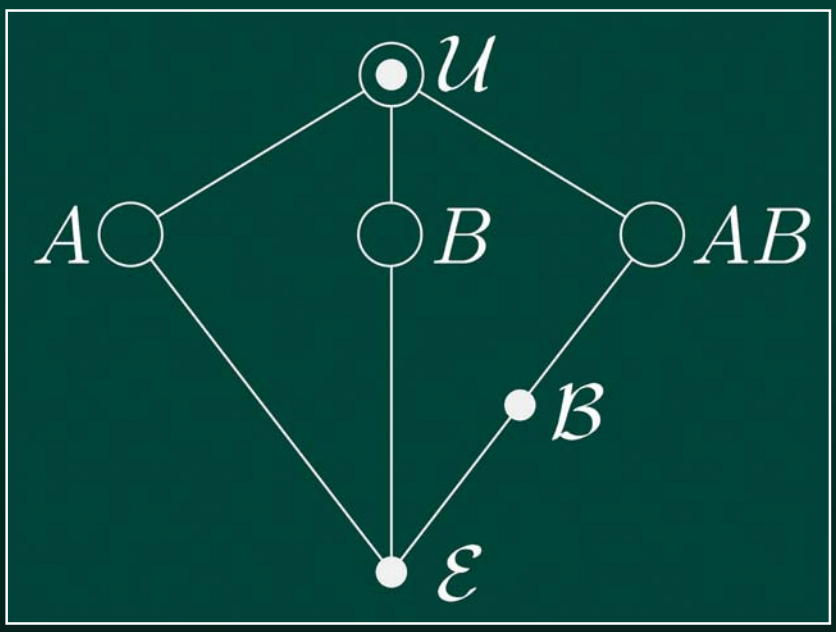
Monographs on Statistics and Applied Probability 131

Theory of Factorial Design

Single- and Multi-Stratum Experiments



Ching-Shui Cheng

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Theory of Factorial Design

Single- and Multi-Stratum Experiments

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

General Editors

F. Bunea, V. Isham, N. Keiding, T. Louis, R. L. Smith, and H. Tong

1. Stochastic Population Models in Ecology and Epidemiology *M.S. Barlett* (1960)
2. Queues *D.R. Cox and W.L. Smith* (1961)
3. Monte Carlo Methods *J.M. Hammersley and D.C. Handscomb* (1964)
4. The Statistical Analysis of Series of Events *D.R. Cox and P.A.W. Lewis* (1966)
5. Population Genetics *W.J. Ewens* (1969)
6. Probability, Statistics and Time *M.S. Barlett* (1975)
7. Statistical Inference *S.D. Silvey* (1975)
8. The Analysis of Contingency Tables *B.S. Everitt* (1977)
9. Multivariate Analysis in Behavioural Research *A.E. Maxwell* (1977)
10. Stochastic Abundance Models *S. Engen* (1978)
11. Some Basic Theory for Statistical Inference *E.J.G. Pitman* (1979)
12. Point Processes *D.R. Cox and V. Isham* (1980)
13. Identification of Outliers *D.M. Hawkins* (1980)
14. Optimal Design *S.D. Silvey* (1980)
15. Finite Mixture Distributions *B.S. Everitt and D.J. Hand* (1981)
16. Classification *A.D. Gordon* (1981)
17. Distribution-Free Statistical Methods, 2nd edition *J.S. Maritz* (1995)
18. Residuals and Influence in Regression *R.D. Cook and S. Weisberg* (1982)
19. Applications of Queueing Theory, 2nd edition *G.F. Newell* (1982)
20. Risk Theory, 3rd edition *R.E. Beard, T. Pentikäinen and E. Pesonen* (1984)
21. Analysis of Survival Data *D.R. Cox and D. Oakes* (1984)
22. An Introduction to Latent Variable Models *B.S. Everitt* (1984)
23. Bandit Problems *D.A. Berry and B. Fristedt* (1985)
24. Stochastic Modelling and Control *M.H.A. Davis and R. Vinter* (1985)
25. The Statistical Analysis of Composition Data *J. Aitchison* (1986)
26. Density Estimation for Statistics and Data Analysis *B.W. Silverman* (1986)
27. Regression Analysis with Applications *G.B. Wetherill* (1986)
28. Sequential Methods in Statistics, 3rd edition *G.B. Wetherill and K.D. Glazebrook* (1986)
29. Tensor Methods in Statistics *P. McCullagh* (1987)
30. Transformation and Weighting in Regression *R.J. Carroll and D. Ruppert* (1988)
31. Asymptotic Techniques for Use in Statistics *O.E. Bandorff-Nielsen and D.R. Cox* (1989)
32. Analysis of Binary Data, 2nd edition *D.R. Cox and E.J. Snell* (1989)
33. Analysis of Infectious Disease Data *N.G. Becker* (1989)
34. Design and Analysis of Cross-Over Trials *B. Jones and M.G. Kenward* (1989)
35. Empirical Bayes Methods, 2nd edition *J.S. Maritz and T. Lwin* (1989)
36. Symmetric Multivariate and Related Distributions *K.T. Fang, S. Kotz and K.W. Ng* (1990)
37. Generalized Linear Models, 2nd edition *P. McCullagh and J.A. Nelder* (1989)
38. Cyclic and Computer Generated Designs, 2nd edition *J.A. John and E.R. Williams* (1995)
39. Analog Estimation Methods in Econometrics *C.F. Manski* (1988)
40. Subset Selection in Regression *A.J. Miller* (1990)
41. Analysis of Repeated Measures *M.J. Crowder and D.J. Hand* (1990)
42. Statistical Reasoning with Imprecise Probabilities *P. Walley* (1991)
43. Generalized Additive Models *T.J. Hastie and R.J. Tibshirani* (1990)
44. Inspection Errors for Attributes in Quality Control *N.L. Johnson, S. Kotz and X. Wu* (1991)
45. The Analysis of Contingency Tables, 2nd edition *B.S. Everitt* (1992)

