels, "name brand" and "generic brand" (coded 1 and 2),
- "Age" (factor C) at three levels, "fresh," "24 hour air exposure," "48 hour air exposure."

The cigarettes were to be burned in whole plots of size six. This was to help with the difficulty of recording burning times and to help keep the amount of smoke in the room at a reasonable level.

Table 19.34   Burning times for the cigarette experiment

Whole plot (Time)	A (Tar)	Levels of BC (Burning times in seconds)					
1	1	22 (301)	11 (326)	23 (260)	13 (290)	12 (312)	21 (292)
2	2	11 (329)	12 (331)	13 (285)	21 (306)	22 (258)	23 (276)
3	2	22 (290)	11 (380)	12 (335)	13 (309)	23 (243)	21 (334)
4	2	11 (321)	21 (337)	23 (275)	12 (316)	13 (307)	22 (250)
5	2	22 (308)	11 (345)	21 (307)	23 (288)	13 (321)	12 (330)
6	1	11 (344)	23 (283)	21 (281)	22 (261)	13 (307)	12 (292)
7	1	21 (274)	13 (310)	12 (304)	22 (279)	23 (277)	11 (330)
8	1	13 (302)	12 (325)	22 (301)	11 (338)	23 (270)	21 (297)
9	2	12 (323)	13 (334)	23 (265)	11 (326)	22 (269)	21 (297)
10	1	23 (309)	13 (314)	22 (259)	11 (344)	21 (310)	12 (322)

There were ten whole plots, and these were assigned at random to the tar levels so that each tar level was assigned five whole plots.

The six split plots (time slots) in each whole plot were assigned at random to the six brand/age treatment combinations. Marks were made across the seam of each cigarette at a given distance apart. Each cigarette was lit at the beginning of its allotted time slot, and the time taken to burn between the two marks was recorded. The data are shown in Table 19.34.

(a) Write down a model for this experiment.
(b) Construct an analysis of variance table and test the hypotheses that you think are of interest. State your conclusions.
(c) Use Tukey's method to examine the pairwise differences in the effects on burning time of the six BC treatment combinations (averaged over tar levels).
(d) Examine the linear and quadratic trends of burning time due to different ages, for each brand separately.
(e) State your conclusions about the experiment, including your choice of overall significance levels and overall confidence levels.

4. **Injection molding experiment, continued**

The injection molding experiment, introduced in Exercise 10 of Chap. 15, was run in order to examine the effect of six factors on the shrinkage of a part produced by injection molding. The six factors were injection velocity (factor A), cooling time (factor B), barrel zone temperature (factor C), mold temperature (factor D), hold pressure (factor E), and back pressure (factor F). Each factor had two levels coded 0 and 1. The treatment combinations, which are shown in Table 15.56, p. 555, were not completely randomized. The levels of factor D were time-consuming to change, so for the first four observations D was held at its low level and the combinations of the other factors were randomly ordered. For the last four observations, D was held at its high level and the combinations of the other factors were randomly ordered. Thus we can think of this design as having two whole plots assigned to the two levels of D, and having four split plots nested within each whole plot assigned at random to combinations of levels of A, B, C, E, and F. Table 19.35 contains the corresponding split plot design and mean lengths. Let σ_W^2 and σ_S^2 denote the variances of the random whole-plot and split-plot effects, respectively.

(a) Which fixed effects can be estimated? Compute their estimates.
(b) Which variance components can be estimated? Compute their estimates.
(c) Is it possible to analyze this experiment in a sensible way? If so, present your results.

Table 19.35 Mean lengths for the injection molding experiment

Whole plot	Split plot	A	B	C	D	E	F	Mean length
1	1	0	0	0	0	0	0	2
	2	0	1	1	0	0	1	46
	3	1	0	1	0	1	0	108
	4	1	1	0	0	1	1	128
2	1	0	0	0	1	1	1	73
	2	0	1	1	1	1	0	105
	3	1	0	1	1	0	1	53
	4	1	1	0	1	0	0	58

Table 19.36 Split-plot design and times (in seconds) for the UAV experiment: situation awareness comprehension

B (Similarity)		1				2			
C (Complexity)		1	2	3	4	1	2	3	4
A (Cue)	W (Subject)								
1	1	20	21	27	32	23	21	30	28
	2	24	19	33	35	24	23	32	30
	3	19	22	28	33	21	20	34	31
	4	30	27	37	30	26	24	31	33
	5	25	24	35	38	28	27	33	30
	6	22	20	40	41	26	25	29	37
	7	28	23	34	36	28	26	37	34
	8	21	26	33	34	24	28	34	40
2	1	8	9	11	9	7	7	8	8
	2	10	8	13	13	12	8	18	10
	3	11	9	15	14	9	10	15	18
	4	8	9	12	18	8	12	8	9
	5	7	10	17	13	12	12	12	17
	6	9	7	13	12	10	7	12	10
	7	11	12	10	11	13	8	10	8
	8	10	11	15	13	7	13	9	8

Source Mahadevan (2009). Copyright © 2009 Sriram Mahadevan. Reprinted with permission

5. **UAV experiment, continued**

The UAV experiment, introduced in Sect. 19.6, was run using a $2^2 \times 4$ split-plot design with 16 subjects (W) as whole plots, with two levels of cue condition (A) assigned to whole-plots, and with 2×4 combinations of levels of task similarity (B) and task complexity (C) assigned to split plots. Another response variable measured by the experimenter was the time taken to perform situation awareness comprehension tasks in the primary task, yielding the data shown in Table 19.36. Using model (19.6.16), p. 715, conduct the following analyses.

(a) Construct an analysis of variance table. For each main effect and interaction in the treatment factors, determining which effects are significant at the 1% level.
(b) Construct a 95% confidence interval for the main effect of cue, and interpret the results.
(c) If the three-factor interaction is significant, then compare the effect of cue at each combination of the other two factors. Otherwise, if the factor cue interacts significantly with either of the other treatment factors, then compare the effect of cue at each level of each factor with which cue interacts significantly. Use individual 99% confidence intervals.

6. **UAV switch experiment, continued**

The UAV switch experiment, introduced in Sect. 19.8.3, was run using a 2×2 split-plot design with 16 subjects as whole plots, with two levels of alert type (A) assigned to whole-plots, and with two levels of complexity (B) assigned to split plots. Main effects and interactions for A and B were all significant, but the primary interest is in comparing the effects of the levels of A. Use the approach analogous to the first call of PROC GLM in Table 19.19 to do the following.

(a) Construct a 95% confidence interval for the main effect of alert type, and interpret the results.
(b) For each level of complexity, estimate the difference in effects of the two levels of alert type. Determine the variance of each of the two corresponding pairwise comparison estimators,

Table 19.37 Alert detection times (in seconds) for UAV switch experiment

A (Alert type)	B (Complexity)	W (Subject)							
		1	2	3	4	5	6	7	8
1	1	5	6	4	6	5	5	5	6
	2	10	10	6	8	6	6	6	6
2	1	6	5	4	5	5	4	6	6
	2	5	6	4	5	5	4	6	5

Source Mahadevan (2009). Copyright © 2009 Sriram Mahadevan. Reprinted with permission

compute the corresponding estimated standard errors, and use Satterthwaite's approximation to compute the approximate number of degrees of freedom associated with each estimated standard error.

(c) Using the results of part (b), construct individual 99% confidence intervals for the difference in effects of the two levels of alert type for each level of complexity. Interpret the results.

7. **UAV switch experiment, continued**

The UAV switch experiment, introduced in Sect. 19.8.3, was run using a 2×2 split-plot design with 16 subjects as whole plots, with two levels of alert type (A) assigned to whole-plots, and with two levels of complexity (B) assigned to split plots. Another response variable measured by the experimenter was the alert detection time (in seconds), yielding the data shown in Table 19.37. Using model (19.8.20), p. 734, conduct the following analyses.

(a) Construct an analysis of variance table. For each main effect and interaction in the treatment factors, determining which effects are significant at the 1% level.
(b) Construct a 95% confidence interval for the main effect of alert type, and interpret the results.
(c) For each level of complexity, compare the effects of the two levels of alert type, using individual 99% confidence intervals.

8. **Mobile Computing Field Study, continued**

The Mobile Computing Field Study, introduced in Sect. 19.7, was run using a 2×3 split-plot design with 12 subjects as blocks, 24 days as whole plots, three runs per day as split plots, two display types (A) assigned to whole-plots, three visual presentation formats (B) assigned to split-plots, six runs per subject, and utilizing six paths. Another response variable measured by the experimenter was the time (in minutes) to navigate the path, yielding the data shown in Table 19.38. Using model (19.7.18), p. 718, conduct the following analyses, estimating the missing value and utilizing it in the analysis.

(a) Estimate the missing value.
(b) Construct an analysis of variance table analogous to Table 19.12. For each main effect and interaction in the treatment factors, determining which effects are significant at the 1% level.
(c) Compare the visual presentation formats using simultaneous 99% confidence intervals and Tukey's method. Interpret the results.

Table 19.38 Mobile computing field study: time (in minutes) for each subject (Subj), day, run order, and path-treatment combination (*PAB*), with one observation missing

Subj	Day 1						Day 2					
	Run 1		Run 2		Run 3		Run 4		Run 5		Run 6	
	PAB	Time	*PAB*	Time	*PAB*	Time	*PAB*	Time	*PAB*	Time	*PAB*	Time
1	111	15.633	212	11.717	313	13.550	421	11.900	522	11.817	623	16.517
2	213	13.767	311	13.867	112	11.233	523	10.533	621	11.966	422	10.633
3	312	11.150	113	13.167	211	12.633	622	9.367	423	10.350	521	9.750
4	121	11.867	222	8.600	323	8.700	411	10.400	512	9.117	613	10.067
5	223	10.183	321	11.600	122	8.350	513	16.667	611	9.783	412	7.600
6	322	7.650	123	9.667	221	12.417	612	11.400	413	14.900	511	10.350
7	421	9.233	522	8.700	623	12.083	111	10.700	212	7.833	313	9.050
8	523	11.130	621	8.717	422	7.967	213	8.867	311	8.917	112	6.667
9	622	12.133	423	12.800	521	12.433	312	11.767	113	12.367	211	12.983
10	411	18.150	512	10.483	613	13.370	121	14.100	222	10.217	323	11.583
11	513	10.883	611	15.117	412	10.833	223	16.783	321	15.333	122	12.483
12	612	9.567	413	16.450	511	11.217	322	.	123	10.567	221	10.517

Source Wesler (2001). Copyright © 2001 Mary Mc. Wesler. Research was performed under U.S. Army Research Laboratory, Federated Laboratory Research Consortium (DAAL01-96-0003) directed by Mr. Bernie Corona. Reprinted with permission

9. **Mobile Computing Field Study, continued**

The Mobile Computing Field Study was introduced in Sect. 19.7, and data on an additional response variable, time, was provided in Exercise 8. Using model (19.7.18), p. 718, conduct the following analyses of the response variable time, using only the 71 observations available—namely, without estimating and using the missing value. If using SAS software, you may adapt the program in Table 19.21; if using R, you may adapt the program in Tables 19.30 and 19.31.

(a) Provide the resulting variance component estimates.
(b) Test for significance of each treatment factor main effect and interaction. Report the observed significance level of each test, and interpret the results.
(c) Using an appropriate method of multiple comparisons, construct simultaneous 99% confidence intervals for pairwise comparison of the visual presentation formats (*B*). Interpret the results.

Computer Experiments

20

20.1 Introduction

All of the experiments described in the previous chapters are *physical experiments* where the experimenter can work directly with the treatments and experimental material. The experimenter makes an assignment of experimental units to the levels of the treatment factors, possibly including blocks, and then measures the corresponding responses. Since it is not possible to control every single variable that influences the scientific process under consideration, all measurements necessarily include random variability. Statisticians have developed numerous methods to mitigate the effects of uncontrolled variation on the study conclusions. For example, Chap. 1 introduced replication, blocking, and randomization. These standard techniques of physical experiments allow an experimenter to increase precision and decrease bias.

In some situations, however, it is economically, ethically, or temporally not possible to run a statistically appropriate physical experiment. Instead, the following scenario might be feasible. Suppose that

  (i) the physical process can be described by a mathematical model (for example, a system of differential equations),
 (ii) computer code (called a *simulator*) can be written to compute the response from the mathematical model, and
(iii) computational methods exist for working with this model (for example, solving differential equations) within a reasonable timeframe.

In this scenario, a researcher can conduct a *computer experiment* by running the computer code, which serves as a proxy for the physical process, to compute a "response" at any combination of values of the treatment factors, which are now called *input combinations*, or *input points*, or sometimes shortened to *inputs*. We use x to represent an input combination.

Example 20.1.1 A Real Experiment—Prosthetic Elbow Experiment

Many biomedical engineers use computer experiments to help with the engineering design of prosthetic devices. For example, Hayeck (2009) studied the effects of four implant position variables on the functioning of a total elbow replacement prosthetic device. One of the responses of interest was a measure of principal compressive strain in the bone.

The input variables were the tip displacement (x_1), the rotation of the implant axis about the lateral axis at the tip (x_2), the rotation of the implant axis about the anterior axis at the tip (x_3), and the rotation about the implant axis (x_4). The variable x_1 was measured in millimeters, while x_2, x_3, and x_4 were measured in degrees. To conduct an experiment, one would need to observe the response y at various combinations $\boldsymbol{x} = (x_1, \ldots, x_4)$ of the input variables. It would be ethically questionable to carry out this experiment on patients, since some input combinations of interest may prove harmful. However, as described by Hayeck (2009), the influence of these four variables on the functioning of the prosthetic device could be described by a mathematical model and could be implemented in computer code, and so a computer experiment could be conducted. □

A key difference between computer experiments and physical experiments is that many computer experiments are "deterministic". In other words, if an experimenter runs the computer code twice at the same input combination $\boldsymbol{x}$, the same response y will be observed both times. A deterministic computer code has no random variability, and may exhibit an unknown, and/or highly complex, functional relationship, $y = f(\boldsymbol{x})$. The lack of variability eliminates any benefit of replication and, further, since *all* the input variables to the code are known and can be controlled, randomization and blocking are likewise unnecessary.

Some deterministic computer codes are very computationally intensive—perhaps requiring hours or days for a computer to solve the underlying mathematical model to produce each response value—so time constraints limit the number of runs or observations that can be taken. Given outputs $y_1, y_2, \ldots, y_n$ from a limited number of runs of the computer code (taken at input combinations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, respectively), a goal of a computer experiment is to fit a model that can generate good response predictions relatively rapidly throughout the set of possible input combinations. The set of possible input combinations is called the *input space* or *design space*. If the fitted model, $\hat{f}(\mathbf{x})$ say, provides a good approximation to the computer code function $f(\boldsymbol{x})$ throughout the design space, and if $\hat{f}(\boldsymbol{x})$ can be computed quickly as compared to running the cumbersome computer code, then the fitted model $\hat{f}(\boldsymbol{x})$ can be used to investigate the behavior of the computer function $f(\boldsymbol{x})$ over the input space. The fitted model is called an *emulator* since it "emulates" the output of the computer code.

In the prosthetic elbow experiment in Example 20.1.1, it would be possible to use the fitted model $\hat{f}(\boldsymbol{x})$ to find a region of inputs $\boldsymbol{x}$ that yield relatively low estimates $\hat{f}(\boldsymbol{x})$ of principal compressive strains in the bone, and finding this good region with respect to $\hat{f}(\boldsymbol{x})$ may be the end of the study. That said, having pinned down this good region of inputs $\boldsymbol{x}$ with respect to $\hat{f}(\boldsymbol{x})$, but knowing that $\hat{f}(\boldsymbol{x})$ merely approximates the computer code function $f(\boldsymbol{x})$, one might make some additional runs of the computer code using input combinations $\boldsymbol{x}$ in or around this good region. With this additional data, a refitted model $\hat{f}(\boldsymbol{x})$ should better approximate the computer code function $f(\boldsymbol{x})$ in the previously identified good region of inputs, allowing one to re-examine the effects of the inputs in this region. Better yet, since the computer code function $y = f(\boldsymbol{x})$ is only a proxy for the true physical situation, having pinned down a good region of inputs with respect to either $f(\boldsymbol{x})$ or $\hat{f}(\boldsymbol{x})$, it may then be feasible, both morally and practically, to run a physical experiment with variable settings $\boldsymbol{x}$ in this good region to study the effects of the input variables on principal compressive strains in the bone directly. This would be done using methods of physical experiments presented in earlier chapters.

For computer experiments, statisticians have developed highly flexible statistical models to emulate simulator output. These are typically smooth "interpolators", in the sense that they provide a fitted model $\hat{f}(\boldsymbol{x})$ which is relatively smooth (no sudden jumps) and, since the observed data are deterministic, $\hat{f}(\boldsymbol{x})$ passes through, or interpolates, the n observed data points $(\boldsymbol{x}_i, y_i)$; in other words, $\hat{f}(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$ for the n observed input combinations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

Fig. 20.1 True unknown relationship between x and y, including seven observed points
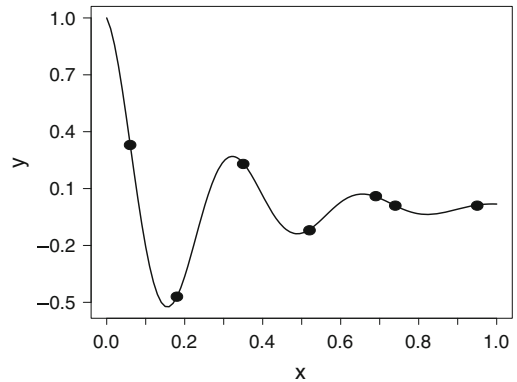
20.2   Models for Computer Experiments

Since the precise functional relationship $y = f(x)$ between the inputs and the response in a deterministic computer experiment is usually unknown a priori, and is potentially highly non-linear, we require the model for the outputs to be very flexible as well as to interpolate the data. While the complete treatment of the technical details behind the statistical models of computer experiment data is beyond the scope of this book, the intuition behind them is quite appealing and is explained in the following example.

Example 20.2.1 `One Predictor`

Suppose $n = 7$ runs of a computer code yield the following data points, written as the pairs (x_i, y_i), $i = 1, \ldots, 7$: $(0.06, 0.33)$, $(0.18, -0.47)$, $(0.35, 0.23)$, $(0.52, -0.12)$, $(0.69, 0.06)$, $(0.74, 0.01)$, $(0.95, 0.01)$. These seven points are plotted in Fig. 20.1 as dark circles. While the relationship $y = f(x)$ intrinsic in the computer code would be unknown to us, for sake of illustration, let it be $y = e^{-4x} \cos(6\pi x)$. The solid curve in Fig. 20.1 represents this true but unknown relationship between x and y, and we want to estimate it based on the seven data points. Suppose we could use the seven data points to somehow obtain a fitted model (emulator), $\hat{y} = \hat{f}(x)$, which we could compute for any value of x in the input space. Then to "estimate" this unknown relationship $y = f(x)$, we could use the fitted model to make predictions $\hat{f}(x)$ at a grid of x values, such as $x = 0, 0.01, 0.02, \ldots, 1$. Plotting these points $(x, \hat{f}(x))$ would approximate the unknown solid curve.

To fit a model with the desired characteristics, consider how we might predict y at any unobserved input point x. We want the fitted model to be an interpolator—namely, for any of the observed points (x, y), we want the model prediction $\hat{y} = \hat{f}(x)$ to match the observed value $y = f(x)$ of the computer code. We also want the fitted model to be relatively smooth. Consider obtaining a prediction $\hat{y}_a$ at, say, the unobserved input point $x_a = 0.36$. In the absence of knowing the nature of the true relationship, the model will "look around" in the x space and it will notice that $x_a = 0.36$ is very close to the observed input point $x_3 = 0.35$, and is between x_3 and $x_4 = 0.52$. Since x_a is very close to x_3, y_a should be very close to the observed value y_3 if the response is smooth and without "spikes". Hence, the model will estimate y_a by a value $\hat{y}_a$ that is close to $y_3 = 0.23$, and perhaps a little smaller to fall between y_3 and $y_4 = -0.12$.

Consider now calculating $\hat{y}$ at a different input point $x_a = 0.45$. The two input points closest to $x_a = 0.45$ for which the computer code was observed are again $x_3 = 0.35$ and $x_4 = 0.52$. Since x_a is approximately halfway between x_4 and x_3, the model could assign roughly the average of y_3 and y_4

as a value for $\hat{y}(x_a)$. Or, since x_a is not too close to either x_3 or x_4, $\bar{y}$ might be a reasonable prediction of y_a. Consider this latter notion further.

To formulate a simple predictor, we might consider $\bar{y}$ and the deviations $y_i - \bar{y}$, $i = 1, 2, \ldots, n$. For example, to make a prediction for a particular input x_a, suppose we use

$$\hat{y}(x_a) = \bar{y} + \sum_{i=1}^{n} w_i (y_i - \bar{y}),$$

where $\sum_{i=1}^{n} w_i (y_i - \bar{y})$ represents a weighted average of the deviations from the mean. The weight w_i should depend on the relative distance between x_a and x_i, with more weight associated with the inputs x_i that are closer to x_a. If only one input point, say x_p, is very close to x_a, this could result in $w_p \approx 1$ with the rest of the weights negligible, in which case one would obtain $\hat{y}_a \approx \bar{y} + (y_p - \bar{y}) = y_p$. For an input x_a sufficiently far away from all of the x_i, all weights w_i could be very small, in which case one would obtain $\hat{y}(x_a) \approx \bar{y} + 0 = \bar{y}$. The model described in Sect. 20.3 has features similar to these. $\qquad \square$

20.3 Gaussian Stochastic Process Model

The model most commonly employed in the analysis of computer experiments is called the *Gaussian Stochastic Process model* (GaSP). The model specifies the deterministic computer code function $y = f(x)$ as the realization of a "Gaussian stochastic process". Although the code is deterministic, this statistical model provides a probabilistic framework for the response at unobserved input combinations while modeling the unknown output surface as being relatively smooth and responses at nearby inputs as being highly correlated. The GaSP model takes the form

$$Y(x) = \beta_0 + Z(x), \tag{20.3.1}$$

which is similar to a regression model in Chap. 8 but with a different type of error variable. Here, $Y(x)$ denotes the response at input combination $x = (x_1, x_2, \ldots, x_d)$, and β_0 is an unknown constant. $Z(x)$ is assumed to be a *Gaussian stochastic process*, which means that, for any choice of ℓ input combinations, $x_1, x_2, \ldots, x_\ell$, the random variables $Z(x_1), Z(x_2), \ldots, Z(x_\ell)$ have a multivariate normal distribution (so, in particular, they are generally not independent). The assumptions on the model are as follows:

(i) $Z(x)$ has mean zero and constant variance σ^2 for any input combination x, so $Z(x) \sim N(0, \sigma^2)$ and

$$Y(x) \sim N(\beta_0, \sigma^2).$$

(ii) For any two inputs x_i and x_j, the correlation between the responses $Y(x_i)$ and $Y(x_j)$ is denoted by $R(x_i - x_j | \xi)$, where $R(x_i - x_j | \xi)$ is a function of the distance between the inputs x_i and x_j (in d-dimensional space), and depends on a set of unknown parameters which, for simplicity, we write as ξ.

Thus,

$$Cov(Y(x_i), Y(x_j)) = Cov(Z(x_i), Z(x_j)) = \sigma^2 R(x_i - x_j | \xi). \tag{20.3.2}$$

Quantifying the correlation between $Y(x_i)$ and $Y(x_j)$ by taking into account the distance between x_i and x_j is a non-trivial problem. One possible structure that one might use is the "Gaussian correlation function", which is given by

$$R(x_i - x_j|\xi) = \prod_{k=1}^{d} e^{-\theta_k(x_{ik}-x_{jk})^2} \tag{20.3.3}$$

$$= e^{-[\theta_1(x_{i1}-x_{j1})^2+\theta_2(x_{i2}-x_{j2})^2+\cdots+\theta_d(x_{id}-x_{jd})^2]},$$

where all $\theta_k > 0$, and where ξ is the set of parameters $(\theta_1, \theta_2, \ldots, \theta_d)$ that need to be estimated from the data. This correlation structure yields a relatively smooth model. Replacing the exponent 2 in $(x_{i1} - x_{j1})^2$ by some positive number p smaller than 2, would result in a less smooth model.

Consider the case of one predictor ($d = 1$). Then the notation simplifies, and the correlation function (20.3.3) reduces to $R(x_i - x_j|\theta) = e^{-\theta(x_i-x_j)^2}$. For any fixed value of $\theta > 0$, the correlation $e^{-\theta(x_i-x_j)^2}$ starts to approach zero as the absolute value $|x_i - x_j|$ of the distance between x_i and x_j gets very large, indicating that $Y(x_i)$ and $Y(x_j)$ are essentially independent so need not be similar. On the other hand, as the distance between x_i and x_j gets very small, the correlation starts to approach 1, indicating that $Y(x_i)$ and $Y(x_j)$ are strongly correlated and so should be similar. Taken to the extreme, if $x_i = x_j$, the correlation between $Y(x_i)$ and $Y(x_j)$ is 1, so $Y(x_i) = Y(x_j)$, i.e. two responses at the same input yield the same value, as desired. Exercise 1 asks the reader to investigate the effect of θ and the absolute distance $|x_i - x_j|$ on the size of the correlation between two responses while using the Gaussian correlation function. Exercises 2–4 show some alternatives to the Gaussian correlation function which are sometimes used in practice.

Under model (20.3.1) and assumption (i), the mean response is simply a constant, β_0. While β_0 could be replaced with a more complicated regression model as in Chap. 8 or 16, doing so typically does not improve the predictive performance of the model for reasonably smooth surfaces (For further reading, see Sacks et al. 1989). Using appropriate methods of prediction, the correlation structure of the Gaussian stochastic process model with constant mean apparently provides adequate model flexibility for reasonably smooth surfaces such as that of Fig. 20.1.

The parameters $(\beta_0, \sigma^2, \theta_1, \ldots, \theta_d)$ in (20.3.1)–(20.3.3) are unknown so need to be estimated. There are a few different ways to estimate these, but we shall use the most common one—maximum likelihood estimation, since such estimators have excellent statistical properties. An in-depth discussion of maximum likelihood estimation is beyond the scope of this book, but maximum likelihood estimates can be computed using appropriate statistical software, and we will rely on software for the computations (as in Sects. 20.6 and 20.7).

For the interested reader, here is the idea of maximum likelihood estimation. Under model (20.3.1), the response variables $Y(x_1), \ldots, Y(x_n)$ follow a multivariate normal distribution which depends on the values of the parameters $\beta_0, \sigma^2, \theta_1, \ldots, \theta_d$. For given values of these parameters, the probability density function (pdf) is a function of the possible data (output) points $y(x_1), \ldots, y(x_n)$. One is more likely to observe data values where the pdf is larger, and most likely to get data where the pdf is at or near its maximum. Conversely, if we have the data $y = (y(x_1), \ldots, y(x_n))$ but the parameters are unknown, it seems reasonable to estimate the parameters by values that would correspond to the observed data being as likely as possible. If we estimate the parameters by values that maximize the likelihood of the observed data, these are *maximum likelihood estimates*, and this is *maximum likelihood estimation*.

Once estimates of the model parameters have been obtained, the next step is to construct a predictor $\hat{Y}(x_a)$ of a response $Y(x_a)$ at any new unobserved input point x_a (cf. Chap. 8). The

predictor should be easily computable but sufficiently sophisticated to utilize the observed data $\boldsymbol{y} = (y(\boldsymbol{x}_1), y(\boldsymbol{x}_2), \ldots, y(\boldsymbol{x}_n))$. The predictor $\hat{Y}(\boldsymbol{x}_a) = \hat{\beta}_0$, for example, would be too simplistic. If the correlation parameters θ_k were known, we could compute and use the *best linear unbiased predictor*—namely, the unbiased predictor that minimizes the mean squared error of prediction. Since the correlation parameters θ_k are unknown, what we can do is compute their maximum likelihood estimates, then use what would be the best linear unbiased predictor if the estimated correlations θ_k were the true values. The resulting predictor is called the *empirical best linear unbiased predictor* (eBLUP).

It can be shown that the eBLUP of $Y(\boldsymbol{x}_a)$ is $\hat{Y}(\boldsymbol{x}_a) = E[Y(\boldsymbol{x}_a)|\boldsymbol{y}]$, which is the mean of $Y(\boldsymbol{x}_a)$ conditional on the observed data $\boldsymbol{y}$ at the observed input combinations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. For model (20.3.1), the formula for the eBLUP of $Y(\boldsymbol{x}_a)$ is most easily written in terms of vectors and matrices. Readers who do not have an algebra background can jump to (20.3.6) and leave it to computer software to do the needed calculations. The formula for the eBLUP is

$$\hat{Y}(\boldsymbol{x}_a) = \hat{\beta}_0 + \hat{\boldsymbol{r}}' \hat{\boldsymbol{R}}^{-1} (\boldsymbol{y} - \mathbf{1}\hat{\beta}_0), \tag{20.3.4}$$

where $\hat{\boldsymbol{r}}$ is an $n \times 1$ vector whose ith element $\hat{r}_i = R(\boldsymbol{x}_a - \boldsymbol{x}_i|\hat{\boldsymbol{\xi}})$ is the estimated correlation between the response $Y(\boldsymbol{x}_a)$ at a new input combination $\boldsymbol{x}_a$ and the ith previous response $Y(\boldsymbol{x}_i)$, $\hat{\boldsymbol{R}}$ is a $n \times n$ matrix whose ijth element $\hat{R}_{ij} = R(\boldsymbol{x}_i - \boldsymbol{x}_j|\hat{\boldsymbol{\xi}})$ is the estimated correlation between observed responses $Y(\boldsymbol{x}_i)$ and $Y(\boldsymbol{x}_j)$, $\mathbf{1}$ is $n \times 1$ vector of 1's, and $\hat{\beta}_0 = (\mathbf{1}'\hat{\boldsymbol{R}}^{-1}\mathbf{1})^{-1}\mathbf{1}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{y}$ is the generalized least squares estimate of β_0.

Under model (20.3.1), the uncertainty about the predicted value can be estimated via the estimated variance of the eBLUP given by

$$\hat{s}^2(\boldsymbol{x}_a) = \hat{\sigma}^2 \left[1 - \hat{\boldsymbol{r}}'\hat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{r}} + \frac{(1 - \mathbf{1}'\hat{\boldsymbol{R}}^{-1}\mathbf{1})^2}{\mathbf{1}'\hat{\boldsymbol{R}}^{-1}\mathbf{1}} \right]. \tag{20.3.5}$$

It can be shown algebraically that $s^2(\boldsymbol{x}) = 0$ for any $\boldsymbol{x}$ for which the response $Y(\boldsymbol{x})$ was already observed, i.e. the model will interpolate the data.

For each possible unobserved input combination $\boldsymbol{x}_a$, the $100(1 - \alpha)\%$ prediction interval can be calculated as

$$\hat{Y}(\boldsymbol{x}_a) \pm z_{\alpha/2}\hat{s}(\boldsymbol{x}_a). \tag{20.3.6}$$

There are two key observations to be made about the predictor $\hat{Y}(\boldsymbol{x}_a)$, whose form is shown in (20.3.4). First, the predicted value at a new input site $\boldsymbol{x}_a$ is the sum of two parts. The first part is the estimated overall mean $\hat{\beta}_0$. The second part is a weighted average of the differences $y(\boldsymbol{x}_i) - \hat{\beta}_0$ between each observed response $y(\boldsymbol{x}_i)$ and the estimated overall mean $\hat{\beta}_0$. In general, the largest weight that would be calculated in (20.3.4) is associated with the term $y(\boldsymbol{x}_i) - \hat{\beta}_0$ corresponding to the $\boldsymbol{x}_i$ that is closest to $\boldsymbol{x}_a$; the second largest weight is associated with the term corresponding to $\boldsymbol{x}_i$ that is second closest to $\boldsymbol{x}_a$, etc. So, similar to Example 20.2.1, the quality of the prediction at $\boldsymbol{x}_a$ is a function of the distances between $\boldsymbol{x}_a$ and the n points $\boldsymbol{x}_i$ for which we have already observed the response. If we have not observed any responses in the vicinity of an input combination $\boldsymbol{x}_a$, then our prediction $\hat{Y}(\boldsymbol{x}_a)$ will default to a value close to $\hat{\beta}_0$ and may be a poor estimate of $Y(\boldsymbol{x}_a)$. On the other hand, if we have collected quite a few observations in a particular region of the input space, we can expect to predict the unknown surface quite well in that region.

Fig. 20.2 True unknown, observed, and predicted relationships between x and y for Example 20.3.1
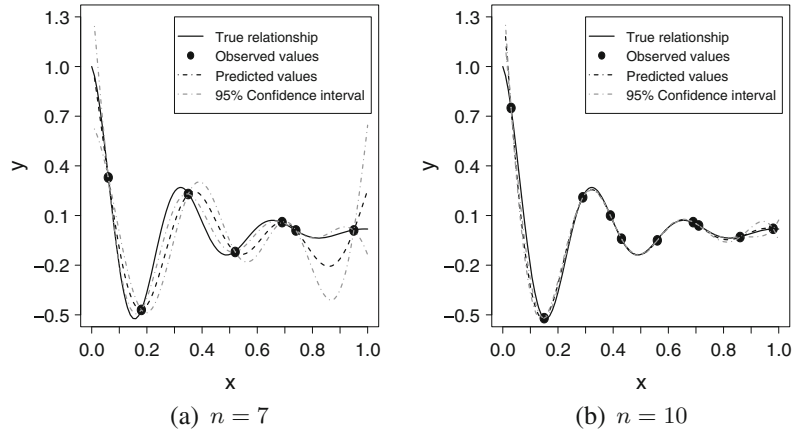
The second important observation is that the predictor $\hat{Y}(\boldsymbol{x}_a)$ shown in (20.3.4) is an *interpolating* predictor. In other words, suppose that $\boldsymbol{x}_a$ is one of the inputs for which we have observed the response, say $\boldsymbol{x}_a = \boldsymbol{x}_p$. Then our predictor will return the observed value $y(\boldsymbol{x}_p)$; that is,

$$\hat{Y}(\boldsymbol{x}_p) = y(\boldsymbol{x}_p) .$$

This behavior from our predictor is desirable because of the deterministic nature of the computer code. If we know that we have observed a response without an error and that we will observe the same response every time we run the computer code for the same input $\boldsymbol{x}_i$, then we would like our predictor to predict the same value for the response.

Example 20.3.1 One Predictor, continued

As in Example 20.2.1, suppose the true but unknown input-output relationship is described by a dampened cosine curve, $y = f(x) = e^{-4x} \cos(6\pi x)$, for $0 \le x \le 1$, as shown in Fig. 20.1. This relationship is shown again as the solid line in Fig. 20.2. Suppose we try to estimate the unknown relationship based on the seven observations (x_i, y_i) that were considered in Example 20.2.1—namely, $(0.06, 0.33), (0.18, -0.47), (0.35, 0.23), (0.52, -0.12), (0.69, 0.06), (0.74, 0.01), (0.95, 0.01)$. These seven data points are shown as the solid dots in Fig. 20.2(a). Using software (see Sects. 20.6 and 20.7) to fit the GaSP model from (20.3.1) with the Gaussian correlation function (20.3.3) and $d = 1$, we obtain the maximum likelihood estimates $\hat{\sigma} = 0.0582$, $\hat{\beta}_0 = 0.0047$, and $\hat{\theta} = 271.95$. Using (20.3.4)–(20.3.6), the predictions and 95% prediction intervals can then be calculated for $x = 0.01, 0.02, \ldots, 1$. The three dashed lines in Fig. 20.2(a) show the predicted curve (the darker middle curve) and the prediction intervals. We can see that our model does a particularly poor job of predicting the true relationship for x values roughly between 0.8 and 0.9. Since there are no observations in this area, the prediction tends to regress to (i.e. predict values closer to) the estimated mean $\hat{\beta}_0 = 0.0047$.

Figure 20.2(b) shows the same information as Fig. 20.2(a), but having fit the model using a set of $n = 10$ different observations. The maximum likelihood estimates are now $\hat{\sigma} = 0.0942$, $\hat{\beta}_0 = 0.0551$, and $\hat{\theta} = 196.17$. Figure 20.2(b) indicates a dramatic improvement in the model's prediction performance. Thus, we see that good design of the computer experiment can be crucial to good prediction. □

20.4 Design

For a computer experiment involving n runs and d input variables, a *design* consists of the n input combinations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ to be run, where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$, $i = 1, \ldots, n$. It is customary to collect these input combinations into an $n \times d$ array or "matrix" $\boldsymbol{X}$, where the kth row of $\boldsymbol{X}$ shows the kth input combination, and we refer to such an array as an $n \times d$ *design* $\boldsymbol{X}$, or an $n \times d$ *design matrix* $\boldsymbol{X}$.

The interpretation of the magnitude of the parameters θ_k in (20.3.3) is related to the scale of the input x_k. More importantly, assessing the sensitivity of the response to the changes in inputs x_k and x_j by comparing the magnitudes of θ_k and θ_j associated with those inputs x_k and x_j can be deceiving if those inputs are defined on vastly different scales. To avoid the potential pitfalls it is common to transform the range of each input variable to become [0,1]. This is done by subtracting the minimum and then dividing by maximum−minimum. For example, if x_1 has an original range of possible values 3.5 to 9.6, we subtract 3.5 from every value of x_1 and divide by $(9.6 - 3.5) = 6.1$. From now on, we will assume that this transformation has been done and that every input variable now has values in the range [0,1].

20.4.1 Space-Filling and Non-collapsing Designs

The design of a computer experiment—namely, the choice of input combinations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ for the computer code—is guided by the deterministic nature of a computer experiment and by the particular characteristics of the predictor discussed in Sect. 20.3. We expect to predict the unknown surface well in the regions of input space where we have observed some nearby responses. Therefore, we would like our initial design to explore the input space "evenly", so we would like our design to be *space-filling*. Designs that are not space-filling will neglect one or more regions of the input space, and we can anticipate poor prediction of the response in such regions.

For example, consider a computer experiment involving $d = 2$ input variables and input combinations $\boldsymbol{x_i} = (x_{i1}, x_{i2})$, with $0 < x_{ik} < 1$ for $i = 1, \ldots, n; k = 1, 2$, listed as the rows of the design $\boldsymbol{X}$. Such input combinations correspond to points in a unit square, where a response can be observed. Consider the two 20-run experimental designs $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ whose design points (x_{i1}, x_{i2}) are plotted in Fig. 20.3. Although we have not yet quantified the space-filling idea, the design $\boldsymbol{X}_1$ seems to do a much better job of exploring the two-dimensional input space than $\boldsymbol{X}_2$. In particular, $\boldsymbol{X}_2$ does not contain any points (x_{i1}, x_{i2}) such that $x_{i2} > x_{i1} + 0.4$ leaving the upper left region of the input space unexplored. Hence, based on our previous discussion, the predictor based on $\boldsymbol{X}_1$ should do a better job of predicting the unknown surface in the upper left region of the input space than the predictor based on $\boldsymbol{X}_2$.

As previously noted, multiple runs of the computer code at the same input combination $\boldsymbol{x} = (x_1, x_i, \ldots, x_d)$ will yield the same output $y(\boldsymbol{x})$ due to the deterministic nature of the computer code, so replicating input points is wasteful. Moreover, when one or more input variables has no effect on the response, it is beneficial to avoid replicating any input combination for *any subset* of the input variables (as explained below). One way to avoid this replication is to ensure that, for each column (i.e. input variable) of such a design $\boldsymbol{X}$, the n input values are distinct. Such a design is called a *noncollapsing design*. To illustrate this notion, consider the following two design matrices, each for four runs (rows) with two inputs (columns):

Fig. 20.3 Two potential
20-run designs for a
computer experiment with
2 input variables

(a) X_1 (b) X_2
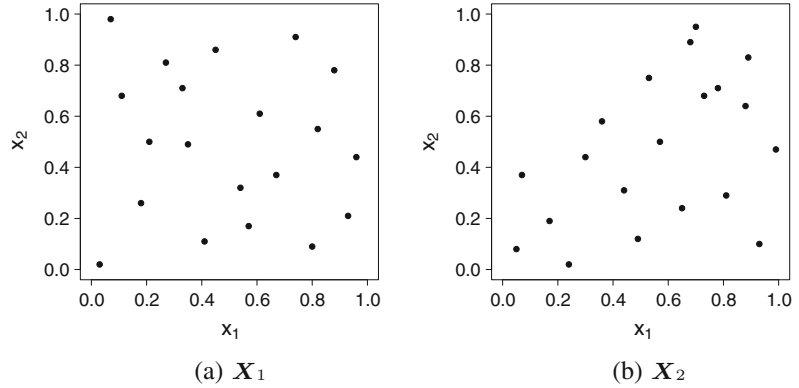
Fig. 20.4 A collapsing
(X_3) and a noncollapsing
(X_4) design

(a) X_3 (b) X_4
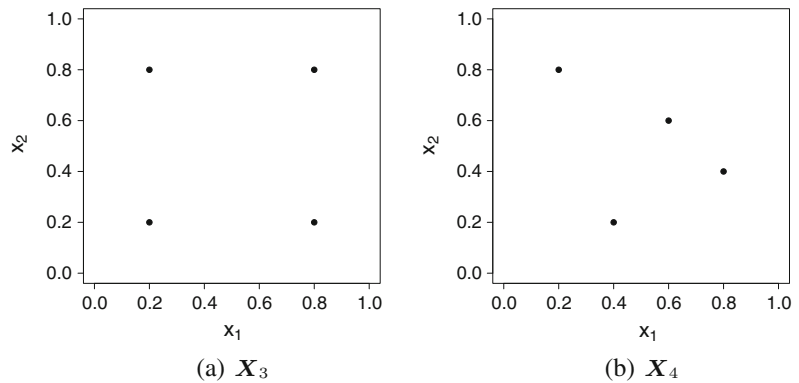
$$X_3 = \begin{bmatrix} 0.2\ 0.2 \\ 0.2\ 0.8 \\ 0.8\ 0.2 \\ 0.8\ 0.8 \end{bmatrix} \text{ and } X_4 = \begin{bmatrix} 0.2\ 0.8 \\ 0.4\ 0.2 \\ 0.6\ 0.6 \\ 0.8\ 0.4 \end{bmatrix}.$$

These designs are depicted in Fig. 20.4. We can see that X_3 is a collapsing design since both of its columns contain replicated values, i.e. if we collapse the design points onto either the x_1- or x_2- axis, we obtain only two distinct values from the four distinct points. X_4 is a noncollapsing design since neither x_1 nor x_2 values are replicated. Without any further considerations, one might wonder what possible problem a collapsing design could cause. Suppose the true, but unknown, response function is given by $y(x_1, x_2) = \exp(-2x_2)\cos(3\pi x_2)$. We see that the response does not depend on the input x_1; a fact not known to a researcher before the experiment. Based on the designs X_3 and X_4, the set of four corresponding observed values are given by $y_3 = (0.21, 0.06, 0.21, 0.06)$ and $y_4 = (0.06, 0.21, -0.24, -0.36)$, respectively. The design X_3 did not contain any replicated input combinations—no two rows were the same. Still, due to replication of values of x_2 and the inactivity of x_1, we ended up with undesirable replicated responses. One might argue that half of our computational resources were wasted, since two of the four runs of the computer code did not add any new information, except perhaps to notice that x_1 may not affect the response. To prevent potential computational wastefulness, and in the light of the fact that we have other tools to identify inactive inputs, we prefer the design X_4.

(a) $\boldsymbol{X}_1$ (b) $\boldsymbol{X}_2$ (c) $\boldsymbol{X}_3$
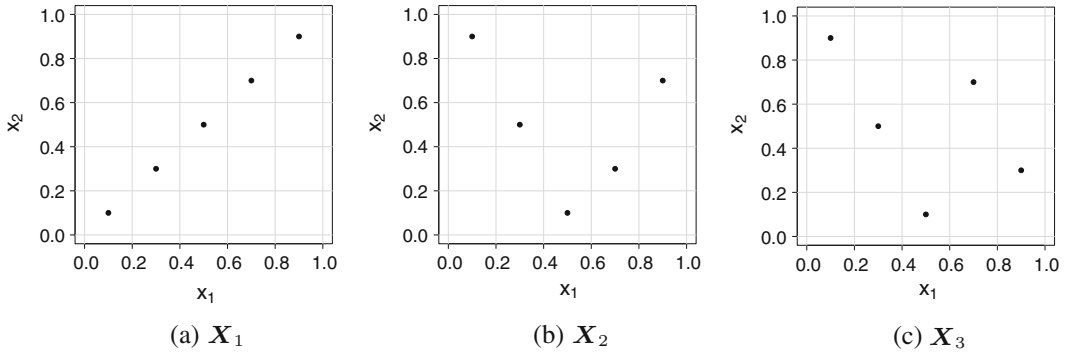
Fig. 20.5 Latin hypercube designs

There are several strategies for constructing designs for computer experiments that are noncollapsing and/or space-filling. The most commonly used approach is that of Latin Hypercube designs, which are inherently noncollapsing and many are also space-filling. We illustrate these designs below.

20.4.2   Construction of Latin Hypercube Designs

To construct an $n \times d$ Latin Hypercube design (LHD), we start by dividing the range $[0, 1]$ of each input variable into n equal-length "bins", with divisions at

$$0, 1/n, 2/n, \ldots, (n-1)/n, 1\,.$$

These bins have midpoints at $(2k-1)/2n$ for $k = 1, 2, \ldots, n$ and, for simplicity, we can use the midpoints to label the bins. For example, in each of Fig. 20.5(a)–(c), we have $n = 5$ input points, and so the bin mid-points for each of x_1 and x_2 are at

$$\frac{2-1}{10} = 0.1, \quad \frac{4-1}{10} = 0.3, \quad \frac{6-1}{10} = 0.5, \quad \frac{8-1}{10} = 0.7, \quad \frac{10-1}{10} = 0.9\,. \qquad (20.4.7)$$

In total, there are 5^2 two-dimensional cells which can be labeled by the pairs of midpoints of the (x_1, x_2) bins. Similarly, in general, dividing the ranges of each of d input variables $x_1, \ldots, x_d$ into n bins leads to n^d possible d-dimensional cells; these are called *hypercubes* and can be labeled using the sets of bin midpoints for the d input variables.