46. The Analysis of Quantal Response Data *B.J.T. Morgan* (1992)
47. Longitudinal Data with Serial Correlation—A State-Space Approach *R.H. Jones* (1993)
48. Differential Geometry and Statistics *M.K. Murray and J.W. Rice* (1993)
49. Markov Models and Optimization *M.H.A. Davis* (1993)
50. Networks and Chaos—Statistical and Probabilistic Aspects
O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall (1993)
51. Number-Theoretic Methods in Statistics *K.-T. Fang and Y. Wang* (1994)
52. Inference and Asymptotics *O.E. Barndorff-Nielsen and D.R. Cox* (1994)
53. Practical Risk Theory for Actuaries *C.D. Daykin, T. Pentikäinen and M. Pesonen* (1994)
54. Biplots *J.C. Gower and D.J. Hand* (1996)
55. Predictive Inference—An Introduction *S. Geisser* (1993)
56. Model-Free Curve Estimation *M.E. Tarter and M.D. Lock* (1993)
57. An Introduction to the Bootstrap *B. Efron and R.J. Tibshirani* (1993)
58. Nonparametric Regression and Generalized Linear Models *P.J. Green and B.W. Silverman* (1994)
59. Multidimensional Scaling *T.F. Cox and M.A.A. Cox* (1994)
60. Kernel Smoothing *M.P. Wand and M.C. Jones* (1995)
61. Statistics for Long Memory Processes *J. Beran* (1995)
62. Nonlinear Models for Repeated Measurement Data *M. Davidian and D.M. Giltinan* (1995)
63. Measurement Error in Nonlinear Models *R.J. Carroll, D. Rupert and L.A. Stefanski* (1995)
64. Analyzing and Modeling Rank Data *J.J. Marden* (1995)
65. Time Series Models—In Econometrics, Finance and Other Fields
D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen (1996)
66. Local Polynomial Modeling and its Applications *J. Fan and I. Gijbels* (1996)
67. Multivariate Dependencies—Models, Analysis and Interpretation *D.R. Cox and N. Wermuth* (1996)
68. Statistical Inference—Based on the Likelihood *A. Azzalini* (1996)
69. Bayes and Empirical Bayes Methods for Data Analysis *B.P. Carlin and T.A Louis* (1996)
70. Hidden Markov and Other Models for Discrete-Valued Time Series *I.L. MacDonald and W. Zucchini* (1997)
71. Statistical Evidence—A Likelihood Paradigm *R. Royall* (1997)
72. Analysis of Incomplete Multivariate Data *J.L. Schafer* (1997)
73. Multivariate Models and Dependence Concepts *H. Joe* (1997)
74. Theory of Sample Surveys *M.E. Thompson* (1997)
75. Retrial Queues *G. Falin and J.G.C. Templeton* (1997)
76. Theory of Dispersion Models *B. Jørgensen* (1997)
77. Mixed Poisson Processes *J. Grandell* (1997)
78. Variance Components Estimation—Mixed Models, Methodologies and Applications *P.S.R.S. Rao* (1997)
79. Bayesian Methods for Finite Population Sampling *G. Meeden and M. Ghosh* (1997)
80. Stochastic Geometry—Likelihood and computation
O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout (1998)
81. Computer-Assisted Analysis of Mixtures and Applications—Meta-Analysis, Disease Mapping and Others
D. Böhning (1999)
82. Classification, 2nd edition *A.D. Gordon* (1999)
83. Semimartingales and their Statistical Inference *B.L.S. Prakasa Rao* (1999)
84. Statistical Aspects of BSE and vCJD—Models for Epidemics *C.A. Donnelly and N.M. Ferguson* (1999)
85. Set-Indexed Martingales *G. Ivanoff and E. Merzbach* (2000)
86. The Theory of the Design of Experiments *D.R. Cox and N. Reid* (2000)
87. Complex Stochastic Systems *O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg* (2001)
88. Multidimensional Scaling, 2nd edition *T.F. Cox and M.A.A. Cox* (2001)
89. Algebraic Statistics—Computational Commutative Algebra in Statistics
G. Pistone, E. Riccomagno and H.P. Wynn (2001)
90. Analysis of Time Series Structure—SSA and Related Techniques
N. Golyandina, V. Nekrutkin and A.A. Zhigljavsky (2001)
91. Subjective Probability Models for Lifetimes *Fabio Spizzichino* (2001)
92. Empirical Likelihood *Art B. Owen* (2001)