Next, we want to choose n of these n^d cells (hypercubes) to contain a design point in such a way that there is only one design point in each bin for each individual input variable. The design is then non-collapsing. For example, if the 5 input points in each of the designs depicted in Figs. 20.5(a)–(c) are projected onto each axis, it can be seen that they rest on the midpoints (20.4.7) of the 5 bins for both x_1 and x_2. One way of achieving such a choice is as follows.

(a) Start with an array X with n rows and d columns where each column contains the integers $1, 2, \ldots, n$.
(b) For each column separately, randomly order the integers.
(c) Replace each integer k by the corresponding bin midpoint $(2k-1)/2n$.

(d) The ith row of of the randomly ordered X then determines the ith input combination for the experiment.

Example 20.4.1 `5 × 2 Latin hypercube designs`

Suppose that a design is required with $n = 5$ input points for a computer experiment where the simulator has $d = 2$ input variables. Following the above procedure, each of the d columns of X starts with the values 1, 2, 3, 4, 5 in order. Each column is randomly ordered and the integers k are replaced by the bin midpoints $(2k - 1)/2n = 0.1, 0.3, 0.5, 0.7, 0.9$ to give the final design. Three possible designs are

$$X_1 = \begin{bmatrix} 0.1 & 0.1 \\ 0.3 & 0.3 \\ 0.5 & 0.5 \\ 0.7 & 0.7 \\ 0.9 & 0.9 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 0.3 & 0.5 \\ 0.5 & 0.1 \\ 0.9 & 0.7 \\ 0.1 & 0.9 \\ 0.7 & 0.3 \end{bmatrix}, \quad \text{and } X_3 = \begin{bmatrix} 0.5 & 0.1 \\ 0.3 & 0.5 \\ 0.1 & 0.9 \\ 0.9 & 0.3 \\ 0.7 & 0.7 \end{bmatrix},$$

where each row gives the coordinates of an input combination $x_i = (x_{i1}, x_{i2})$. Designs X_1, X_2, and X_3 are the three designs depicted in Fig. 20.5a–c. Since all these designs are LHDs, they are noncollapsing.

By visual inspection, we notice that the design X_1 is the least space-filling and does the poorest job of exploring the input space. Although this design does a really good job of exploring the space along the main diagonal, that is all that it does well. Since no points are placed away from the main diagonal, there is no exploration of the input space towards the upper left and lower right corners of the space. If we use this design to run our experiment, we can expect very poor predictions the further we get away from the main diagonal. Even though X_1 is an LHD, it is clearly not space-filling, so is a poor design for a computer experiment. Both X_2 and X_3 do better at space filling, and which is the better of these two designs is less clear (although a preference will be expressed for X_3 in Example 20.4.2). □

To seek a design which is both noncollapsing and space-filling, experimenters often look for a Latin hypercube design that best satisfies some space-filling property. There is a vast literature describing many attempts to define what we mean by a "best" design with respect to space-fillingness. One of the most common approaches is to choose a design which maximizes the distance between the two closest points in the design. A common measure of distance is the familiar Euclidean distance between $x_i = (x_{i1}, \dots, x_{id})$ and $x_j = (x_{j1}, \dots, x_{jd})$; that is,

$$\sqrt{\sum_{\ell=1}^{d} (x_{i\ell} - x_{j\ell})^2} \,. \tag{20.4.8}$$

Suppose that we have two $n \times d$ designs X_1 and X_2. For each design, we can determine the minimum distance between pairs of input points. The design with the larger value of this minimum distance is considered preferable, since choosing a design in this way forces the design points to be far away from each other and so tends to be more space-filling. Among all possible designs, any design that *max*imizes this *min*imum interpoint distance is called a *maximin design*.

Example 20.4.2 `5 × 2 Latin hypercube designs, continued`

One could compare the designs X_1, X_2, and X_3 of Fig. 20.5 based on the maximin criterion. The corresponding interpoint distances (rounded to two decimal places) are shown in Table 20.1. Notice that the smallest interpoint distance is 0.28 for both X_1 and X_2, compared with the minimum interpoint distance of 0.45 for X_3. Hence, among these three Latin hypercube designs, X_3 is the maximin design.

Table 20.1 Interpoint distances for the designs of Fig. 20.5

	X_1				X_2				X_3			
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
x_2	0.28				0.45				0.45			
x_3	0.57	0.28			0.63	0.72			0.89	0.45		
x_4	0.85	0.57	0.28		0.45	0.89	0.82		0.45	0.63	1.00	
x_5	1.13	0.85	0.57	0.28	0.45	0.28	0.45	0.85	0.63	0.45	0.63	0.45

If we now look at X_1 in Fig. 20.5(b) again, we can see that the two points nearest the x_1 axis are close together in the two-dimensional space, whereas all the points lie further apart in X_3 in Fig. 20.5(c), leading to a slightly better coverage of the design space. $\qquad\square$

Exercises 7 and 8 show alternative measures of space-fillingness ("minimax" and "average reciprocal distance"). There are other measures, too, in the literature. But, in this chapter, we will concentrate on the most common measure of "maximin". (For further types of designs, see Santner et al. 2003, Chaps. 5 and 6).

In practice, to construct a space-filling, noncollapsing design, one would use the aid of a computer. For example, given the number of runs n and inputs d, one could use the computer to generate lots (several hundred) Latin hypercube designs X, by using many random orderings of the columns, compute the minimum interpoint distance for each design, and keep the design with the largest minimum interpoint distance. Although this may not result in *the* maximin design, it should result in one that is close to maximin and which has good space-filling properties. Sections 20.6.1 and 20.7.1 show how to generate maximin LHDs using the SAS and R software, respectively.

As a final note, design points do not necessarily need to be placed at the midpoint of the cell. One can place a design point in a randomly chosen location in the cell by adding a random number between $-(2n)^{-1}$ and $+(2n)^{-1}$ to every x_{ij} in X.

20.5   A Real Experiment—Neuron Experiment

One of the authors (Danel Draguljić) was involved in a study aimed at modeling the performance of neurons. The study, which is described by Rumbell and coauthors in *Journal of Computational Neuroscience* in 2016, aimed to improve the fits of conductance-based models to in vitro whole cell recordings from pyramidal neurons of layer 3 of the prefrontal cortex from young and aged rhesus monkeys. The neuron's performance was measured by its firing rate (response, Y) which was modeled with either 4 or 8 ion channels resulting in 10 or 23 explanatory variables, respectively, to be considered in the model.

Table 20.2 shows the data from a simplified model for the firing rates of a neuron at +380 pA current injection of a young monkey. The simplified model contained two input variables; x_1 was the maximal conductance of the transient sodium, denoted g_{NaF}, and x_2 was the maximal conductance of the delayed-rectifier potassium, denoted g_{KDR}. Both of these variables affect the neuron firing rate. Both maximal conductances had original ranges between 0.05 and 0.5 mS/cm^2, but the g_{NaF} and g_{KDR} values in Table 20.2 have been scaled to the [0, 1] range, as throughout this chapter. The design chosen for the input variables was a 30×2 Latin hypercube design and is shown in Fig. 20.6(a).

Columns 1 and 2 of Table 20.2 show the first 15 values of the input variables, and columns 4 and 5 show the next 15 values. For a given value of g_{NaF} and g_{KDR} the computer simulator was run for 2

Table 20.2 The data for the neuron experiment

g_{NaF}	g_{KDR}	y_i	g_{NaF}	g_{KDR}	y_i
0.38594	0.21207	33	0.53994	0.83528	41
0.04667	0.45947	0	0.13236	0.93565	1
1.00000	0.44733	46	0.90811	0.58219	46
0.95468	0.33514	44	0.88221	0.09805	39
0.53335	0.79813	41	0.39918	0.55403	36
0.59167	0.60427	41	0.22713	0.67680	34
0.18570	0.37995	31	0.82599	0.00000	13
0.49928	0.24442	36	0.69149	0.06387	36
0.74609	0.39496	42	0.79710	0.01235	19
0.07269	1.00000	0	0.31326	0.12559	30
0.27908	0.27845	32	0.62431	0.63717	41
0.34985	0.51466	35	0.97649	0.86049	48
0.64658	0.89122	43	0.46554	0.76087	39
0.20972	0.71452	34	0.11439	0.31166	34
0.00000	0.95918	0	0.77811	0.15813	39

(a) 30×2 LHD   (b) Firing rates
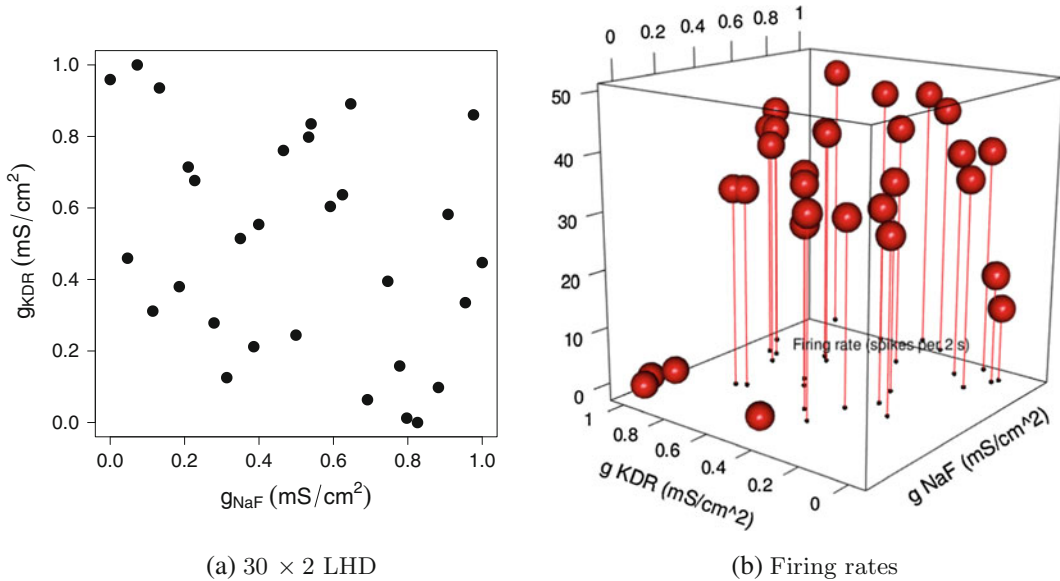
Fig. 20.6 The design and the data for the neuron experiment

seconds and the number of spikes, y_i, was recorded. These observed firing rates are shown in columns 3 and 6 of Table 20.2 and displayed in Fig. 20.6(b). We can observe that relatively low firing rates occur for large values of g_{KDR} and small values of g_{NaF}. For example, if we are trying to maximize neuron's firing rate, it seems that our best chance lies with values in the upper ranges of both variables.

We would like to use our data to construct an estimator $\hat{y}$ of $Y = f(g_{NaF}, g_{KDR})$. The estimator $\hat{y}$ would then enable us to identify the regions in the (g_{NaF}, g_{KDR}) space that are associated with either low or high neuron activity. After fitting the GaSP model (20.3.1) with Gaussian correlation function (20.3.3), we obtain the maximum likelihood estimates $\hat{\sigma} = 15.87$, $\hat{\theta}_1 = \hat{\theta}_{NaF} = 5.03$, and

Fig. 20.7 Observed values and predicted surface for the neuron experiment
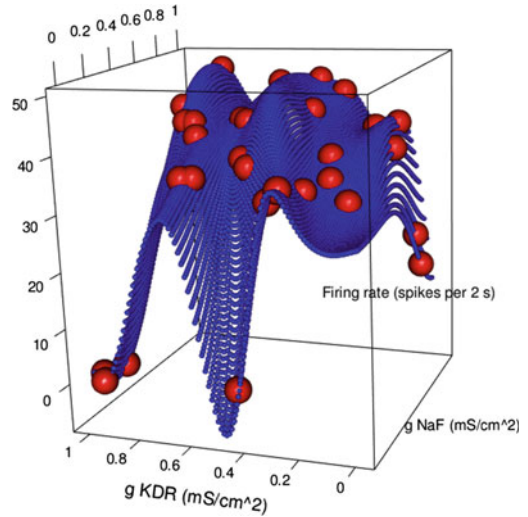
$\hat{\theta}_2 = \hat{\theta}_{\text{KDR}} = 50.22$, giving $\hat{\beta}_0 = 27.61$. Using the predictor in (20.3.4), we calculated the values of $\hat{Y}$ at a dense grid of input points over a $[0, 1] \times [0, 1]$ square. The predicted outputs are shown as the blue points in Fig. 20.7, together with the original data from the simulator (red points). The predicted maximum firing rate is 49.11 spikes per 2 seconds and it occurs at the point $(g_{\text{NaF}}, g_{\text{KDR}}) = (0.86, 0.86)$. The predicted minimum firing rate is -7.52 spikes per 2 seconds and it occurs at the point $(g_{\text{NaF}}, g_{\text{KDR}}) = (0, 0.49)$. We know that firing rate cannot be negative, but there is no intrinsic mechanism in the GaSP model to prevent it from making negative predictions. At this point, we could run the simulator at $(g_{\text{NaF}}, g_{\text{KDR}}) = (0, 0.49)$ and obtain $y(0, 0.49)$, then refit our model using all $30 + 1$ observations, and look for the minimum again. Alternatively, if we had been concerned with avoiding negative predictions, we could have considered transforming our data, say to $\ln(y)$, to fit our model and then exponentiate the predicted values.

In this example, suppose we take two input points x_1 and x_2 such that $x_1 = (x_{11}, x_{12})$ and $x_2 = (x_{11} + h, x_{12} + h)$ for some $h > 0$. Then, the estimated correlation between $Y(x_1)$ and $Y(x_2)$ is given by

$$\hat{R}(x_1 - x_2 | \hat{\theta}_1, \hat{\theta}_2) = e^{-5.03h^2 - 50.22h^2} = e^{-5.03h^2} e^{-50.22h^2} = \hat{R}_1(h)\hat{R}_2(h).$$

$\hat{R}_1(h)$ and $\hat{R}_2(h)$ are shown in Fig. 20.8. Since $\hat{R}(x_1 - x_2 | \hat{\theta}_1, \hat{\theta}_2)$ is the product of $\hat{R}_1(h)$ and $\hat{R}_2(h)$, Fig. 20.8 indicates that, for all h larger than 0.1, $\hat{R}(x_1 - x_2 | \hat{\theta}_1, \hat{\theta}_2) \approx 0$ because $\hat{R}_2(h) \approx 0$. In other words, suppose that two points are more than 0.1 mS/cm^2 apart in the g_{KDR} space then, even if they are very close in the g_{NaF} space, the observations at those points are essentially treated as uncorrelated.

Here, and in general, we conclude that changes in an input variable x_k associated with a large value of θ_k will tend to have little impact on the size of the correlation. In other words, as seen by the dotted line in Fig. 20.8, if θ_k is large, the two design points have to be extremely close to each other for their corresponding observations to have sizable correlation.

20.6 Using SAS Software

In this section, we illustrate use of the SAS software for creating Latin hypercube designs (LHDs) and for analyzing data obtained from computer experiments.

Fig. 20.8 Correlation as a
function of distance h
between two input points
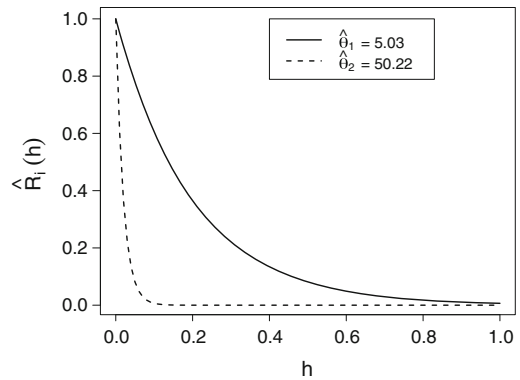
20.6.1   Maximin Latin Hypercube Designs

This section shows how to use SAS software to get an LHD that is approximately maximin. The idea
behind constructing a maximin LHD (Mm LHD) is described in Sect. 20.4.2. First, we construct an
$n \times d$ LHD. Second, for the newly constructed LHD we calculate all $\binom{n}{2}$ interpoint Euclidean distances
and then find the smallest one, i.e. we identify the distance between two closest points. And third,
we iterate this process many times and, among many created LHDs, we identify the one with largest
minimum interpoint distance. It is worth noting that the "best" design obtained after running the SAS
code from this section contains the maximum minimum interpoint distance among the number of
designs examined by the code. Another run of the same code with the same inputs might produce a
design with even larger maximum minimum interpoint distance.

The SAS programs shown here are in the form of SAS "macros". A macro is a piece of SAS
code that can be stored separately and is extremely useful when SAS software has to repeat the same
calculations many times, as it does in the search for the best LHD. Table 20.3 shows a SAS macro.
A macro begins and ends with SAS statements `%macro` and `%mend`, respectively. The macro in
Table 20.3 is called `createLHD` and it has two inputs n and d, which will be provided when the
macro is "called" (implemented). The body of a macro contains SAS statements that can be run many
times by calling the macro which eliminates copying and pasting the statements themselves. Before
being able to use a macro, we have to run the whole macro. Once this is done, the SAS software has
this macro available for use. To then execute the SAS statements within the macro `createLHD`, we
run for example `%createLHD(n = 20, d = 2)` where, here, values 20 and 2 have been selected
for n and d. These values can be changed to suit the user's needs.

Within this macro, the `%` symbol indicates to the SAS software that this is a segment of the program;
here we have two (nested) loops starting at `%do` and ending at `%end`. The `&` symbol in front of the
variables `n, d, i` and `j` indicates that, within the macro, the variables have been assigned specific
values (the last two change as we proceed around the loops). Writing macros is beyond the scope of
this book, but we note that SAS code that is inside of a macro can be taken out and be used with some
necessary clean-up, such as removing the `%` and `&` symbols.

The SAS code to find an approximately maximin LHD is shown in the form of three macros in
Tables 20.3, 20.4, and 20.5, respectively. The macro `createLHD`, shown in Table 20.3, constructs an
LHD with n rows and d columns, where d is the number of inputs to the computer simulator and
n is the required number of outputs. The LHD is constructed by filling each of the d columns with
$(2k - 1)/2n$ for $k = 1, 2, \ldots, n$ and then the rows of each column are randomly rearranged using the
`ranuni(0)` command described in Sect. 3.8.1. To then construct, say, a 20×2 LHD, one simply

Table 20.3 A SAS program for construction of an LHD

```
%macro createLHD(n=, d=);
  %do i = 1 %to &d;
    data temp&i;
      %do j = 1 %to &n;
        x&i = (2*&j - 1)/(2*&n) + (1/&n)*ranuni(-1) - 1/(2*&n);
        ranno&i = ranuni(0);
        output;
      %end;
    run;
    proc sort data = temp&i; by ranno&i; run;
    data temp&i; set temp&i; drop ranno&i; run;
  %end;
  data lhd; merge temp1 - temp&d; run;
  proc print data = Lhd; run;
%mend createLHD;
```

runs `%createLHD(n = 20, d = 2)` line. A 20×2 LHD will then be created, printed out to the SAS *Results Viewer* window, and stored in the SAS data set called `Lhd`.

The macro shown in Table 20.4 is `mipd`. This macro calculates minimum interpoint distance for a user-provided LHD. For example, if we created an LHD using `createLHD` macro, we can then calculate this LHD's minimum interpoint distance by running `%mipd(x = Lhd, d = 2)` where input *x* is the SAS data set that contains the LHD and *d* is the number of columns in this LHD. The minimum interpoint distance will be printed to the *Results Viewer* window and it will be saved in the SAS data set `Mindist`.

In order to create a maximin LHD, we do not have to call the `createLHD` and `mipd` macros explicitly. Macro `MmLHD` shown in Table 20.5 does all the work for us. In addition to other SAS statements, `MmLHD` calls macros `createLHD` and `mipd` with the appropriate inputs for each. Even though we do not call `createLHD` and `mipd` directly, we do have to run them before we run and call `MmLHD`. The macro `MmLHD` requires three inputs itself: `nrow` and `ncol` define the size of the LHD (*n* and *d*) and the integer `iter` tells the SAS software how many LHDs we want the SAS software to create from which to choose the one with largest minimum interpoint distance. If we run `%MmLHD(nrow = 30, ncol = 3, iter = 100)` the SAS software will create one hundred 30×3 LHDs and it will find the one with the largest minimum interpoint distance. This design with largest minimum interpoint distance is stored in the SAS data set `Best`. To see the design `Best`'s smallest interpoint distance, we can run `%mipd(x = Best, d = 3)`.

All designs created by `MmLHD` have input points placed at random within the selected LHD cells. If we want points placed at the midpoints of the selected cells, we have to remove `+ (1/&n)*ranuni(-1) - 1/(2*&n)` from the fifth line in the macro `createLHD`.

After creating the approximate maximin LHD, the simulator can be run with input combinations **x** defined by the *n* rows of the design stored in `Best`.

Table 20.4 A SAS program for calculation of minimum interpoint distance

```
%macro mipd(x=, d=);
  proc distance data = &x out = dist method = Euclid nostd;
    var interval(x1 - x&d); run;
  data dist;
    set dist;
    array var _numeric_;
    do over var;
      if var = 0 then var = .;
    end;
  run;
  proc means nolabels data = dist min noprint;
    output out = colmins; run;
  data colmins;
    set colmins;
    where _STAT_ = "MIN";
    drop _TYPE_ _FREQ_ _STAT_;
  run;
  proc transpose data = colmins out = colminslong; run;
  proc means nolabels data = colminslong min;
    var COL1;
    output out = Mindist min = minimum; run;
%mend mipd;
```

20.6.2 Fitting the GaSP Model

To use the SAS software, we view GaSP model as a mixed model where all observations are taken on the same subject and are therefore correlated. This approach allows us to fit the GaSP model using PROC MIXED (cf. Sect. 17.10).

Table 20.6 shows a SAS program for the analysis of the neuron computer experiment data of Table 20.2, p. 777. The goal of the experiment was to quantify the effect of g_{NaF} and g_{KDR} on a neuron's firing rate.

The 30 sets of data values are read in line by line via the first set of DATA statements in Table 20.6. In the second set of DATA statements a $101^2 \times 3$ data set PREDS is constructed, whose first two columns contain the values of g_{NaF} and g_{KDR} for which we would like to predict the firing rate once the model is fitted. These values consist of a 2-dimensional grid with "mesh size" of 0.01. The mesh size defines the spacing on the grid, so a grid with mesh size 0.01 has points (0, 0), (0, 0.01), (0, 0.02), …, (1, 0.99), (1, 1). The third column in PREDS is y and it has all values set to missing. The third DATA statement constructs the new data set FORANALYSIS by appending PREDS to the original data, so that the data set FORANALYSIS has $30 + 101^2$ rows of data. The GaSP model (20.3.1) is fitted with PROC MIXED. When fitting the model, the SAS software will use only the observations whose y value is available. If the value for y is missing, the SAS software will use the fitted model to calculate $\hat{y}$ for the observations with missing y values.

There are several options in PROC MIXED that require some clarification. By default, SAS software will choose the restricted maximum likelihood as the method of estimation (see Sect. 19.8.3). The METHOD = ML option instructs the SAS software to use maximum likelihood estimation (p. 769). By leaving the right side of the MODEL equation empty, we instruct the SAS software to fit an intercept only model, i.e. estimate only β_0. If we assumed a richer mean structure, say $E[Y] = \beta_0 + \beta_1 g_{\text{NaF}}$, the

Table 20.5 A SAS program for construction of maximin LHD

```
%macro MmLHD(nrow=, ncol=, iter=);
  %createLHD(n = &nrow, d = &ncol)
  %let design = lhd;
  %mipd(x = &design, d = &ncol)
  data Best; set &design; run; quit;
  data _null_;
    set mindist;
    if _N_ = 1 then call symput("mipd", minimum);
    else stop;
  run; quit;
  %do k = 2 %to &iter;
    %createLHD(n = &nrow, d = &ncol)
    %mipd(x = &design, d = &ncol)
    data _null_;
      set mindist;
      if _N_ = 1 then call symput("mipdnew", minimum);
      else stop;
    run; quit;
    %if &mipdnew > &mipd %then %do;
      %let mipd = &mipdnew;
      data Best; set &design; run; quit;
    %end;
  %end;
  proc datasets library = work noprint;
    delete temp1 - temp&ncol mindist colmins colminslong dist lhd;
  run; quit;
  proc print data = Best noobs; run;
%mend MmLHD;
```

statement would have been MODEL = g_NaF. The option OUTP = PREDICTIONS will calculate predictions and save them in a data set called PREDICTIONS.

The REPEATED statement is used to specify the covariance structure. The SP(EXPA) choice in the TYPE option specifies the "anisotropic exponential spatial" covariance structure which has the structure of (20.3.2) with $R(\boldsymbol{x}_i - \boldsymbol{x}_j | \boldsymbol{\xi})$ being the Power exponential correlation function introduced in Exercise 2. The Gaussian correlation function from (20.3.3) is a special case of the Power exponential correlation function where $p_k = 2$ for $k = 1, 2, \ldots, d$, and will be specified below. The SP(EXPA) statement is followed by the list of the input variables ($g_{NaF} \; g_{KDR}$). Then SUBJECT = INTERCEPT identifies all data as coming from one subject and forces the SAS software to model the covariance between any two observations, so that no covariance is set to zero.

The PARMS statement specifies the initial values for the covariance parameters. We can request a grid of values for each parameter to initiate the optimization algorithm which then searches for the parameter values that maximize the likelihood function. For the Power exponential correlation function, the SAS software expects $2d + 1$ initial values ($\theta_1, \ldots, \theta_d$, $p_1, \ldots, p_d$, and σ^2). In this example, the covariance parameters to be estimated are $\theta_1 (= \theta_{\text{NaF}})$, $\theta_2 (= \theta_{\text{KDR}})$, p_1, p_2, and σ^2. To select the Gaussian correlation function, we need to set p_1 and p_2 equal to 2. Following the PARMS statement are the initial values. We must specify the values in the order in which they appear in the SAS output Covariance Parameter Estimates table (see Fig. 20.9). The first set of initial values is a set for θ_1 and, here, we provided a grid of 10 initial values, $\{1, 2, \ldots, 10\}$. For θ_2 we provided a grid of 41 initial

Table 20.6 A SAS program for analysis of the neuron experiment

```
DATA NEURON;
  INPUT g_NaF g_KDR y;
  LINES;
  0.38594 0.21207 33
  0.04667 0.45947 0
   :   : :   :       :
  0.77811 0.15813 39
;
* Create a grid of values for prediction and append to the data set;
DATA PREDS;
  DO i = 0 TO 100;
    DO j = 0 TO 100;
      g_NaF = i/100; g_KDR = j/100; y = .;
      OUTPUT;
    END;
  END;
  KEEP g_NaF g_KDR y;

DATA FORANALYSIS; SET NEURON PREDS;

PROC MIXED METHOD = ML;
  MODEL y = / OUTP = PREDICTIONS;
  REPEATED / TYPE = SP(EXPA)(g_NaF g_KDR) SUBJECT = INTERCEPT;
  PARMS (1 TO 10 BY 1) (48 TO 52 BY 0.1) (2) (2) (60) / HOLD = 3, 4
    LOWERB = 1, 45, 2, 2, . UPPERB = 10, 55, 2, 2, .;
  ESTIMATE 'INTERCEPT' INTERCEPT 1;

DATA FORPLOT; SET PREDICTIONS; IF Y = .;
PROC G3D; PLOT g_NaF*g_KDR = Pred; RUN;
```
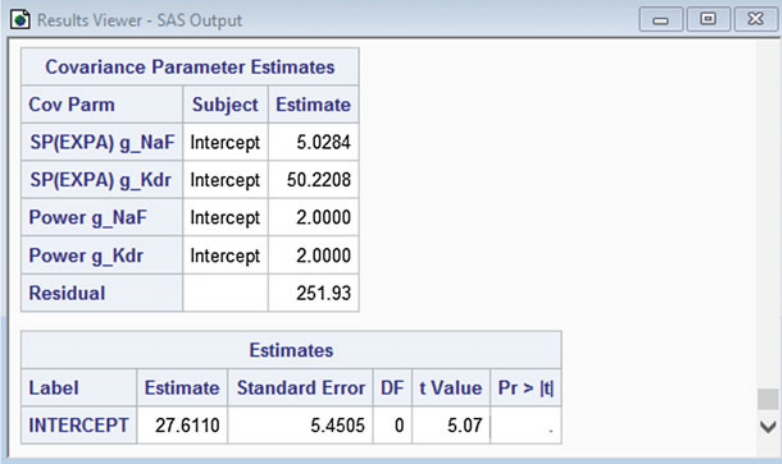
values, $\{48, 48.1, 48.2, \ldots 52\}$. Next, single starting values of 2, 2, and 60 were provided for p_1, p_2, and σ^2, respectively. The HOLD=3 , 4 option tells the SAS software to keep values for the third and fourth parameter fixed, in this case p_1 and p_2, thereby insuring that we are using the Gaussian correlation function. The LOWERB and UPPERB options set the lower and upper bounds for each parameter. If a bound is missing, then the values for that parameter are unconstrained in that direction. In Table 20.6, the SAS software will run its optimization algorithm for $10 \times 41 \times 1^3$ different starting sets of values with lower bounds 1 and 45, and upper bounds 10 and 55, for θ_1 and θ_2, respectively.

Choosing the starting values is somewhat of a trial and error process. We should choose a relatively coarse grid of starting values that covers the space for each parameter and let the SAS software provide the estimates based on this coarse grid. This initial coarse search will hopefully prevent the algorithm from returning a "local maximum" (e.g. the top of a small hill but missing the top of the nearby mountain). Once these initial estimates are provided, we can construct a fine grid around these estimates to run the search again in this region to obtain the final estimates.

By default the SAS software will not show $\hat{\beta}_0$ in its output window. The ESTIMATE statement requires it to do so. The partial output from PROC MIXED displaying the parameter estimates for the GaSP model is shown in Fig. 20.9. It shows the maximum likelihood estimates of θ_1 and θ_2, the fixed p_1 and p_2, and the maximum likelihood estimate of σ^2 (labeled "Residual").

The predictions for the data in PREDS are calculated and stored in data set PREDICTIONS. The last two lines in Table 20.6 are used to make a plot of the predicted response surface plot similar to the

Fig. 20.9 GaSP model
estimates—the neuron
experiment

one in Fig. 20.7. The DATA statement creates a data set to be plotted by keeping only the predictions.
The PROC G3D creates the 3D plot (the actual plot is not shown here).

20.7 Using R Software

In this section, we illustrate use of the R software for creating Latin hypercube designs (LHDs) and
for analyzing data obtained from computer experiments.

20.7.1 Maximin Latin Hypercube Designs

This section shows how to use R to get an LHD that is approximately maximin (denoted Mm LHD).
The idea behind constructing an Mm LHD is described in Sects. 20.4.2 and 20.6.1. While it would have
been fairly straight-forward to write our own R code similar to the SAS code shown in Sect. 20.6.1 to
construct an approximate Mm LHD, we can use R's package lhs for this purpose. To construct an
appropriate Mm LHD, we use lhs's function maximinLHS which has three inputs. These are n, the
planned number of outputs of the computer simulator, k, the number of inputs to the computer simulator,
and dup, an integer tuning parameter that determines the number of candidate points considered by
maximinLHS while constructing an Mm LHD. Larger values of dup lead to more points considered
but also increase the time needed to construct a design. dup $= 5$ is suggested as a reasonable choice.

For example, to construct an approximate 30×3 Mm LHD named Best, we run the R command
Best <- maximinLHS(30, 3, 5). After creating the approximate Mm LHD, the simulator
can be run with input combinations x defined by the $n = 30$ rows of the design stored in Best.
Currently, maximinLHS does not support construction of designs whose points are at the midpoint
of the cells, so the points in Best will be randomly placed within the selected cells.

Table 20.7   An R program and selected output for analysis of the neuron experiment

```
> neuron <- read.table("data/neuron.txt", header = T)
> head(neuron, 3)

        g_NaF   g_KDR  y
[1,] 0.38594 0.21207 33
[2,] 0.04667 0.45947  0
[3,] 1.00000 0.44733 46

> # install.packages("mlegp")
> library(mlegp)
> gasp <- mlegp(neuron[, 1:2], neuron[, 3])
> summary(gasp)

Total observations = 30
Dimensions = 2
mu = 27.6111
sig2: 251.9121

Correlation parameters:
       beta a
1  5.027256 2
2 50.233487 2

Log likelihood = -104.4487

> predictedX <- expand.grid(g_NaF = seq(0, 1, 0.01),
+                           g_KDR = seq(0, 1, 0.01))
> yhats <- predict(gasp, predictedX)
> # install.packages("rgl")
> library(rgl)
> plot3d(neuron[, 1], neuron[, 2], neuron[, 3],
+        col = "red", size = 3, type = "s",
+        xlab = "g NaF (mS/cm^2)", ylab = "g KDR (mS/cm^2)",
+        zlab = "Firing rate (spikes per 2 s)")
> plot3d(predictedX[, 1], predictedX[, 2], yhats, col = "blue",
>        size = 0.5, type = "s", add = TRUE)
```

20.7.2   Fitting the GaSP Model

In this section, we illustrate the analysis of computer experiment data with R software, using the neuron experiment data in Table 20.2, p. 777. The goal of the experiment was to quantify the effect of g_{NaF} and g_{KDR} on a neuron's firing rate.

Table 20.7 contains the R program and selected output illustrating the fitting of the GaSP model to the neuron experiment data. The first lines of Table 20.7 read the data from the file neuron.txt and print out the first three rows of the data. The model is then fit using the mlegp function from the R package mlegp, which computes maximum likelihood estimates (mle) for Gaussian (stochastic) processes (gp). To calculate the maximum likelihood estimates of the GaSP parameters, the mlegp function requires two inputs. The first one is the design matrix X defined here by the first two columns of neuron data. The second one is the vector of the observed responses y given here in the third column of the neuron data set. The mlegp function has the capability to fit noisy data, data with

multiple responses, etc. For more information on the `mlegp` function see R's help file by typing `?mlegp`. The model details are stored in `gasp`.

The maximum likelihood parameter estimates can be viewed by invoking the `summary` command. In the resulting output, $\hat{\beta}_0 \approx 27.61$ is labeled as `mu`, and $\hat{\sigma}^2 \approx 251.91$ is labeled as `sig2`. The correlation parameters are given in a $d \times 2$ array where the first and second columns correspond to $\hat{\theta}_k$ and $\hat{p}_k$ for $k = 1, 2, \ldots, d$ but are labeled `beta` and `a`, respectively. $\hat{\theta}_1 \approx 5.03$ and $\hat{\theta}_2 \approx 50.23$. Since we fitted a Gaussian correlation function, $p_1 = p_2 = 2$.

Having fitted the model, we would like to construct the estimated response surface by predicting the response at a dense grid over the $[0, 1] \times [0, 1]$ experimental region. The function `expand.grid` takes two sequences $\{0, 0.01, 0.02, \ldots, 1\}$ of length 101 as inputs and creates a "data frame" whose 101^2 rows represent all combinations of the members of the two sequences (e.g. all combinations of two sequences $\{.3, .6\}$ would be $(.3, .3), (.3, .6), (.6, .3), (.6, .6)$). The resulting grid is stored in `predictedX`. To make the predictions `yhats`, we use R's `predict` function, first specifying `gasp` as the object that holds the estimated parameters and then specifying `predictedX` as the object that holds the new prediction sites. Depending upon the computer being used, calculating the `yhats` could take some time. The `plot3d` function from the R graphics library package `rgl` is used to produce the plot shown in Fig. 20.7, p. 778.

Exercises

1. **Gaussian correlation function**

   The Gaussian correlation function $R(\boldsymbol{x}_i - \boldsymbol{x}_j | \boldsymbol{\xi})$ introduced in Sect. 20.3, p. 768, quantifies the correlation between outputs at two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ based on the distance between them. For parts (a) and (b) below, consider the case of $d = 1$ input variable.

   (a) To investigate the effect of the value of parameter θ on the correlation between outputs at two points, calculate the twenty five correlations $R(x_i - x_j | \theta)$ for $\theta \in \{0.5, 2, 5, 20, 100\}$ and $|x_i - x_j| \in \{0.1, 0.2, 0.4, 0.6, 0.7\}$, where $|x_i - x_j|$ is the absolute value of the distance between x_i and x_j.

   (b) Construct a plot with $|x_i - x_j|$ on the x-axis and $R(x_i - x_j | \theta)$ on the y-axis, plot $R(x_i - x_j | \theta)$ for each value of θ on the same plot, and comment on the relationship between θ and $R(x_i - x_j | \theta)$.

2. **Power exponential correlation function**

   The Gaussian correlation function (20.3.3) is a special case of a Power exponential correlation function. The latter is given by

   $$R(\boldsymbol{x}_i - \boldsymbol{x}_j | \boldsymbol{\xi}) = \prod_{k=1}^{d} \exp(-\theta_k |x_{ik} - x_{jk}|^{p_k})$$

   where $\boldsymbol{\xi}$ represents the parameters $(\theta_1, \ldots, \theta_d, p_1, \ldots, p_d)$, with all $\theta_k \geq 0$, and $0 < p_k \leq 2$.

   (a) Suppose there is $d = 1$ input variable. To investigate the effect of θ on the correlation between outputs at two points x_i and x_j for the Power exponential correlation function, calculate the nine

correlations $R(x_i - x_j | \theta)$ for $\theta \in \{0.5, 5, 100\}$ and $|x_i - x_j| \in \{0.1, 0.4, 0.7\}$, for each value of $p \in \{0.5, 1, 1.5, 2\}$.

(b) Construct four plots, one for each value of p, with $|x_i - x_j|$ on the x-axis and $R(x_i - x_j | \theta)$ on the y-axis. Plot $R(x_i - x_j | \theta)$ for each value of θ on the same plot, and comment on the relationship between θ and $R(x_i - x_j | \theta)$ for each value of p.

3. **Cubic correlation function**

The cubic correlation function is given by

$$R(\boldsymbol{x}_i - \boldsymbol{x}_j | \boldsymbol{\xi}) = \prod_{k=1}^{d} R(x_{ik} - x_{jk} | \boldsymbol{\xi})$$

where

$$R(x_{ik} - x_{jk} | \boldsymbol{\xi}) = \begin{cases} 1 - 6 \left(\frac{x_{ik} - x_{jk}}{\theta_k} \right)^2 + 6 \left(\frac{|x_{ik} - x_{jk}|}{\theta_k} \right)^3 & \text{if } |x_{ik} - x_{jk}| \leq \frac{\theta_k}{2} \\ 2 \left[1 - \left(\frac{|x_{ik} - x_{jk}|}{\theta_k} \right)^3 \right] & \text{if } \frac{\theta_k}{2} < |x_{ik} - x_{jk}| \leq \theta_k \\ 0 & \text{if } \theta_k < |x_{ik} - x_{jk}| \end{cases} ,$$

where $\boldsymbol{\xi}$ represents the parameters $(\theta_1, \theta_2 \ldots, \theta_d)$, and $\theta_k > 0$. Repeat Exercise 1 with the cubic correlation function and $d = 1$.

4. **Bohman correlation function**

The Bohman correlation function is given by

$$R(\boldsymbol{x}_i - \boldsymbol{x}_j | \boldsymbol{\xi}) = \prod_{k=1}^{d} R(x_{ik} - x_{jk} | \boldsymbol{\xi})$$

where $R(x_{ik} - x_{jk} | \boldsymbol{\xi})$ is given by

$$\begin{cases} 1 - \frac{x_{ik} - x_{jk}}{\theta_k} \cos \left(\frac{\pi(x_{ik} - x_{jk})}{\theta_k} \right) + \frac{1}{\pi} \sin \left(\frac{\pi(x_{ik} - x_{jk})}{\theta_k} \right) & \text{if } |x_{ik} - x_{jk}| < \theta_k \\ 0 & \text{if } \theta_k \leq |x_{ik} - x_{jk}| \end{cases} ,$$

and $\boldsymbol{\xi}$ represents the parameters $(\theta_1, \theta_2 \ldots, \theta_d)$, and $\theta_k > 0$. Repeat Exercise 1 with the Bohman correlation function and $d = 1$.

5. **Euclidean interpoint distance**

For two points $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ and $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jd})$ in a d-dimensional space, the Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined by (20.4.8), p. 775. For the questions below assume that the range of each input variable is [0, 1].

(a) What is the largest possible distance between two points in (i) a 1-dimensional space? (ii) a 2-dimensional space? (iii) a d-dimensional space?

(b) What is the smallest possible distance between two points in a d-dimensional space? Why would it be undesirable to have two input points with this minimum distance between them in a computer experiment?

6. **Maximin Latin hypercube designs**

Consider the three LHDs shown below. Note that the location of the points is not in the center of each cell but is randomly chosen.

$$X_1 = \begin{bmatrix} 0.66 & 0.65 \\ 0.23 & 0.38 \\ 0.78 & 0.21 \\ 0.38 & 0.98 \end{bmatrix}, \; X_2 = \begin{bmatrix} 0.50 & 0.68 \\ 0.24 & 0.39 \\ 0.89 & 0.19 \\ 0.35 & 0.92 \end{bmatrix}, \text{ and } X_3 = \begin{bmatrix} 0.98 & 0.10 \\ 0.39 & 0.54 \\ 0.18 & 0.76 \\ 0.57 & 0.39 \end{bmatrix}$$

We can think of *maximin designs* in the following way. Suppose we need to place n grocery stores (design points, corresponding to rows in X) in a particular county, where the county is the experimental region, taken as a rectangle determined by the ranges of the input variables x_i and then scaled to $[0, 1]^2$. We would like to choose the store locations in a way that prevents any two stores from being close together. In other words, we are maximizing the minimum distance between the stores and are constructing a maximin design.

(a) For each design, calculate all $\binom{4}{2}$ Euclidean interpoint distances (20.4.8), p. 775, i.e. calculate the distances between each pair of proposed store locations.
(b) For each design, identify the two closest points and their distance.
(c) Between X_1, X_2, and X_3 identify the design that maximizes the minimum interpoint distance; i.e. identify the maximin design.

7. **Minimax designs**

Referring back to the intuitive explanation of the maximin designs in Exercise 6, consider now the distance between each customer and the location of the stores in a county. A reasonable placement of the stores could be such that no customer is too far from the closest store. In other words, we are minimizing the maximum distance of each customer to the closest store. When we are minimizing the maximum distance of any point in the experimental region (input space) from the closest design point, we are constructing a *minimax design.*

Minimax designs are notoriously difficult to construct. When building a maximin design we have to calculate only the $\binom{n}{2}$ distances among the design points but, when constructing a minimax design, we have to consider infinitely many distances (since there are infinitely many points in the experimental region).

Consider the three LHDs from Exercise 6 (i.e. the potential grocery store locations) and suppose that we are interested in the points in the experimental region (i.e. the customer locations) given by

$$C = \begin{bmatrix} 0.59 & 0.10 \\ 0.89 & 0.55 \\ 0.14 & 0.35 \\ 0.38 & 0.90 \\ 0.72 & 0.63 \end{bmatrix}.$$

(a) For each of five points (customer locations) in C, determine the closest of the four design points in X_1 by calculating four appropriate Euclidean distances (20.4.8), p. 775. Identify a point (customer) in C with the largest distance to its closest point in X_1.
(b) Repeat part (a) for X_2 and X_3.
(c) Identify which of the three designs is minimax for this scenario.

8. **Minimum average reciprocal distance designs**

   An alternative measure of space-fillingness is that of *minimum average reciprocal distance*. If design points are spaced out, then the distance between any pair of points will not be small, and so their reciprocal distance will not be large. Thus an alternative to constructing a maximin design is to construct a design with smallest possible average reciprocal distance between pairs of design points. The Euclidean distances between the points are calculated as in (20.4.8), p. 775, and the average of their reciprocals is the measure of goodness of the design.

   (a) Of the three designs X_1, X_2, X_3 in Exercise 6, which has the minimum average reciprocal distance? Does this coincide with the maximin design?
   (b) Of the three designs X_1, X_2, X_3 in Example 20.4.2, p. 775, which has the minimum average reciprocal distance? Does this coincide with the maximin design?

9. **Prediction**

   Suppose that a computer simulator with one input variable was run at 3 input points and the GaSP model with Gaussian correlation function (20.3.3) was fitted to the data:

x	y
0.20	−0.3635
0.50	−0.1353
0.80	−0.0330

   leading to parameter estimates $\hat{\beta}_0 = -0.2104$, $\hat{\sigma}^2 = 0.0264$, and $\hat{\theta} = 4.9003$. Based on this information, we would like to predict Y at $x_a = 0.20$ using the predictor in (20.3.4), p. 770.

   (a) Calculate $\hat{R}$ (defined below (20.3.4)).
   (b) Calculate $\hat{r}$ (defined below (20.3.4)).
   (c) Calculate $\hat{Y}(x_a)$ in (20.3.4) at $x_a = 0.20$. Is this the answer you expected? Explain.
   (d) Calculate the estimated variance $\hat{s}^2(x_a)$ (20.3.5) of your prediction. Is this the answer you expected? Explain.
   (e) Repeat parts (b), (c), and (d) for $x_a = 0.49$ and $x_a = 0.65$.

10. **Tool coating experiment**

    D. Draguljić, S. Nekkanty, T. J. Santner, A. M. Dean, and R. Shivpuri, in *Quality Engineering* in 2015, described a computer experiment used to develop multilayer coatings which are used to protect tools, drills, cutting blades, bearings, etc. The experiment consisted of modeling the effect of the number of coating layers and the thicknesses of those layers on the normalized measures of maximum normal radial stress, Y_1, and the maximum shear stress, Y_2. The two responses were modeled independently. Large values of either stress would lead to coating failures (either peeling of of the coating or occurrence of cracks in the coating). Here we will focus on coatings with only two layers. Therefore we have input variables x_1 and x_2 (both in μm), the thicknesses of the top and the bottom layer, respectively. The data for this experiment are given in Table 20.8. Note that x_1 and x_2, as shown in Table 20.8, are scaled from their original (0, 6) μm scale to (0, 1) μm scale.

    (a) Estimate θ_1 and θ_2 from the GaSP model that relates x_1 and x_2 to y_1 using the data from Table 20.8.

Table 20.8 Data for the tool coating experiment

x_1	x_2	y_1	y_2
0.6250	0.0833	0.87771	0.13567
0.8750	0.1250	0.84920	0.12590
0.2083	0.2083	0.53123	0.13394
0.0417	0.4583	0.26112	0.15829
0.3750	0.0833	0.75124	0.13708
0.0417	0.8333	2.56179	0.15464
0.1667	0.5417	1.09043	0.22275
0.3750	0.5000	1.75458	0.19860
0.6250	0.3750	1.40447	0.12491
0.4167	0.2917	1.19420	0.13259

(b) For this experiment the total thickness of the coating was required to be between 1/3 and 1, i.e. $1/3 \leq x_1 + x_2 \leq 1$ while the thickness of any individual layer needed to be a multiple of 1/24. Construct a grid with mesh size of 1/24 that satisfies this constraint and predict the values of Y_1 for this grid.

(c) Based on your predictions, what pair of coating thicknesses (x_1, x_2) seems to minimize Y_1?

(d) Repeat parts (a)–(c) for Y_2.

(e) Based on your predictions for Y_1 and Y_2, what single pair of coating thicknesses (x_1, x_2) would you suggest to try to minimize both Y_1 and Y_2 simultaneously (as well as you can)? It may help to make a plot of the predicted values of Y_1 and Y_2.

11. **Borehole function**

A borehole is a narrow tunnel drilled in the ground. Boreholes serve numerous purposes, including extraction of water or gases, mineral exploration, temperature measurement, etc. The *borehole function* (see Surjanovic and Bingham 2013) models water flow through a borehole and is given by

$$y = \frac{2\pi T_u (H_u - H_l)}{a \left(1 + \frac{2LT_u}{ar_w^2 K_w} + \frac{T_u}{T_l} \right)} \tag{20.7.9}$$

where $a = \ln(r/r_w)$. The output y measures the water flow rate in m^3/year. There are eight inputs to the borehole function. Their names and ranges are given in Table 20.9.

(a) Construct an 80×8 maximin LHD, X, where each element of X is in [0, 1]. Let the elements x_{i1} in the first column of X represent the scaled values of the first variable r_w for which we will "observe" (calculate) y from the simulator, the elements x_{i2} in the second column of X represent the scaled values of the second variable r, and so on for all 8 columns.

(b) To be able to calculate the simulator data $y(x_i)$ using (20.7.9), where x_i is the ith row (input combination) of X, we need to transform the value of each input variable in X (which has range [0, 1]) to the variables and matching ranges given in Table 20.9. This will allow the appropriate values to be entered into (20.7.9). For transforming x_1 to r_w with range in Table 20.9, the scaling is done via

$$r_w^{[0.05, 0.15]} = (0.15 - 0.05)x_1^{[0,1]} + 0.05 \,.$$

Table 20.9 Input variables for the borehole function

Variable	Variable name	Range
r_w	Radius of the borehole (m)	[0.05, 0.15]
r	Radius of influence (m)	[100, 50000]
T_u	Transmissivity of the upper aquifer (m^2/yr)	[63070, 115600]
H_u	Potentiometric head of the upper aquifer (m)	[990, 1110]
T_ℓ	Transmissivity of the lower aquifer (m^2/yr)	[63.1, 116]
H_ℓ	Potentiometric head of the lower aquifer (m)	[700, 820]
L	Length of the borehole (m)	[1120, 1680]
K_w	Hydraulic conductivity of the borehole (m/yr)	[9855, 12045]

The superscripts represent the ranges for the input variable x_1 and transformed variable r_w. The scaling for other variables is done in a similar fashion. Scale each value in each row $\boldsymbol{x}_i$ of X to the appropriate range. Calculate the output $y(\boldsymbol{x}_i)$, $i = 1, 2, \ldots, 80$ using (20.7.9).

(c) One of the simple ways to assess how sensitive Y is to the changes in a particular input x_k is to plot y versus the x_{ik}'s and examine the plot for any obvious patterns. Construct eight scatterplots, one for each input variable x_1–x_8, and identify the variables that seem to influence the response the most.

(d) Using X and $y(\boldsymbol{x}_i)$, fit the GaSP model. What are the maximum likelihood estimates of the parameters? Do the values of $\hat{\theta}_k$, $k = 1, 2, \ldots, 8$, support your conclusion from part (c)?

(e) Construct a grid of mesh size roughly equal to 0.5 (i.e. $x_{ik} \in \{0, 0.5, 1\}$) and predict the outputs $y(\boldsymbol{x}_i)$ over the grid. Based on your predictions, what are the values of $x_1, \ldots, x_8$ that result in the predicted maximal water flow? How about the minimal water flow? Using the ranges in Table 20.9, again, scale these values to show the values of the original variables that result in predicted maximal and minimal water flow.

Contents

Table A.1 Random numbers* (Section 1)

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

*Random numbers were generated using the SAS statements "retain seed 1613126064;" and "rn = floor(10*ranuni(seed));"

Table A.1   Random numbers (Section 2)

772357	993916	529577	678243	561863	847933
432529	970061	557500	798091	179138	481792
307505	525070	258159	554655	575046	678559
466350	487907	299729	248259	480128	346994
617303	468101	318577	923429	038402	757219
593585	243209	814395	862822	746040	791440
001642	339004	160458	600772	662567	368585
149273	042555	774276	484976	819195	397235
650084	170717	096777	646438	049121	435982
291696	199860	255968	977965	704832	152207
380204	676820	172541	838781	875496	866057
601789	268945	637114	763853	972534	321807
034840	046770	352998	091564	836769	679665
021222	269557	705784	227686	602822	502110
657269	645233	803339	576422	510363	650292
496756	485525	641125	398402	351104	116450
837444	628196	558438	029761	119038	624675
488075	739348	349563	070527	096075	150099
704658	615314	955986	559666	475521	954801
102178	988809	827881	636004	439186	939951
693252	078895	586050	545691	461858	159346
252611	441081	786315	281005	010070	205751
280273	117608	080447	202151	108901	325238
453702	029011	657141	620700	604050	013203
327097	317434	109367	377563	036065	203413
786141	861409	263375	518852	087913	861715
194910	643115	993964	022892	046775	061654
875988	516714	699665	303752	334663	145398
370225	083754	123844	056596	275830	997716
060627	822010	449146	256578	444509	074713
536122	366557	689934	795977	120966	293430
413614	370683	263343	215245	320757	951902
722209	875974	091346	411939	116215	304007
661019	124710	939043	865547	430354	627979
631837	027762	468015	102214	231263	950303
751081	600240	089651	253141	586782	632880

Table A.1 Random numbers (Section 3)

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

Table A.1 Random numbers (Section 4)

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

Table A.1 Random numbers (Section 5)

178620	984332	234149	062802	203306	409996
857451	696719	245360	219945	554014	514398
363613	427959	500852	392452	867274	977098
934682	279698	307888	039229	143251	417387
355775	341604	728003	783298	021655	069731
665699	212463	461254	662458	822056	868018
490532	385220	245113	034522	661336	735326
177715	678060	444410	984458	394908	203395
247042	584562	285965	331489	536429	150486
247037	647390	958757	889429	320390	733635
037684	319850	319453	708172	600706	116880
163936	682211	276583	695090	141138	775518
936059	110744	230502	313261	037153	962437
723498	531743	968071	809740	728333	510996
260338	879915	375157	296812	522971	126148
224336	392250	500589	802202	461732	596494
983601	286360	962976	318965	952288	371597
950090	003498	944526	890667	177182	557799
517339	436412	783049	641166	652511	767894
192027	132666	760861	620786	271242	132728
503888	846019	856424	233882	942325	511328
796439	106427	930753	519367	261714	969157
354978	019946	325140	390753	674999	265593
881853	487399	093950	592023	463922	642370
834658	788932	959353	528824	971467	890240
429335	115816	387007	698632	332875	685767
502244	063360	540982	294989	737673	576578
230094	859807	408288	251055	307629	402631
108970	046489	583823	027768	600042	788227
985730	336414	654852	476272	011614	049166
797389	447994	431552	294961	293555	180306
967876	584097	346493	644527	708534	540623
488214	736136	666810	990017	550198	332353
594677	673425	481734	749014	380123	200747
216666	893249	173621	966579	934953	101352
829921	075112	978105	427576	707853	535755

Table A.1 Random numbers (Section 6)

797468	484252	023853	192278	937112	102503
869406	106834	234802	302737	222508	069132
084703	456756	742761	693385	713106	704798
953014	558897	076970	286515	902394	008147
068334	355648	814540	713770	474093	866359
291019	695796	692442	099917	354198	854868
829045	784428	553248	413099	160541	531210
347351	078917	067538	591284	925325	240708
961281	639335	808637	097914	370752	132493
237408	602057	784998	569469	812566	754613
949564	679405	303591	345128	074802	919895
611262	590822	844714	210655	727420	947239
382111	838410	235673	240015	226367	794195
017136	100540	728145	640255	869932	716829
832572	001164	362585	877812	958998	300684
559820	106016	954619	384380	539268	309304
567125	265846	256533	543082	341655	423495
135117	247924	017309	392054	724319	938577
491976	446939	602958	446781	684067	223566
031030	690179	893112	500077	640800	683948
821529	204137	910310	756711	114949	579204
889160	294936	024169	683546	518509	569679
089569	665239	349069	158437	093839	659940
713873	707840	086598	612305	446369	210679
436460	508406	967487	781618	226710	547386
613674	173160	898727	687279	638647	536756
366699	338577	843373	151717	430829	944266
828010	825546	202510	695671	505800	680682
340755	280452	630101	428665	695421	280009
358767	365611	845098	195830	471533	344402
495054	142723	141577	566053	231325	044934
495165	641454	169446	498379	947861	680617
326075	085211	962907	515048	074505	256236
897734	334967	269056	297336	570543	835922
039744	115260	301407	795887	731089	488919
279055	728952	717435	600372	561076	481948

Table A.2 Coefficients c_i for orthogonal polynomial trend contrasts

$v = 3$			
Trend	c_1	c_2	c_3
Linear	-1	0	1
Quadratic	1	-2	1

$v = 4$				
Trend	c_1	c_2	c_3	c_4
Linear	-3	-1	1	3
Quadratic	1	-1	-1	1
Cubic	-1	3	-3	1

$v = 5$					
Trend	c_1	c_2	c_3	c_4	c_5
Linear	-2	-1	0	1	2
Quadratic	2	-1	-2	-1	2
Cubic	-1	2	0	-2	1
Quartic	1	-4	6	-4	1

$v = 6$						
Trend	c_1	c_2	c_3	c_4	c_5	c_6
Linear	-5	-3	-1	1	3	5
Quadratic	5	-1	-4	-4	-1	5
Cubic	-5	7	4	-4	-7	5
Quartic	1	-3	2	2	-3	1
Quintic	-1	5	-10	10	-5	1

$v = 7$							
Trend	c_1	c_2	c_3	c_4	c_5	c_6	c_7
Linear	-3	-2	-1	0	1	2	3
Quadratic	5	0	-3	-4	-3	0	5
Cubic	-1	1	1	0	-1	-1	1
Quartic	3	-7	1	6	1	-7	3
Quintic	-1	4	-5	0	5	-4	1
Sextic	1	-6	15	-20	15	-6	1

Table A.3 Standard normal distribution:[*] upper α critical coefficients, z_α, and upper-tail probabilities, $\alpha = P(Z > z_\alpha)$

α	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005	0.00025	0.0001
z_α	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291	3.481	3.719
z_α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005

[*]Values were generated using the SAS statements "z_alpha = probit(1-alpha);" and "alpha = 1 - probnorm(z_alpha);"

Table A.4 Student's t-distribution:* upper α critical coefficients, $t_{df,\alpha}$, where $\alpha = P(t_{df} > t_{df,\alpha})$

df	α								
	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005	0.0001
1	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6	3183
2	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60	70.70
3	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92	22.20
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610	13.03
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869	9.678
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959	8.025
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408	7.063
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041	6.442
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781	6.010
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587	5.694
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437	5.453
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318	5.263
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221	5.111
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140	4.985
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073	4.880
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015	4.791
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965	4.714
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922	4.648
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883	4.590
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850	4.539
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819	4.493
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792	4.452
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768	4.415
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745	4.382
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725	4.352
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707	4.324
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690	4.299
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674	4.275
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659	4.254
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646	4.234
35	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591	4.153
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551	4.094
45	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520	4.049
50	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496	4.014
55	1.297	1.673	2.004	2.396	2.668	2.925	3.245	3.476	3.986
60	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460	3.962
70	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435	3.926
80	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416	3.899
90	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402	3.878
100	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390	3.862
110	1.289	1.659	1.982	2.361	2.621	2.865	3.166	3.381	3.848
120	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373	3.837
∞	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291	3.719

*Values $t_{df,\alpha}$ were generated using the SAS statements "t = tinv(1-alpha,df);" for $df < \infty$ and "t = probit(1-alpha)" for $df = \infty$

Table A.5 Chi-squared distribution:* upper α critical coefficients, $\chi^2_{df,\alpha}$, where $\alpha = P(\chi^2_{df} > \chi^2_{df,\alpha})$

df	α									
	0.999	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.001
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	10.83
2	0.002	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	13.82
3	0.024	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34	16.27
4	0.091	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28	18.47
5	0.210	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09	20.52
6	0.381	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81	22.46
7	0.598	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	24.32
8	0.857	1.646	2.180	2.733	3.490	13.36	15.51	17.53	20.09	26.12
9	1.152	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67	27.88
10	1.479	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21	29.59
11	1.834	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72	31.26
12	2.214	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22	32.91
13	2.617	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69	34.53
14	3.041	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14	36.12
15	3.483	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58	37.70
16	3.942	5.812	6.908	7.962	9.312	23.54	26.30	28.85	32.00	39.25
17	4.416	6.408	7.564	8.672	10.09	24.77	27.59	30.19	33.41	40.79
18	4.905	7.015	8.231	9.390	10.86	25.99	28.87	31.53	34.81	42.31
19	5.407	7.633	8.907	10.12	11.65	27.20	30.14	32.85	36.19	43.82
20	5.921	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57	45.31
21	6.447	8.897	10.28	11.59	13.24	29.62	32.67	35.48	38.93	46.80
22	6.983	9.542	10.98	12.34	14.04	30.81	33.92	36.78	40.29	48.27
23	7.529	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	49.73
24	8.085	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	51.18
25	8.649	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	52.62
26	9.222	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	54.05
27	9.803	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	55.48
28	10.39	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	56.89
29	10.99	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	58.30
30	11.59	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	59.70
35	14.69	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34	66.62
40	17.92	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	73.40
45	21.25	25.90	28.37	30.61	33.35	57.51	61.66	65.41	69.96	80.08
50	24.67	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	86.66
55	28.17	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29	93.17
60	31.74	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	99.61
70	39.04	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	112.3
80	46.52	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	124.8
90	54.16	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	137.2
100	61.92	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	149.4
120	77.76	86.92	91.57	95.70	100.6	140.2	146.6	152.2	159.0	173.6

*Values were generated using the SAS statement "chi2 = cinv(1-alpha,df);"

Table A.6 F-distribution:* upper α critical coefficients, $F_{\nu_1,\nu_2,\alpha}$, where $\alpha = P(F_{\nu_1,\nu_2} > F_{\nu_1,\nu_2,\alpha})$

ν_2	α	ν_1								
		1	2	3	4	5	6	7	8	9
1	0.100	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9
	0.050	161	200	216	225	230	234	237	239	241
	0.010	4052	5000	5403	5625	5764	5859	5928	5981	6022
2	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	0.050	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4
	0.010	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4
	0.001	999	999	999	999	999	999	999	999	999
3	0.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	0.050	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	0.010	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.4
	0.001	167	149	141	137	135	133	132	131	130
4	0.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	0.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	0.010	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7
	0.001	74.1	61.3	56.2	53.4	51.7	50.5	49.7	49.0	48.5
5	0.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	0.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	0.010	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2
	0.001	47.2	37.1	33.2	31.1	29.8	28.8	28.2	27.7	27.2
6	0.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	0.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	0.010	13.8	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	0.001	35.5	27.0	23.7	21.9	20.8	20.0	19.5	19.0	18.7
7	0.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	0.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	0.010	12.3	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	0.001	29.3	21.7	18.8	17.2	16.2	15.5	15.0	14.6	14.3
8	0.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	0.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	0.010	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	0.001	25.4	18.5	15.8	14.4	13.5	12.9	12.4	12.1	11.8
9	0.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	0.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	0.010	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	0.001	22.9	16.4	13.9	12.6	11.7	11.1	10.7	10.4	10.1
10	0.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	0.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	0.010	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	0.001	21.0	14.9	12.6	11.3	10.5	9.93	9.52	9.20	8.96

(continued)

Table A.6 (F-distribution, continued)

ν_2	α	ν_1								
		1	2	3	4	5	6	7	8	9
11	0.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	0.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	0.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	0.001	19.7	13.8	11.6	10.4	9.58	9.05	8.66	8.35	8.12
12	0.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	0.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	0.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	0.001	18.6	13.0	10.8	9.63	8.89	8.38	8.00	7.71	7.48
13	0.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	0.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	0.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	0.001	17.8	12.3	10.2	9.07	8.35	7.86	7.49	7.21	6.98
14	0.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	0.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	0.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	0.001	17.1	11.8	9.73	8.62	7.92	7.44	7.08	6.80	6.58
15	0.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	0.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	0.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	0.001	16.6	11.3	9.34	8.25	7.57	7.09	6.74	6.47	6.26
16	0.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	0.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	0.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	0.001	16.1	11.0	9.01	7.94	7.27	6.80	6.46	6.19	5.98
17	0.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	0.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	0.010	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
	0.001	15.7	10.7	8.73	7.68	7.02	6.56	6.22	5.96	5.75
18	0.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	0.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	0.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	0.001	15.4	10.4	8.49	7.46	6.81	6.35	6.02	5.76	5.56
19	0.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	0.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	0.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	0.001	15.1	10.2	8.28	7.27	6.62	6.18	5.85	5.59	5.39
20	0.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	0.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	0.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	0.001	14.8	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24

(continued)

Table A.6 (F-distribution, continued)

ν_2	α	ν_1								
		1	2	3	4	5	6	7	8	9
22	0.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	0.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	0.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	0.001	14.4	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99
25	0.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	0.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	0.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	0.001	13.9	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71
30	0.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	0.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	0.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	0.001	13.3	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39
35	0.100	2.85	2.46	2.25	2.11	2.02	1.95	1.90	1.85	1.82
	0.050	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16
	0.010	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96
	0.001	12.9	8.47	6.79	5.88	5.30	4.89	4.59	4.36	4.18
40	0.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	0.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
	0.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	0.001	12.6	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02
60	0.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	0.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	0.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	0.001	12.0	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
80	0.100	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71
	0.050	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00
	0.010	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64
	0.001	11.7	7.54	5.97	5.12	4.58	4.20	3.92	3.70	3.53
100	0.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69
	0.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
	0.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
	0.001	11.5	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44
120	0.100	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68
	0.050	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
	0.010	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
	0.001	11.4	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38
1000	0.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64
	0.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
	0.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
	0.001	10.9	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13

(continued)

Table A.6 (F-distribution, continued)

ν_2	α	ν_1									
		10	12	15	20	25	30	40	60	120	1000
1	0.100	60.2	60.7	61.2	61.7	62.1	62.3	62.5	62.8	63.1	63.3
	0.050	242	244	246	248	249	250	251	252	253	254
	0.010	6056	6106	6157	6209	6240	6261	6287	6313	6339	6363
2	0.100	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
	0.050	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5
	0.010	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5
	0.001	999	999	999	999	999	999	999	999	999	1000
3	0.100	5.23	5.22	5.20	5.18	5.17	5.17	5.16	5.15	5.14	5.13
	0.050	8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.57	8.55	8.53
	0.010	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
	0.001	129	128	127	126	126	125	125	124	124	124
4	0.100	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
	0.050	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	0.010	14.6	14.4	14.2	14.0	13.9	13.8	13.8	13.7	13.6	13.5
	0.001	48.1	47.4	46.8	46.1	45.7	45.4	45.1	44.8	44.4	44.1
5	0.100	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.11
	0.050	4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.43	4.40	4.37
	0.010	10.1	9.89	9.72	9.55	9.45	9.38	9.29	9.20	9.11	9.03
	0.001	26.9	26.4	25.9	25.4	25.1	24.9	24.6	24.3	24.1	23.8
6	0.100	2.94	2.90	2.87	2.84	2.81	2.80	2.78	2.76	2.74	2.72
	0.050	4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.74	3.70	3.67
	0.010	7.87	7.72	7.56	7.40	7.30	7.23	7.14	7.06	6.97	6.89
	0.001	18.4	18.0	17.6	17.1	16.9	16.7	16.4	16.2	16.0	15.8
7	0.100	2.70	2.67	2.63	2.59	2.57	2.56	2.54	2.51	2.49	2.47
	0.050	3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.30	3.27	3.23
	0.010	6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.82	5.74	5.66
	0.001	14.1	13.7	13.3	12.9	12.7	12.5	12.3	12.1	11.9	11.7
8	0.100	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.30
	0.050	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.01	2.97	2.93
	0.010	5.81	5.67	5.52	5.36	5.26	5.20	5.12	5.03	4.95	4.87
	0.001	11.5	11.2	10.8	10.5	10.3	10.1	9.92	9.73	9.53	9.36
9	0.100	2.42	2.38	2.34	2.30	2.27	2.25	2.23	2.21	2.18	2.16
	0.050	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.79	2.75	2.71
	0.010	5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.48	4.40	4.32
	0.001	9.89	9.57	9.24	8.90	8.69	8.55	8.37	8.19	8.00	7.84
10	0.100	2.32	2.28	2.24	2.20	2.17	2.16	2.13	2.11	2.08	2.06
	0.050	2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.62	2.58	2.54
	0.010	4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.08	4.00	3.92
	0.001	8.75	8.45	8.13	7.80	7.60	7.47	7.30	7.12	6.94	6.78

(continued)

Table A.6 (F-distribution, continued)

ν_2	α	ν_1									
		10	12	15	20	25	30	40	60	120	1000
11	0.100	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.98
	0.050	2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.49	2.45	2.41
	0.010	4.54	4.40	4.25	4.10	4.01	3.94	3.86	3.78	3.69	3.61
	0.001	7.92	7.63	7.32	7.01	6.81	6.68	6.52	6.35	6.18	6.02
12	0.100	2.19	2.15	2.10	2.06	2.03	2.01	1.99	1.96	1.93	1.91
	0.050	2.75	2.69	2.62	2.54	2.50	2.47	2.43	2.38	2.34	2.30
	0.010	4.30	4.16	4.01	3.86	3.76	3.70	3.62	3.54	3.45	3.37
	0.001	7.29	7.00	6.71	6.40	6.22	6.09	5.93	5.76	5.59	5.44
13	0.100	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
	0.050	2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.30	2.25	2.21
	0.010	4.10	3.96	3.82	3.66	3.57	3.51	3.43	3.34	3.25	3.18
	0.001	6.80	6.52	6.23	5.93	5.75	5.63	5.47	5.30	5.14	4.99
14	0.100	2.10	2.05	2.01	1.96	1.93	1.91	1.89	1.86	1.83	1.80
	0.050	2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.22	2.18	2.14
	0.010	3.94	3.80	3.66	3.51	3.41	3.35	3.27	3.18	3.09	3.02
	0.001	6.40	6.13	5.85	5.56	5.38	5.25	5.10	4.94	4.77	4.62
15	0.100	2.06	2.02	1.97	1.92	1.89	1.87	1.85	1.82	1.79	1.76
	0.050	2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.16	2.11	2.07
	0.010	3.80	3.67	3.52	3.37	3.28	3.21	3.13	3.05	2.96	2.88
	0.001	6.08	5.81	5.54	5.25	5.07	4.95	4.80	4.64	4.47	4.33
16	0.100	2.03	1.99	1.94	1.89	1.86	1.84	1.81	1.78	1.75	1.72
	0.050	2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.11	2.06	2.02
	0.010	3.69	3.55	3.41	3.26	3.16	3.10	3.02	2.93	2.84	2.76
	0.001	5.81	5.55	5.27	4.99	4.82	4.70	4.54	4.39	4.23	4.08
17	0.100	2.00	1.96	1.91	1.86	1.83	1.81	1.78	1.75	1.72	1.69
	0.050	2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.06	2.01	1.97
	0.010	3.59	3.46	3.31	3.16	3.07	3.00	2.92	2.83	2.75	2.66
	0.001	5.58	5.32	5.05	4.78	4.60	4.48	4.33	4.18	4.02	3.87
18	0.100	1.98	1.93	1.89	1.84	1.80	1.78	1.75	1.72	1.69	1.66
	0.050	2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.02	1.97	1.92
	0.010	3.51	3.37	3.23	3.08	2.98	2.92	2.84	2.75	2.66	2.58
	0.001	5.39	5.13	4.87	4.59	4.42	4.30	4.15	4.00	3.84	3.69
19	0.100	1.96	1.91	1.86	1.81	1.78	1.76	1.73	1.70	1.67	1.64
	0.050	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
	0.010	3.43	3.30	3.15	3.00	2.91	2.84	2.76	2.67	2.58	2.50
	0.001	5.22	4.97	4.70	4.43	4.26	4.14	3.99	3.84	3.68	3.53
20	0.100	1.94	1.89	1.84	1.79	1.76	1.74	1.71	1.68	1.64	1.61
	0.050	2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.95	1.90	1.85
	0.010	3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.61	2.52	2.43
	0.001	5.08	4.82	4.56	4.29	4.12	4.00	3.86	3.70	3.54	3.40

(continued)

Table A.6 (F-distribution, continued)

v_2	α	v_1									
		10	12	15	20	25	30	40	60	120	1000
22	0.100	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
	0.050	2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.89	1.84	1.79
	0.010	3.26	3.12	2.98	2.83	2.73	2.67	2.58	2.50	2.40	2.32
	0.001	4.83	4.58	4.33	4.06	3.89	3.78	3.63	3.48	3.32	3.17
25	0.100	1.87	1.82	1.77	1.72	1.68	1.66	1.63	1.59	1.56	1.52
	0.050	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.72
	0.010	3.13	2.99	2.85	2.70	2.60	2.54	2.45	2.36	2.27	2.18
	0.001	4.56	4.31	4.06	3.79	3.63	3.52	3.37	3.22	3.06	2.91
30	0.100	1.82	1.77	1.72	1.67	1.63	1.61	1.57	1.54	1.50	1.46
	0.050	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.74	1.68	1.63
	0.010	2.98	2.84	2.70	2.55	2.45	2.39	2.30	2.21	2.11	2.02
	0.001	4.24	4.00	3.75	3.49	3.33	3.22	3.07	2.92	2.76	2.61
35	0.100	1.79	1.74	1.69	1.63	1.60	1.57	1.53	1.50	1.46	1.42
	0.050	2.11	2.04	1.96	1.88	1.82	1.79	1.74	1.68	1.62	1.57
	0.010	2.88	2.74	2.60	2.44	2.35	2.28	2.19	2.10	2.00	1.90
	0.001	4.03	3.79	3.55	3.29	3.13	3.02	2.87	2.72	2.56	2.40
40	0.100	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
	0.050	2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.64	1.58	1.52
	0.010	2.80	2.66	2.52	2.37	2.27	2.20	2.11	2.02	1.92	1.82
	0.001	3.87	3.64	3.40	3.14	2.98	2.87	2.73	2.57	2.41	2.25
60	0.100	1.71	1.66	1.60	1.54	1.50	1.48	1.44	1.40	1.35	1.30
	0.050	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.53	1.47	1.40
	0.010	2.63	2.50	2.35	2.20	2.10	2.03	1.94	1.84	1.73	1.62
	0.001	3.54	3.32	3.08	2.83	2.67	2.55	2.41	2.25	2.08	1.92
80	0.100	1.68	1.63	1.57	1.51	1.47	1.44	1.40	1.36	1.31	1.25
	0.050	1.95	1.88	1.79	1.70	1.64	1.60	1.54	1.48	1.41	1.34
	0.010	2.55	2.42	2.27	2.12	2.01	1.94	1.85	1.75	1.63	1.51
	0.001	3.39	3.16	2.93	2.68	2.52	2.41	2.26	2.10	1.92	1.75
100	0.100	1.66	1.61	1.56	1.49	1.45	1.42	1.38	1.34	1.28	1.22
	0.050	1.93	1.85	1.77	1.68	1.62	1.57	1.52	1.45	1.38	1.30
	0.010	2.50	2.37	2.22	2.07	1.97	1.89	1.80	1.69	1.57	1.45
	0.001	3.30	3.07	2.84	2.59	2.43	2.32	2.17	2.01	1.83	1.64
120	0.100	1.65	1.60	1.55	1.48	1.44	1.41	1.37	1.32	1.26	1.20
	0.050	1.91	1.83	1.75	1.66	1.60	1.55	1.50	1.43	1.35	1.27
	0.010	2.47	2.34	2.19	2.03	1.93	1.86	1.76	1.66	1.53	1.40
	0.001	3.24	3.02	2.78	2.53	2.37	2.26	2.11	1.95	1.77	1.57
1000	0.100	1.61	1.55	1.49	1.43	1.38	1.35	1.30	1.25	1.18	1.08
	0.050	1.84	1.76	1.68	1.58	1.52	1.47	1.41	1.33	1.24	1.11
	0.010	2.34	2.20	2.06	1.90	1.79	1.72	1.61	1.50	1.35	1.16
	0.001	2.99	2.77	2.54	2.30	2.14	2.02	1.87	1.69	1.49	1.22

*Values $F_{v_1, v_2, \alpha}$ were generated using the SAS statement "f = round(finv(1-alpha,df1,df2),0.01);"

Table A.7 Power of the F-test: $\pi(\phi) = P(F_{v_1,v_2,\phi} > F_{v_1,v_2,\alpha})$

$v_1 = 1, \alpha = 0.05$

v_2	ϕ										
	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
5	0.211	0.301	0.405	0.515	0.623	0.721	0.805	0.870	0.919	0.952	0.973
6	0.223	0.320	0.430	0.546	0.657	0.755	0.836	0.897	0.939	0.966	0.982
7	0.232	0.333	0.449	0.568	0.681	0.779	0.856	0.913	0.951	0.974	0.987
8	0.239	0.344	0.463	0.585	0.698	0.795	0.871	0.924	0.959	0.979	0.990
9	0.245	0.352	0.474	0.598	0.712	0.808	0.881	0.932	0.964	0.982	0.992
10	0.249	0.359	0.483	0.608	0.722	0.817	0.889	0.938	0.968	0.985	0.993
12	0.256	0.370	0.496	0.623	0.738	0.831	0.900	0.946	0.973	0.988	0.995
15	0.263	0.380	0.510	0.639	0.753	0.844	0.910	0.953	0.977	0.990	0.996
20	0.270	0.391	0.524	0.654	0.767	0.857	0.919	0.959	0.981	0.992	0.997
30	0.278	0.402	0.537	0.668	0.781	0.868	0.928	0.964	0.984	0.994	0.998
60	0.285	0.413	0.551	0.683	0.795	0.879	0.936	0.969	0.987	0.995	0.998
120	0.289	0.418	0.557	0.690	0.801	0.884	0.939	0.971	0.988	0.995	0.998
1000	0.293	0.423	0.563	0.696	0.807	0.889	0.942	0.973	0.989	0.996	0.999

$v_1 = 2, \alpha = 0.05$

v_2	ϕ										
	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
5	0.196	0.283	0.386	0.499	0.611	0.714	0.802	0.870	0.920	0.954	0.975
6	0.211	0.307	0.421	0.543	0.660	0.764	0.847	0.907	0.948	0.973	0.987
7	0.223	0.327	0.449	0.576	0.696	0.798	0.876	0.930	0.963	0.983	0.992
8	0.233	0.343	0.470	0.602	0.723	0.822	0.896	0.944	0.973	0.988	0.995
9	0.241	0.356	0.487	0.622	0.743	0.840	0.910	0.954	0.979	0.991	0.997
10	0.248	0.366	0.502	0.638	0.759	0.854	0.920	0.961	0.983	0.993	0.998
12	0.258	0.383	0.523	0.662	0.783	0.874	0.934	0.969	0.987	0.995	0.998
15	0.270	0.400	0.546	0.687	0.805	0.892	0.947	0.977	0.991	0.997	0.999
20	0.282	0.418	0.569	0.711	0.827	0.908	0.957	0.982	0.994	0.998	0.999
30	0.294	0.437	0.592	0.735	0.847	0.923	0.966	0.987	0.996	0.999	1.000
60	0.308	0.457	0.615	0.758	0.866	0.935	0.973	0.990	0.997	0.999	1.000
120	0.314	0.467	0.627	0.769	0.875	0.941	0.976	0.992	0.998	0.999	1.000
1000	0.321	0.476	0.637	0.778	0.882	0.946	0.979	0.993	0.998	1.000	1.000

$v_1 = 3, \alpha = 0.05$

v_2	ϕ										
	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
5	0.191	0.279	0.384	0.499	0.614	0.719	0.808	0.876	0.925	0.958	0.978
6	0.209	0.308	0.426	0.552	0.674	0.779	0.861	0.919	0.956	0.978	0.990
7	0.224	0.332	0.460	0.593	0.717	0.819	0.894	0.944	0.973	0.988	0.995
8	0.236	0.352	0.487	0.625	0.750	0.848	0.916	0.959	0.982	0.993	0.997
9	0.246	0.368	0.509	0.651	0.775	0.869	0.932	0.968	0.987	0.995	0.998
10	0.255	0.382	0.528	0.671	0.794	0.885	0.943	0.975	0.990	0.997	0.999
12	0.269	0.404	0.556	0.703	0.822	0.907	0.957	0.983	0.994	0.998	1.000
15	0.284	0.428	0.586	0.734	0.849	0.926	0.968	0.988	0.996	0.999	1.000
20	0.300	0.453	0.617	0.764	0.874	0.942	0.978	0.993	0.998	1.000	1.000
30	0.318	0.479	0.648	0.794	0.897	0.956	0.985	0.995	0.999	1.000	1.000
60	0.338	0.507	0.680	0.822	0.917	0.968	0.990	0.997	0.999	1.000	1.000
120	0.348	0.522	0.696	0.835	0.926	0.972	0.992	0.998	1.000	1.000	1.000
1000	0.357	0.535	0.709	0.847	0.933	0.976	0.993	0.998	1.000	1.000	1.000

(continued)

Table A.7 (Power, continued)

$v_1 = 4, \alpha = 0.05$

					ϕ						
v_2	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
5	0.190	0.278	0.385	0.502	0.619	0.726	0.814	0.882	0.930	0.961	0.980
6	0.209	0.311	0.433	0.562	0.686	0.791	0.872	0.928	0.962	0.982	0.992
7	0.226	0.338	0.471	0.609	0.735	0.836	0.908	0.953	0.978	0.991	0.997
8	0.240	0.361	0.503	0.646	0.771	0.867	0.931	0.968	0.987	0.995	0.998
9	0.252	0.381	0.529	0.675	0.799	0.890	0.946	0.977	0.991	0.997	0.999
10	0.262	0.398	0.551	0.699	0.821	0.906	0.957	0.983	0.994	0.998	1.000
12	0.279	0.424	0.586	0.736	0.852	0.929	0.970	0.990	0.997	0.999	1.000
15	0.298	0.454	0.622	0.771	0.881	0.948	0.981	0.994	0.998	1.000	1.000
20	0.319	0.485	0.659	0.806	0.907	0.963	0.988	0.997	0.999	1.000	1.000
30	0.342	0.519	0.697	0.840	0.930	0.975	0.993	0.998	1.000	1.000	1.000
60	0.368	0.555	0.735	0.870	0.949	0.984	0.996	0.999	1.000	1.000	1.000
120	0.382	0.574	0.754	0.885	0.957	0.987	0.997	1.000	1.000	1.000	1.000
1000	0.394	0.591	0.771	0.896	0.963	0.990	0.998	1.000	1.000	1.000	1.000

$v_1 = 5, \alpha = 0.05$

					ϕ						
v_2	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
5	0.189	0.278	0.387	0.506	0.624	0.731	0.820	0.887	0.934	0.964	0.981
6	0.210	0.314	0.438	0.571	0.696	0.801	0.880	0.934	0.967	0.985	0.994
7	0.228	0.344	0.481	0.622	0.749	0.849	0.918	0.960	0.982	0.993	0.998
8	0.244	0.370	0.517	0.663	0.789	0.882	0.941	0.974	0.990	0.997	0.999
9	0.257	0.392	0.546	0.696	0.819	0.905	0.956	0.983	0.994	0.998	1.000
10	0.269	0.411	0.571	0.722	0.842	0.922	0.967	0.988	0.996	0.999	1.000
12	0.289	0.442	0.610	0.762	0.875	0.944	0.979	0.993	0.998	1.000	1.000
15	0.311	0.477	0.652	0.802	0.905	0.962	0.988	0.997	0.999	1.000	1.000
20	0.336	0.514	0.694	0.839	0.931	0.976	0.993	0.999	1.000	1.000	1.000
30	0.365	0.555	0.738	0.874	0.952	0.986	0.997	0.999	1.000	1.000	1.000
60	0.397	0.598	0.781	0.906	0.969	0.992	0.999	1.000	1.000	1.000	1.000
120	0.414	0.621	0.802	0.920	0.975	0.994	0.999	1.000	1.000	1.000	1.000
1000	0.431	0.642	0.821	0.931	0.980	0.996	0.999	1.000	1.000	1.000	1.000

$v_1 = 6, \alpha = 0.05$

					ϕ						
v_2	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
5	0.189	0.279	0.388	0.509	0.628	0.736	0.824	0.891	0.936	0.966	0.983
6	0.211	0.317	0.443	0.578	0.704	0.809	0.887	0.939	0.970	0.986	0.994
7	0.231	0.349	0.490	0.633	0.760	0.859	0.925	0.965	0.985	0.994	0.998
8	0.248	0.378	0.528	0.677	0.802	0.893	0.949	0.978	0.992	0.997	0.999
9	0.262	0.402	0.560	0.712	0.834	0.916	0.964	0.986	0.996	0.999	1.000
10	0.276	0.423	0.588	0.741	0.858	0.933	0.973	0.991	0.997	0.999	1.000
12	0.298	0.458	0.631	0.784	0.892	0.955	0.984	0.996	0.999	1.000	1.000
15	0.323	0.497	0.677	0.826	0.922	0.972	0.992	0.998	1.000	1.000	1.000
20	0.352	0.540	0.724	0.865	0.947	0.984	0.996	0.999	1.000	1.000	1.000
30	0.385	0.587	0.772	0.901	0.967	0.992	0.998	1.000	1.000	1.000	1.000
60	0.424	0.637	0.819	0.931	0.981	0.996	0.999	1.000	1.000	1.000	1.000
120	0.445	0.663	0.842	0.944	0.986	0.997	1.000	1.000	1.000	1.000	1.000
1000	0.465	0.687	0.860	0.954	0.989	0.998	1.000	1.000	1.000	1.000	1.000

(continued)

Table A.7 (Power, continued)

$\nu_1 = 7, \alpha = 0.05$

ν_2	ϕ										
	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
5	0.189	0.280	0.390	0.511	0.631	0.739	0.828	0.894	0.939	0.967	0.984
6	0.212	0.319	0.448	0.584	0.710	0.815	0.892	0.943	0.972	0.988	0.995
7	0.233	0.354	0.497	0.642	0.769	0.867	0.931	0.968	0.987	0.995	0.998
8	0.251	0.384	0.538	0.688	0.813	0.901	0.954	0.982	0.993	0.998	0.999
9	0.267	0.410	0.573	0.726	0.846	0.925	0.969	0.989	0.997	0.999	1.000
10	0.281	0.434	0.602	0.756	0.871	0.942	0.978	0.993	0.998	1.000	1.000
12	0.305	0.472	0.649	0.801	0.906	0.963	0.988	0.997	0.999	1.000	1.000
15	0.333	0.515	0.699	0.845	0.935	0.978	0.994	0.999	1.000	1.000	1.000
20	0.366	0.562	0.750	0.886	0.959	0.989	0.998	1.000	1.000	1.000	1.000
30	0.405	0.615	0.801	0.921	0.977	0.995	0.999	1.000	1.000	1.000	1.000
60	0.450	0.672	0.850	0.950	0.988	0.998	1.000	1.000	1.000	1.000	1.000
120	0.475	0.701	0.873	0.962	0.992	0.999	1.000	1.000	1.000	1.000	1.000
1000	0.498	0.728	0.892	0.970	0.994	0.999	1.000	1.000	1.000	1.000	1.000

$\nu_1 = 8, \alpha = 0.05$

ν_2	ϕ										
	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
5	0.189	0.280	0.391	0.513	0.634	0.742	0.830	0.896	0.941	0.968	0.984
6	0.213	0.321	0.452	0.589	0.716	0.821	0.897	0.946	0.974	0.989	0.996
7	0.234	0.358	0.503	0.649	0.777	0.873	0.936	0.971	0.988	0.996	0.999
8	0.254	0.390	0.546	0.698	0.822	0.908	0.959	0.984	0.994	0.998	1.000
9	0.271	0.418	0.583	0.737	0.856	0.932	0.973	0.991	0.997	0.999	1.000
10	0.286	0.443	0.614	0.769	0.882	0.949	0.981	0.994	0.999	1.000	1.000
12	0.312	0.484	0.664	0.816	0.916	0.969	0.991	0.998	1.000	1.000	1.000
15	0.343	0.530	0.717	0.861	0.945	0.983	0.996	0.999	1.000	1.000	1.000
20	0.379	0.583	0.772	0.902	0.968	0.992	0.999	1.000	1.000	1.000	1.000
30	0.423	0.640	0.825	0.936	0.983	0.997	1.000	1.000	1.000	1.000	1.000
60	0.474	0.703	0.876	0.963	0.993	0.999	1.000	1.000	1.000	1.000	1.000
120	0.502	0.735	0.898	0.973	0.995	0.999	1.000	1.000	1.000	1.000	1.000
1000	0.530	0.764	0.917	0.981	0.997	1.000	1.000	1.000	1.000	1.000	1.000

(continued)

Table A.7 (Power, continued)

$\nu_1 = 9, \ \alpha = 0.05$

					ϕ						
ν_2	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
5	0.189	0.281	0.393	0.515	0.637	0.745	0.833	0.898	0.942	0.969	0.985
6	0.214	0.323	0.455	0.593	0.721	0.825	0.900	0.948	0.976	0.990	0.996
7	0.236	0.361	0.508	0.656	0.783	0.878	0.939	0.973	0.989	0.996	0.999
8	0.256	0.395	0.553	0.706	0.830	0.914	0.962	0.986	0.995	0.999	1.000
9	0.274	0.424	0.592	0.747	0.864	0.938	0.976	0.992	0.998	0.999	1.000
10	0.290	0.450	0.625	0.780	0.890	0.954	0.984	0.995	0.999	1.000	1.000
12	0.318	0.494	0.678	0.828	0.925	0.973	0.992	0.998	1.000	1.000	1.000
15	0.352	0.544	0.733	0.874	0.953	0.986	0.997	0.999	1.000	1.000	1.000
20	0.391	0.601	0.790	0.915	0.974	0.994	0.999	1.000	1.000	1.000	1.000
30	0.439	0.663	0.846	0.948	0.988	0.998	1.000	1.000	1.000	1.000	1.000
60	0.496	0.730	0.896	0.973	0.995	0.999	1.000	1.000	1.000	1.000	1.000
120	0.529	0.765	0.919	0.982	0.997	1.000	1.000	1.000	1.000	1.000	1.000
1000	0.560	0.795	0.936	0.988	0.999	1.000	1.000	1.000	1.000	1.000	1.000

$\nu_1 = 10, \ \alpha = 0.05$

					ϕ						
ν_2	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
5	0.189	0.281	0.394	0.517	0.639	0.747	0.835	0.900	0.943	0.970	0.985
6	0.215	0.325	0.458	0.597	0.725	0.829	0.903	0.950	0.977	0.990	0.996
7	0.238	0.364	0.513	0.661	0.789	0.883	0.942	0.975	0.990	0.997	0.999
8	0.258	0.399	0.560	0.713	0.836	0.919	0.965	0.987	0.996	0.999	1.000
9	0.277	0.430	0.600	0.755	0.871	0.942	0.978	0.993	0.998	1.000	1.000
10	0.294	0.457	0.634	0.789	0.897	0.958	0.986	0.996	0.999	1.000	1.000
12	0.324	0.504	0.689	0.839	0.932	0.977	0.994	0.999	1.000	1.000	1.000
15	0.359	0.557	0.747	0.885	0.959	0.989	0.998	1.000	1.000	1.000	1.000
20	0.402	0.617	0.806	0.926	0.979	0.996	0.999	1.000	1.000	1.000	1.000
30	0.454	0.683	0.863	0.958	0.991	0.999	1.000	1.000	1.000	1.000	1.000
60	0.517	0.755	0.913	0.980	0.997	1.000	1.000	1.000	1.000	1.000	1.000
120	0.553	0.791	0.935	0.987	0.999	1.000	1.000	1.000	1.000	1.000	1.000
1000	0.588	0.823	0.951	0.992	0.999	1.000	1.000	1.000	1.000	1.000	1.000

*Power was computed using the SAS statements "nc = v*phi**2;",
"Falpha = finv(1-alpha,nu1,nu2);", "power = 1 - probf(Falpha,nu1,nu2,nc);", and "power = round(power,.01);" for values of the parameters "v", "phi", "alpha", "nu1 = v-1" and "nu2"
"Falpha" is the upper-α critical value of the F-distribution with "nu1" and "nu2" degrees of freedom, and "nc" is the noncentrality parameter for the corresponding noncentral F-distribution

Table A.8 Tukey's method:* upper α critical values, $q_{v,df,\alpha}$, of the studentized range distribution

df	α							v							
		2	3	4	5	6	7	8	9	10	12	14	16	18	20
2	0.01	14.0	19.0	22.3	24.7	26.6	28.2	29.5	30.7	31.7	33.4	34.8	36.0	37.0	37.9
	0.05	6.08	8.33	9.80	10.9	11.7	12.4	13.0	13.5	14.0	14.7	15.4	15.9	16.4	16.8
	0.10	4.13	5.73	6.77	7.54	8.14	8.63	9.05	9.41	9.72	10.3	10.7	11.1	11.4	11.7
3	0.01	8.26	10.6	12.2	13.3	14.2	15.0	15.6	16.2	16.7	17.5	18.2	18.8	19.3	19.8
	0.05	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.95	10.3	10.7	11.0	11.2
	0.10	3.33	4.47	5.20	5.74	6.16	6.51	6.81	7.06	7.29	7.67	7.98	8.25	8.48	8.68
4	0.01	6.51	8.12	9.17	9.96	10.6	11.1	11.5	11.9	12.3	12.8	13.3	13.7	14.1	14.4
	0.05	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.21	8.52	8.79	9.03	9.23
	0.10	3.01	3.98	4.59	5.03	5.39	5.68	5.93	6.14	6.33	6.65	6.91	7.13	7.33	7.50
5	0.01	5.70	6.98	7.81	8.42	8.91	9.32	9.67	9.97	10.2	10.7	11.1	11.4	11.7	11.9
	0.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.32	7.60	7.83	8.03	8.21
	0.10	2.85	3.72	4.26	4.66	4.98	5.24	5.46	5.65	5.82	6.10	6.34	6.54	6.71	6.86
6	0.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.48	9.81	10.1	10.3	10.5
	0.05	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.79	7.03	7.24	7.43	7.59
	0.10	2.75	3.56	4.07	4.44	4.73	4.97	5.17	5.34	5.50	5.76	5.98	6.16	6.32	6.47
7	0.01	4.95	5.92	6.55	7.02	7.39	7.70	7.98	8.21	8.43	8.80	9.11	9.38	9.62	9.84
	0.05	3.34	4.16	4.68	5.06	5.36	5.61	5.81	5.99	6.15	6.42	6.65	6.84	7.00	7.15
	0.10	2.68	3.45	3.93	4.28	4.55	4.78	4.97	5.14	5.28	5.53	5.74	5.91	6.06	6.20
8	0.01	4.74	5.64	6.21	6.63	6.97	7.24	7.48	7.69	7.88	8.20	8.46	8.70	8.90	9.08
	0.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.17	6.39	6.57	6.72	6.86
	0.10	2.63	3.37	3.83	4.17	4.43	4.65	4.83	4.99	5.13	5.36	5.56	5.73	5.87	6.00
9	0.01	4.60	5.43	5.96	6.35	6.66	6.92	7.14	7.33	7.50	7.79	8.03	8.23	8.41	8.57
	0.05	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.98	6.19	6.36	6.51	6.64
	0.10	2.59	3.32	3.76	4.08	4.34	4.54	4.72	4.87	5.01	5.23	5.42	5.58	5.72	5.85
10	0.01	4.48	5.27	5.77	6.14	6.43	6.67	6.88	7.05	7.21	7.48	7.71	7.90	8.07	8.22
	0.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.83	6.03	6.19	6.34	6.47
	0.10	2.56	3.27	3.70	4.02	4.26	4.47	4.64	4.78	4.91	5.13	5.32	5.47	5.61	5.73
11	0.01	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.25	7.46	7.64	7.80	7.94
	0.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.71	5.90	6.06	6.20	6.33
	0.10	2.54	3.23	3.66	3.96	4.20	4.40	4.57	4.71	4.84	5.05	5.23	5.38	5.51	5.63

(continued)

Table A.8   (Tukey's method, continued)

df	α	2	3	4	5	6	7	8	9	10	12	14	16	18	20
12	0.01	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.68	6.82	7.07	7.28	7.46	7.62	7.76
	0.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.26	5.39	5.61	5.80	5.95	6.08	6.20
	0.10	2.52	3.20	3.62	3.92	4.16	4.35	4.51	4.65	4.78	4.99	5.16	5.31	5.44	5.55
14	0.01	4.21	4.89	5.32	5.63	5.88	6.09	6.26	6.41	6.55	6.77	6.97	7.13	7.27	7.40
	0.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.46	5.64	5.78	5.91	6.03
	0.10	2.49	3.16	3.56	3.85	4.08	4.27	4.42	4.56	4.68	4.88	5.05	5.19	5.32	5.43
16	0.01	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.56	6.74	6.90	7.03	7.15
	0.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.35	5.52	5.66	5.79	5.90
	0.10	2.47	3.12	3.52	3.80	4.03	4.21	4.36	4.49	4.61	4.80	4.97	5.11	5.23	5.33
18	0.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.41	6.58	6.73	6.85	6.97
	0.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.27	5.43	5.57	5.69	5.79
	0.10	2.45	3.10	3.49	3.77	3.98	4.16	4.31	4.44	4.55	4.75	4.90	5.04	5.16	5.26
20	0.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.28	6.45	6.59	6.71	6.82
	0.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.20	5.36	5.49	5.61	5.71
	0.10	2.44	3.08	3.46	3.74	3.95	4.12	4.27	4.40	4.51	4.70	4.85	4.99	5.10	5.20
24	0.01	3.96	4.55	4.91	5.17	5.37	5.54	5.68	5.81	5.92	6.11	6.26	6.39	6.51	6.61
	0.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.10	5.25	5.38	5.49	5.59
	0.10	2.42	3.05	3.42	3.69	3.90	4.07	4.21	4.34	4.44	4.63	4.78	4.91	5.02	5.12
30	0.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.93	6.08	6.20	6.31	6.41
	0.05	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	5.00	5.15	5.27	5.38	5.47
	0.10	2.40	3.02	3.39	3.65	3.85	4.02	4.16	4.28	4.38	4.56	4.71	4.83	4.94	5.03
40	0.01	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.76	5.90	6.02	6.12	6.21
	0.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.90	5.04	5.16	5.27	5.36
	0.10	2.38	2.99	3.35	3.60	3.80	3.96	4.10	4.21	4.32	4.49	4.63	4.75	4.86	4.95
60	0.01	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45	5.60	5.73	5.84	5.93	6.01
	0.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.81	4.94	5.06	5.15	5.24
	0.10	2.36	2.96	3.31	3.56	3.75	3.91	4.04	4.16	4.25	4.42	4.56	4.67	4.78	4.86
120	0.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.44	5.56	5.66	5.75	5.83
	0.05	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.71	4.84	4.95	5.04	5.13
	0.10	2.34	2.93	3.28	3.52	3.71	3.86	3.99	4.10	4.19	4.35	4.48	4.60	4.69	4.78
∞	0.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.29	5.40	5.49	5.57	5.65
	0.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.62	4.74	4.85	4.93	5.01
	0.10	2.33	2.90	3.24	3.48	3.66	3.81	3.93	4.04	4.13	4.28	4.41	4.52	4.61	4.69

The column group header above the numeric columns is v.

*Values $q_{v,df,\alpha}$ were generated using the SAS statement "qT=probmc('range',.,prob,df,v);", where "prob" $= 1 - \alpha$, and "df=." for $df = \infty$

Table A.9 Dunnett's one-sided method:* upper α critical coefficients $w_{D1} = t_{v-1,df,\alpha}^{(0.5)}$

df	α	$v-1$													
		2	3	4	5	6	7	8	9	10	12	14	16	18	20
2	0.01	8.88	10.0	10.9	11.5	12.0	12.5	12.8	13.2	13.5	14.0	14.4	14.7	15.1	15.3
	0.05	3.80	4.34	4.71	5.00	5.24	5.43	5.60	5.75	5.88	6.11	6.29	6.45	6.59	6.72
	0.10	2.54	2.92	3.19	3.40	3.57	3.71	3.83	3.94	4.03	4.19	4.32	4.44	4.54	4.62
3	0.01	5.48	6.04	6.44	6.74	6.99	7.20	7.38	7.53	7.67	7.91	8.11	8.28	8.43	8.56
	0.05	2.94	3.28	3.52	3.70	3.85	3.97	4.08	4.17	4.25	4.39	4.51	4.61	4.70	4.78
	0.10	2.13	2.41	2.61	2.75	2.87	2.97	3.06	3.13	3.20	3.31	3.41	3.49	3.56	3.62
4	0.01	4.41	4.80	5.07	5.28	5.45	5.59	5.71	5.82	5.92	6.08	6.22	6.34	6.44	6.53
	0.05	2.61	2.88	3.08	3.22	3.34	3.44	3.52	3.59	3.66	3.77	3.86	3.94	4.01	4.07
	0.10	1.96	2.20	2.37	2.50	2.60	2.68	2.75	2.82	2.87	2.97	3.05	3.11	3.17	3.22
5	0.01	3.90	4.21	4.43	4.60	4.73	4.85	4.94	5.03	5.11	5.24	5.34	5.44	5.52	5.59
	0.05	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30	3.36	3.45	3.53	3.60	3.66	3.71
	0.10	1.87	2.09	2.24	2.36	2.45	2.53	2.59	2.65	2.70	2.78	2.86	2.92	2.97	3.02
6	0.01	3.61	3.88	4.06	4.21	4.32	4.42	4.51	4.58	4.64	4.76	4.85	4.93	5.00	5.06
	0.05	2.34	2.56	2.71	2.83	2.92	3.00	3.06	3.12	3.17	3.26	3.33	3.40	3.45	3.50
	0.10	1.82	2.02	2.17	2.27	2.36	2.43	2.49	2.54	2.59	2.67	2.74	2.79	2.84	2.89
7	0.01	3.42	3.66	3.83	3.96	4.06	4.15	4.22	4.29	4.35	4.45	4.53	4.60	4.67	4.72
	0.05	2.27	2.48	2.62	2.73	2.82	2.89	2.95	3.00	3.05	3.13	3.20	3.26	3.31	3.36
	0.10	1.78	1.98	2.11	2.22	2.30	2.37	2.42	2.47	2.52	2.59	2.66	2.71	2.76	2.80
8	0.01	3.29	3.51	3.66	3.78	3.88	3.96	4.03	4.09	4.14	4.23	4.31	4.38	4.43	4.49
	0.05	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92	2.96	3.04	3.11	3.16	3.21	3.25
	0.10	1.75	1.94	2.08	2.17	2.25	2.32	2.38	2.42	2.47	2.54	2.60	2.65	2.70	2.74
9	0.01	3.19	3.40	3.54	3.66	3.75	3.82	3.89	3.94	3.99	4.08	4.15	4.21	4.26	4.31
	0.05	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86	2.90	2.97	3.04	3.09	3.14	3.18
	0.10	1.73	1.92	2.05	2.14	2.22	2.28	2.34	2.39	2.43	2.50	2.56	2.61	2.65	2.69
10	0.01	3.11	3.31	3.45	3.56	3.64	3.72	3.78	3.83	3.88	3.96	4.03	4.08	4.14	4.18
	0.05	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81	2.85	2.92	2.98	3.03	3.08	3.12
	0.10	1.71	1.90	2.02	2.12	2.19	2.26	2.31	2.35	2.40	2.46	2.52	2.57	2.61	2.65
11	0.01	3.06	3.25	3.38	3.48	3.56	3.63	3.69	3.74	3.79	3.86	3.93	3.99	4.03	4.08
	0.05	2.13	2.31	2.43	2.53	2.60	2.67	2.72	2.77	2.81	2.88	2.94	2.99	3.03	3.07
	0.10	1.70	1.88	2.01	2.10	2.17	2.23	2.29	2.33	2.37	2.44	2.49	2.54	2.58	2.62

(continued)

Table A.9 (Dunnett's one-sided method, continued)

df	α	2	3	4	5	6	7	8	9	10	12	14	16	18	20
								$v-1$							
12	0.01	3.01	3.19	3.32	3.42	3.50	3.56	3.62	3.67	3.71	3.79	3.85	3.91	3.95	3.99
	0.05	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74	2.78	2.84	2.90	2.95	2.99	3.03
	0.10	1.69	1.87	1.99	2.08	2.16	2.22	2.27	2.31	2.35	2.42	2.47	2.52	2.56	2.60
14	0.01	2.94	3.11	3.23	3.33	3.40	3.46	3.51	3.56	3.60	3.67	3.73	3.78	3.83	3.87
	0.05	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69	2.73	2.79	2.85	2.89	2.93	2.97
	0.10	1.67	1.85	1.97	2.06	2.13	2.19	2.24	2.28	2.32	2.38	2.44	2.48	2.52	2.56
16	0.01	2.88	3.05	3.17	3.26	3.33	3.39	3.44	3.48	3.52	3.59	3.65	3.70	3.74	3.77
	0.05	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65	2.69	2.75	2.81	2.85	2.89	2.93
	0.10	1.66	1.83	1.95	2.04	2.11	2.17	2.22	2.26	2.30	2.36	2.41	2.46	2.50	2.53
18	0.01	2.84	3.01	3.12	3.21	3.27	3.33	3.38	3.42	3.46	3.53	3.58	3.63	3.67	3.71
	0.05	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62	2.66	2.72	2.78	2.82	2.86	2.89
	0.10	1.65	1.82	1.94	2.02	2.09	2.15	2.20	2.24	2.28	2.34	2.39	2.44	2.48	2.51
20	0.01	2.81	2.97	3.08	3.17	3.23	3.29	3.34	3.38	3.42	3.48	3.53	3.58	3.62	3.65
	0.05	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60	2.64	2.70	2.75	2.80	2.83	2.87
	0.10	1.64	1.81	1.93	2.01	2.08	2.14	2.19	2.23	2.26	2.33	2.38	2.42	2.46	2.49
24	0.01	2.77	2.92	3.03	3.11	3.17	3.22	3.27	3.31	3.35	3.41	3.46	3.50	3.54	3.57
	0.05	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57	2.60	2.66	2.72	2.76	2.80	2.83
	0.10	1.63	1.80	1.91	2.00	2.06	2.12	2.17	2.21	2.24	2.30	2.35	2.40	2.43	2.47
30	0.01	2.72	2.87	2.97	3.05	3.11	3.16	3.21	3.25	3.28	3.34	3.39	3.43	3.46	3.50
	0.05	1.99	2.15	2.25	2.34	2.40	2.45	2.50	2.54	2.57	2.63	2.68	2.72	2.76	2.79
	0.10	1.62	1.79	1.90	1.98	2.05	2.10	2.15	2.19	2.22	2.28	2.33	2.37	2.41	2.44
40	0.01	2.68	2.82	2.92	2.99	3.05	3.10	3.14	3.18	3.21	3.27	3.32	3.36	3.39	3.42
	0.05	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51	2.54	2.60	2.65	2.69	2.72	2.75
	0.10	1.61	1.77	1.88	1.96	2.03	2.08	2.13	2.17	2.20	2.26	2.31	2.35	2.39	2.42
60	0.01	2.64	2.78	2.87	2.94	3.00	3.04	3.08	3.12	3.15	3.20	3.25	3.29	3.32	3.35
	0.05	1.95	2.10	2.21	2.28	2.34	2.40	2.44	2.48	2.51	2.57	2.61	2.65	2.69	2.72
	0.10	1.60	1.76	1.87	1.95	2.01	2.06	2.11	2.15	2.18	2.24	2.29	2.33	2.36	2.39
120	0.01	2.60	2.73	2.82	2.89	2.94	2.99	3.03	3.06	3.09	3.14	3.18	3.22	3.25	3.28
	0.05	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45	2.48	2.53	2.58	2.62	2.65	2.68
	0.10	1.59	1.75	1.85	1.93	1.99	2.05	2.09	2.13	2.16	2.22	2.27	2.31	2.34	2.37
∞	0.01	2.56	2.69	2.77	2.84	2.89	2.93	2.97	3.00	3.03	3.08	3.12	3.15	3.18	3.21
	0.05	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42	2.45	2.50	2.55	2.58	2.62	2.64
	0.10	1.58	1.73	1.84	1.92	1.98	2.03	2.07	2.11	2.14	2.20	2.24	2.28	2.32	2.35

*Values $w_{D1} = t_{v-1,df,\alpha}^{(0.5)}$ were generated using the SAS statement

"wD1 = probmc('dunnett1',.,prob,df,vm1);", where "prob" $= 1 - \alpha$, "vm1" $= v - 1$, $t_{v-1,df,\alpha}^{(0.5)}$ is the upper α critical value for the maximum of a $(v - 1)$-variate t-distribution with common correlation $\rho = 0.5$ and degrees of freedom df, and "df=." for $df = \infty$

Table A.10 Dunnett's two-sided method:* upper α critical coefficients $w_{D2} = |t|^{(0.5)}_{v-1,df,\alpha}$

df	α							$v-1$							
		2	3	4	5	6	7	8	9	10	12	14	16	18	20
2	0.01	12.4	13.8	14.8	15.6	16.2	16.7	17.1	17.5	17.8	18.4	18.8	19.2	19.6	19.9
	0.05	5.42	6.06	6.51	6.85	7.12	7.35	7.54	7.71	7.85	8.10	8.31	8.49	8.64	8.77
	0.10	3.72	4.18	4.50	4.74	4.93	5.09	5.23	5.34	5.45	5.62	5.77	5.89	6.00	6.09
3	0.01	6.97	7.64	8.10	8.46	8.74	8.98	9.19	9.37	9.52	9.79	10.0	10.2	10.4	10.5
	0.05	3.87	4.26	4.54	4.75	4.92	5.06	5.18	5.28	5.37	5.53	5.66	5.77	5.87	5.95
	0.10	2.91	3.23	3.45	3.62	3.75	3.87	3.96	4.04	4.12	4.24	4.34	4.43	4.51	4.58
4	0.01	5.36	5.81	6.12	6.36	6.55	6.72	6.85	6.98	7.08	7.27	7.42	7.55	7.66	7.77
	0.05	3.31	3.62	3.83	3.99	4.13	4.23	4.33	4.41	4.48	4.60	4.71	4.79	4.87	4.94
	0.10	2.60	2.86	3.05	3.18	3.30	3.39	3.47	3.54	3.60	3.70	3.79	3.86	3.92	3.98
5	0.01	4.63	4.97	5.22	5.41	5.56	5.68	5.79	5.89	5.97	6.11	6.24	6.34	6.43	6.51
	0.05	3.03	3.29	3.48	3.62	3.73	3.82	3.90	3.97	4.03	4.14	4.23	4.30	4.37	4.42
	0.10	2.43	2.67	2.83	2.96	3.05	3.14	3.21	3.27	3.32	3.41	3.49	3.56	3.61	3.66
6	0.01	4.21	4.51	4.71	4.87	5.00	5.10	5.20	5.28	5.35	5.47	5.57	5.66	5.74	5.80
	0.05	2.86	3.10	3.26	3.39	3.49	3.57	3.64	3.71	3.76	3.86	3.94	4.00	4.06	4.11
	0.10	2.33	2.55	2.70	2.81	2.91	2.98	3.05	3.10	3.15	3.24	3.31	3.37	3.42	3.47
7	0.01	3.95	4.21	4.39	4.53	4.64	4.74	4.82	4.89	4.96	5.07	5.16	5.24	5.31	5.37
	0.05	2.75	2.97	3.12	3.24	3.33	3.41	3.48	3.53	3.58	3.67	3.75	3.81	3.86	3.91
	0.10	2.26	2.47	2.61	2.72	2.81	2.88	2.94	2.99	3.04	3.12	3.18	3.24	3.29	3.33
8	0.01	3.77	4.00	4.17	4.29	4.40	4.48	4.56	4.62	4.68	4.78	4.86	4.93	5.00	5.05
	0.05	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41	3.46	3.54	3.61	3.67	3.72	3.76
	0.10	2.22	2.41	2.55	2.65	2.73	2.80	2.86	2.91	2.96	3.03	3.10	3.15	3.20	3.24
9	0.01	3.63	3.85	4.01	4.12	4.22	4.30	4.37	4.43	4.48	4.57	4.65	4.71	4.77	4.82
	0.05	2.61	2.81	2.95	3.05	3.14	3.20	3.26	3.32	3.36	3.44	3.51	3.56	3.61	3.65
	0.10	2.18	2.37	2.50	2.60	2.68	2.74	2.80	2.85	2.89	2.97	3.03	3.08	3.13	3.17
10	0.01	3.53	3.74	3.88	3.99	4.08	4.16	4.22	4.28	4.33	4.42	4.49	4.55	4.60	4.65
	0.05	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24	3.29	3.36	3.43	3.48	3.53	3.57
	0.10	2.15	2.34	2.46	2.56	2.64	2.70	2.75	2.80	2.84	2.92	2.98	3.03	3.07	3.11
11	0.01	3.45	3.65	3.79	3.89	3.98	4.05	4.11	4.16	4.21	4.29	4.36	4.42	4.47	4.52
	0.05	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19	3.23	3.30	3.36	3.42	3.46	3.50
	0.10	2.13	2.31	2.43	2.53	2.60	2.66	2.72	2.76	2.80	2.87	2.93	2.98	3.03	3.06

(continued)

Table A.10 (Dunnett's two-sided method, continued)

df	α							$v-1$							
		2	3	4	5	6	7	8	9	10	12	14	16	18	20
12	0.01	3.39	3.58	3.71	3.81	3.89	3.96	4.02	4.07	4.12	4.19	4.26	4.32	4.37	4.41
	0.05	2.50	2.68	2.81	2.90	2.98	3.04	3.09	3.14	3.18	3.25	3.31	3.36	3.41	3.45
	0.10	2.11	2.29	2.41	2.50	2.57	2.64	2.69	2.73	2.77	2.84	2.90	2.95	2.99	3.03
14	0.01	3.29	3.47	3.59	3.69	3.76	3.83	3.88	3.93	3.97	4.05	4.11	4.16	4.20	4.25
	0.05	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07	3.11	3.18	3.23	3.28	3.32	3.36
	0.10	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.68	2.72	2.79	2.84	2.89	2.93	2.97
16	0.01	3.22	3.39	3.51	3.60	3.67	3.73	3.78	3.83	3.87	3.94	4.00	4.05	4.09	4.13
	0.05	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02	3.06	3.12	3.18	3.22	3.26	3.30
	0.10	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65	2.69	2.75	2.80	2.85	2.89	2.93
18	0.01	3.17	3.33	3.44	3.53	3.60	3.66	3.71	3.75	3.79	3.86	3.91	3.96	4.00	4.04
	0.05	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98	3.01	3.08	3.13	3.18	3.22	3.25
	0.10	2.04	2.21	2.32	2.41	2.47	2.53	2.58	2.62	2.66	2.72	2.77	2.82	2.86	2.89
20	0.01	3.13	3.29	3.40	3.48	3.55	3.60	3.65	3.69	3.73	3.80	3.85	3.90	3.94	3.97
	0.05	2.38	2.54	2.65	2.73	2.80	2.86	2.90	2.95	2.98	3.05	3.10	3.14	3.18	3.22
	0.10	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60	2.64	2.70	2.75	2.79	2.83	2.87
24	0.01	3.07	3.22	3.32	3.40	3.47	3.52	3.57	3.61	3.64	3.70	3.76	3.80	3.84	3.87
	0.05	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90	2.94	3.00	3.05	3.09	3.13	3.16
	0.10	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57	2.60	2.66	2.71	2.76	2.79	2.83
30	0.01	3.01	3.15	3.25	3.33	3.39	3.44	3.49	3.52	3.56	3.62	3.66	3.71	3.74	3.77
	0.05	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86	2.89	2.95	3.00	3.04	3.08	3.11
	0.10	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54	2.57	2.63	2.68	2.72	2.76	2.79
40	0.01	2.95	3.09	3.19	3.26	3.32	3.37	3.41	3.44	3.48	3.53	3.58	3.62	3.65	3.68
	0.05	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81	2.85	2.90	2.95	2.99	3.02	3.06
	0.10	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51	2.54	2.60	2.65	2.69	2.72	2.75
60	0.01	2.90	3.03	3.12	3.19	3.25	3.29	3.33	3.37	3.40	3.45	3.49	3.53	3.56	3.59
	0.05	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77	2.80	2.86	2.90	2.94	2.97	3.00
	0.10	1.95	2.10	2.21	2.28	2.34	2.40	2.44	2.48	2.51	2.57	2.61	2.65	2.69	2.72
120	0.01	2.85	2.97	3.06	3.12	3.18	3.22	3.26	3.29	3.32	3.37	3.41	3.45	3.48	3.50
	0.05	2.24	2.38	2.47	2.55	2.60	2.65	2.69	2.73	2.76	2.81	2.86	2.89	2.93	2.95
	0.10	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45	2.48	2.53	2.58	2.62	2.65	2.68
∞	0.01	2.79	2.91	3.00	3.06	3.11	3.15	3.19	3.22	3.25	3.29	3.33	3.37	3.40	3.42
	0.05	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69	2.72	2.77	2.81	2.85	2.88	2.91
	0.10	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42	2.45	2.50	2.55	2.58	2.62	2.64

*Values $w_{D2} = t_{v-1,df,\alpha}^{(0.5)}$ generated using the SAS statement "wD2 = probmc('dunnett2',.,prob, df,vm1);", where "prob" $= 1 - \alpha$, "vm1" $= v - 1$, $|t|_{v-1,df,\alpha}^{(0.5)}$ is the upper α critical value for the maximum absolute value of a $(v-1)$-variate t-distribution with common correlation $\rho = 0.5$ and degrees of freedom df, and "df=." for $df = \infty$

Table A.11 Voss–Wang method: upper α critical coefficients $w_V = v_{m,d,\alpha}$ for m orthogonal contrasts and d sums of squares pooled into each quasi mean squared error

even						odd					
			α						α		
m	d	0.10	0.05	0.01	0.001	m	d	0.10	0.05	0.01	0.001
2	1	12.4	25.0	126.	1384.	3	2	5.31	7.61	17.0	54.4
4	2	9.08	13.1	29.9	105.	5	3	6.35	8.27	14.7	32.0
6	3	8.57	11.2	19.9	50.0	7	4	6.82	8.56	13.8	26.8
8	4	8.37	10.5	16.6	33.2	9	5	7.18	8.73	13.3	24.7
10	5	8.39	10.3	16.0	29.5	11	6	7.42	8.89	13.0	21.4
12	6	8.37	10.1	14.7	24.1	13	7	7.64	8.95	12.4	20.4
14	7	8.43	9.89	14.0	22.5	15	8	7.76	9.04	12.4	19.2
16	8	8.45	9.88	13.5	21.8	17	9	7.85	9.10	12.4	19.1
18	9	8.51	9.77	13.5	19.9	19	10	7.98	9.10	12.2	17.8
20	10	8.50	9.70	12.9	19.0	21	11	8.03	9.17	12.1	16.8
22	11	8.53	9.78	12.9	18.2	23	12	8.12	9.19	11.9	17.2
24	12	8.55	9.76	12.7	18.6	25	13	8.15	9.28	11.9	17.2
26	13	8.60	9.80	12.5	18.2	27	14	8.22	9.33	11.8	16.5
28	14	8.61	9.78	12.6	18.3	29	15	8.27	9.36	11.8	16.7
30	15	8.63	9.78	12.4	17.5	31	16	8.32	9.38	11.8	16.3
32	16	8.70	9.76	12.4	16.9	33	17	8.35	9.38	11.8	15.9
34	17	8.66	9.70	12.2	16.5	35	18	8.38	9.35	11.7	15.8
36	18	8.69	9.71	12.1	16.3	37	19	8.41	9.40	11.6	15.6
38	19	8.70	9.67	12.0	16.5	39	20	8.43	9.37	11.6	16.0
40	20	8.71	9.69	12.0	16.5	41	21	8.45	9.36	11.5	15.5
42	21	8.72	9.64	12.0	15.8	43	22	8.48	9.39	11.6	15.5
44	22	8.73	9.68	12.0	15.7	45	23	8.52	9.40	11.6	15.2
46	23	8.76	9.69	11.9	15.7	47	24	8.53	9.42	11.5	15.1
48	24	8.75	9.65	11.8	15.8	49	25	8.55	9.44	11.5	15.1
50	25	8.76	9.68	11.8	15.7	51	26	8.57	9.42	11.5	15.0
52	26	8.78	9.66	11.8	15.6	53	27	8.60	9.45	11.5	15.0
54	27	8.80	9.69	11.8	15.5	55	28	8.61	9.46	11.4	15.0
56	28	8.82	9.70	11.8	15.7	57	29	8.64	9.51	11.5	15.1
58	29	8.81	9.71	11.7	15.3	59	30	8.63	9.49	11.4	15.0
60	30	8.83	9.70	11.7	15.3	61	31	8.66	9.49	11.4	15.1
62	31	8.84	9.68	11.7	15.9	63	32	8.67	9.50	11.4	15.4

Bibliography

Anderson, V. L., & McLean, R. A. (1974). *Design of experiments: a realistic approach*. New York: Marcel Dekker Inc.

Bainbridge, J. R. (1951). Factorial experiments in pilot plant studies. *Industrial and Engineering Chemistry*, *43*, 1300–1306.

Barnabas, I. J., Dean, J. R., Tomlinson, W. R., & Owen, S. P. (1995). Experimental design approach for the extraction of polycyclic aromatic hydrocarbons form soil using supercritical carbon dioxide. *Analytical Chemistry*, *67*, 2064–2069.

Barnett, M. K., & Mead, F. C, Jr. (1956). A 2^4 factorial experiment in four blocks of eight: a study in radioactive decontamination. *Applied Statistics*, *5*, 122–131.

Baten, W. D. (1956). An analysis of variance applied to screw machines. *Industrial Quality Control*, *12*, 8–9.

Beckman, R. J., Nachtsheim, C. J., & Cook, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, *29*, 413–426.

Blom, G. (1958). *Statistical estimates and transformed beta variables*. New York: Wiley.

Box, G. E. P., & Behnken, D. W. (1960). Some new three level designs for the study of quantitative variables. *Technometrics*, *2*, 455–475.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, B*, *26*, 211–243.

Box, G. E. P., & Hunter, J. S. (1957). Multi-factor experimental designs for exploring response surfaces. *Annals of Mathematical Statistics*, *28*, 195–241.

Box, G. E. P., & Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series B*, *13*, 1–45.

Brickell, J., & Knox, K. (1992). Designed experiments case history—environmental engineering operation of an activated sludge system. In S. R. Schmidt & R. G. Launsby (Eds.), *Understanding industrial designed experiments* (3rd ed., pp. 8.187–8.194). Colorado: Air Academy Press.

Bullough, R. C., & Melby, C. L. (1993). Effect of inpatient versus outpatient measurement protocol on resting metabolic rate and respiratory exchange ratio. *Annals of Nutrition and Metabolism*, *37*, 24–32.

Cameron, J. M. (1951). The use of components of variance in preparing schedules for sampling of baled wool. *Biometrics*, *7*, 83–96.

Clatworthy, W. H. (1973). *Tables of two-associate-class partially balanced designs*., Applied Mathematics Series Washington: National Bureau of Standards. 63.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs* (2nd ed.). New York: Wiley.

Corlett, E. N., & Gregory, G. (1960). The consistency of setting of a machine tool handwheel. *Applied Statistics*, *9*, 92–102.

Cunningham, A. C., & O'Connor, N. (1968). Consumer reaction to retail price and display changes. *European Journal of Marketing*, *2*, 147–151.

Cuq, B., Aymard, C., Cuq, J.-L., & Guilbert, S. (1995). Edible packaging films based on fish myofibrillar proteins: formulation and functional purposes. *Journal of Food Science*, *60*, 1369–1373.

Daniel, C. (1976). *Applications of statistics to industrial experimentation*. New York: Wiley.

Davies, O. L. (Ed.). (1963). *Design and analysis of industrial experiments* (2nd ed.). London: Longman.

Dean, A. M., & Draper, N. R. (1999). Saturated main-effect designs for factorial experiments. *Statistics and Computing*, *9*, 179–185.

Desmond, D. J. (1954). Quality control on the setting of voltage regulators. *Applied Statistics*, *3*, 65–73.

Draguljić, D., Nekkanty, S., Santner, T. J., Dean, A. M., & Shivpuri, R. (2015). Optimizing thin film tool coatings using a finite element computer simulator. *Quality Engineering*, *27*(4), 461–472.

Draper, N. R. (1982). Center points in second-order response surface designs. *Technometrics*, *24*, 127–133.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.

Dunnett, C. W. (1955). A multiple comparisons procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, *50*, 1096–1121.

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association*, *75*, 796–800.

Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression II. *Biometrika*, *38*, 159–178.

Eibl, S., Kess, U., & Pukelsheim, F. (1992). Achieving a target value for a manufacturing process: a case study. *Journal of Quality Technology*, *24*, 22–26.

Erler, A., Mas, N, de, Ramsey, P., & Henderson, G. (2103). Efficient biological process characterization by definitive-screening designs: the formaldehyde treatment of a therapeutic protein as a case study. *Biotechnology Letters*, *35*, 323–329.

Finner, H., & Strassburger, K. (2002). The partitioning principle: a powerful tool in multiple decision theory. *Annals of Statistics*, *30*, 1194–1213.

Fisher, R. A., & Yates, F. (1973). *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver and Boyd.

Games, P. A., & Howell, J. F. (1976). Comparison procedure with unequal N's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, *1*, 113–125.

Graybill, F. A. (1976). *Theory and application of the linear model*. North Scituate: Duxbury Press.

Gupte, M., & Kulkarni, P. (2003). A study of antifungal antibiotic production by Thermomonospora sp MTCC 3340 using full factorial design. *Journal of Chemical Technology and Biotechnology*, *78*, 605–610.

Gutsell, J. S. (1951). The effect of sulfamerazine on the erythrocyte and hemoglobin content of trout blood. *Biometrics*, *7*, 171–179.

Hare, L. B. (1988). In the soup: a case study to identify contributors to filling variability. *Journal of Quality Technology*, *20*, 36–43.

Hayeck, G. T. (2009). The kinematics of the upper extremity and subsequent effects on joint loading and surgical treatment. Ph.D. thesis. Cornell University.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Statistics*, *12*, 61–75.

Hicks, C. R. (1956). Fundamentals of analysis of variance, Part III-nested designs in analysis of variance. *Industrial Quality Control*, *13*(4), 13–16.

Hicks, C. R. (1965). The analysis of covariance. *Industrial Quality Control*, *22*, 282–286.

Hicks, C. R. (1993). *Fundamental concepts in the design of experiments* (4th ed.). New York: Oxford University Press Inc.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.

Hocking, R. R. (1996). *Methods and applications of linear models; regression and the analysis of variance*. New York: Wiley.

Hoerl, A. E. (1988). *The 3–7 phenomena*. January issue: Royal Statistical Society News and Notes.

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric statistical methods*. New York: Wiley.

Hsu, J. C. (1996). *Multiple comparisons: theory and methods*. London: Chapman and Hall.

Inman, J., Ledolter, J., Lenth, R. V., & Niemi, L. (1992). Two case studies involving an optical emission spectrometer. *Journal of Quality Technology*, *24*, 27–36.

Jeffers, J. N. R. (1987). Acid rain and tree roots: an analysis. In D. J. Hand & B. S. Everitt (Eds.), *The statistical consultant in action*. New York: Cambridge University Press.

John, J. A. (1981). Efficient cyclic designs. *Journal of the Royal Statistical Society, Series B*, *43*, 76–80.

John, J. A., Wolock, F. W., & David, H. A. (1972). *Cyclic designs*., Applied Mathematics Series Washington: National Bureau of Standards. 62.

John, P. W. M. (1961). An application of a balanced incomplete design. *Technometrics*, *3*, 51–54.

John, P. W. M. (1971). *Statistical design and analysis of experiments*. New York: Macmillan.

Johnson, N. L., & Leone, F. C. (1977). *Statistics and experimental design in engineering and the physical sciences* (2nd ed.). New York: Wiley.

Jones, B., & Kenward, M. G. (2003). *Design and analysis of crossover trials* (2nd ed.). Boca Raton: Chapman and Hall/CRC.

Jones, B., & Nachtsheim, C. J. (2011). A class of three-level designs for definitive screening in the presence of second-order effects. *Journal of Quality Technology*, *43*, 1–15.

Kackar, R. N., & Shoemaker, A. C. (1986). Robust design: a cost-effective method for improving manufacturing processes. *AT&T Technical Journal*, *65*(2), 39–50.

Kempthorne, O. (1976). *Design and analysis of experiments*. New York: Krieger.

Kempthorne, O. (1977). Why randomize? *Journal of Statistical Planning and Inference*, *1*, 1–25.

Khuri, A. I., & Cornell, J. A. (1987). *Response surfaces: designs and analyses*. New York: Marcel Dekker.

Lamacraft, R. R., & Hall, W. B. (1982). Tables of cyclic incomplete block designs: $r = k$. *Australian Journal of Statistics*, *24*, 350–360.

Lewis, S. M., & Dean, A. M. (1980). Factorial experiments in resolvable generalized cyclic designs. *Bulletin in Applied Statistics*, *7*, 159–167.

Lewis, S. M., Hodgson, B. A., New, R. E., & Sexton, C. J. (1989). The application of Taguchi methods at the design analysis stage. *Proceedings of the Institute of Mechanical Engineering, International Conference on Engineering Design* **1**, (pp. 283–294).

Lin, D. K. J. (1993). A new class of supersaturated designs. *Technometrics*, *35*, 28–31.

Lorenz, R. C., Hsu, J. C., & Tuovinen, O. H. (1982). Performance variability, ranking and selection analysis of membrane filters for enumerating coliform bacteria in river water. *Journal of the American Water Works Association*, *74*, 429–437.

Mahadevan, S. (2009). Visualization methods and user interface design guidelines for rapid decision making in complex multi-task time-critical environments. Ph.D. Dissertation, Department of Biomedical, Industrial, and Human Factors Engineering, Wright State University.

Modi, O. P., Yadav, R. P., Mondal, D. P., Dasgupta, R., Das, S., & Yegneswaran, A. H. (2001). Abrasive wear behaviour of zinc-aluminium alloy - 10% Al_2O_3 composite through factorial design of experiment. *Journal of Materials Science*, *36*, 1601–1607.

Moore, M. A., & Epps, H. H. (1992). Accelerated weathering of marine fabrics. *Journal of Testing and Evaluation*, *20*, 139–143.

Munro, K. J. (1986). Investigation of optokinetic nystagmus under different visual conditions. MSc Dissertation, Department of Audiology, University of Southampton.

Naveena, B. J., Altaf, M., Bhadriah, K., & Reddy, G. (2005). Selection of medium components by Plackett-Burman design for production of L(+) lactic acid by Lactobacillus amylophilus GV6 in SSF using wheat bran. *Bioresource Technology*, *96*, 485–490.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago: Richard D. Irwin Inc.

Olsen, R. E., Bartholomew, C. H., Enfield, D. B., Lawson, J. S., Rohbock, N., Scott, B. S., & Woodfield, B. F. (2014). Optimizing the synthesis and properties of Al-modified anatase catalyst supports by statistical experimental design. *Journal of Porous Materials*, *21*, 827–837.

Peake, R. E. (1953). Planning an experiment in a cotton spinning mill. *Applied Statistics*, *2*, 184–192.

Peiser, A. M. (1943). Asymptotic formulas for significance levels of a certain distributions. *Annals of Mathematical Statistics*, *14*, 56–62. (Correction 1949, *Annals of Mathematical Statistics*, *20*, 128–129).

Plackett, R. L., & Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, *33*, 305–325.

Poon, G. K. K. (1995). Sequential experimental study and optimisation of an acid copper pattern plating process. *Circuit World*, *22*, 7–9.

Pugh, C. (1953). The evaluation of detergent performance in domestic dishwashing. *Applied Statistics*, *2*, 172–179.

Ratkowsky, D. A., Evans, M. A., & Alldredge, J. R. (1993). *Cross-over experiments: design, analysis and application.*, Statistics Textbooks and Monographs, 135, New York: Marcel Dekker.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/

Rumbell, T. H., Draguljić, D., Yadav, A., Hof, P. R., Luebke, J. I., & Weaver, C. M. (2016). Automated evolutionary optimization of ion channel conductances and kinetics in models of young and aged rhesus monkey pyramidal neurons. *Journal of Computational Neuroscience*, *41*(1), 65–90.

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, *4*, 409–435.

Salvadori, M. (1980). *Why buildings stand up: the strength of architecture*. New York: Norton.

Santner, T. J., Williams, B. J., & Notz, W. I. (2003). *The design and analysis of computer experiments*. New York: Springer-Verlag.

Saravanan, P., Selvarajan, V., Joshi, S. V., & Sundararajan, G. (2001). Experimental design and performance analysis of alumina coatings deposited by a detonation spray process. *Journal of Physics D: Applied Physics*, *34*(1), 131–140.

SAS Institute Inc. (2004). SAS/STAT. ® 9.1 *Users Guide*. Cary NC: SAS Institute Inc.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, *2*, 110–114.

Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.

Schmidt, S. R., & Launsby, R. G. (1992). *Experimental design on an injection molded plastic part* (3rd ed.)., In Understanding Industrial Designed Experiments Colorado: Air Academy Press.

Shahani, A. K. (1970). A saturated experiment in sequential determination of operating conditions. *The Statistician*, *19*, 403–408.

Sinha, K. (1991). A list of new group divisible designs. *Journal of Research of the National Institute of Standards and Technology*, *96*, 613–615.

Smith, H. F. (1936). The problem of comparing the results of two experiments with unequal errors. *Journal of the Council of Scientific and Industrial Research*, *9*, 211–212.

Smith, H, Jr. (1969). The analysis of data from a designed experiment. *Journal of Quality Technology*, *1*, 259–263.

Sosada, M. (1993). Optimal conditions for fractionation of rapeseed lecithin with alcohols. *Journal of the American Oil Chemists' Society*, *70*, 405–410.

Stefansson, G., Kim, W.-C., & Hsu, J. C. (1988). On confidence sets in multiple comparisons. In S. S. Gupta & J. O. Berger, (Eds.), *Statistical Decision Theory, I. V. Related Topcis* (2nd ed., pp. 89–104). New York: Academic Press.

Surjanovic, S., & Bingham, D. (2013). Virtual Library of Simulation Experiments: Test Functions and Datasets. Retrieved November 2, 2015, from http://www.sfu.ca/~ssurjano.

Thompson, W. A, Jr., & Moore, J. R. (1963). Non-negative estimates of variance components. *Technometrics*, *5*, 441–449.

Tuck, M. G., Lewis, S. M., & Cottrell, J. I. L. (1993). Response surface methodology and Taguchi: a quality improvement study from the milling industry. *Applied Statistics*, *42*, 671–681.

Tukey, J. W. (1953). The problem of multiple comparisons. *Dittoed manuscript of 396 pages, Department of Statistics,* Princeton University.

Vance, F. P. (1962). Optimization study of lube oil treatment by process 'X'. *Proceedings of the Symposium on Application of Statistics and Computers to Fuel Lubricant Research Problems.* Office of Chief Ordinance, U.S. Army, May 1962.

Voss, D. T. (2008). Partition testing: a generalization of hypothesis testing. *Statistics and Applications*, *6*, 17–24.

Voss, D. T. (2010). Testing is confidence estimation: partition multiple testing. *Journal of Statistical Theory and Practice*, *4*, 559–569.

Voss, D. T., & Wang, W. (1999). Exact simultaneous confidence intervals in the analysis of orthogonal saturated designs. *Journal of Statistical Planning and Inference*, *81*, 383–392.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350–362.

Wesler, M. Mc. (2001). The effects of display type and visual presentation format (frame of reference and perspective) on situation awareness and navigation performance in a mobile computing field study. M.S.Eg. thesis, Human Factors Engineering, Wright State University.

Westlake, W. J. (1974). The use of balanced incomplete block designs in comparative bioavailability trials. *Biometrics*, *30*, 319–327.

Wilkie, D. (1962). A method of analysis of mixed level factorial experiments. *Applied Statistics*, *11*, 184–195.

Wood, S. R., & Hartvigsen, D. E. (1964). Statistical design and analysis of qualification test program for a small rocket engine. *Industrial Quality Control*, *20*, 14–18.

Wooding, W. M. (1973). The split-plot design. *Journal of Quality Technology*, *5*, 16–33.

Wu, S. M. (1964). Analysis of rail steel bar welds by two-level factorial design. *Welding Journal Research Supplement*, *43*, 179s–183s. (Reprinted University of Wisconsin Engineering Experiment Station, Reprint 684.)

Xu, H., Phoa, F. K. H., & Wong, W. K. (2009). Recent developments in nonregular fractional factorial designs. *Statistics Surveys*, *3*, 18–46.

Yates, F. (1935). Complex experiments. *Supplement to the Journal of the Royal Statistical Society*, *2*, 181–247. (Reprinted in *Experimental Design* (1970), Charles Griffin and Company, Ltd., London.

Index of Authors

Index of Experiments

Index of Subjects