93. Statistics in the 21st Century *Adrian E. Raftery, Martin A. Tanner, and Martin T. Wells* (2001)
94. Accelerated Life Models: Modeling and Statistical Analysis
Vilijandas Bagdonavicius and Mikhail Nikulin (2001)
95. Subset Selection in Regression, Second Edition *Alan Miller* (2002)
96. Topics in Modelling of Clustered Data *Marc Aerts, Helena Geys, Geert Molenberghs, and Louise M. Ryan* (2002)
97. Components of Variance *D.R. Cox and P.J. Solomon* (2002)
98. Design and Analysis of Cross-Over Trials, 2nd Edition *Byron Jones and Michael G. Kenward* (2003)
99. Extreme Values in Finance, Telecommunications, and the Environment
Bärbel Finkenstädt and Holger Rootzén (2003)
100. Statistical Inference and Simulation for Spatial Point Processes
Jesper Møller and Rasmus Plenge Waagepetersen (2004)
101. Hierarchical Modeling and Analysis for Spatial Data
Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand (2004)
102. Diagnostic Checks in Time Series *Wai Keung Li* (2004)
103. Stereology for Statisticians *Adrian Baddeley and Eva B. Vedel Jensen* (2004)
104. Gaussian Markov Random Fields: Theory and Applications *Håvard Rue and Leonhard Held* (2005)
105. Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition
Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu (2006)
106. Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood
Youngjo Lee, John A. Nelder, and Yudi Pawitan (2006)
107. Statistical Methods for Spatio-Temporal Systems
Bärbel Finkenstädt, Leonhard Held, and Valerie Isham (2007)
108. Nonlinear Time Series: Semiparametric and Nonparametric Methods *Jiti Gao* (2007)
109. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis
Michael J. Daniels and Joseph W. Hogan (2008)
110. Hidden Markov Models for Time Series: An Introduction Using R
Walter Zucchini and Iain L. MacDonald (2009)
111. ROC Curves for Continuous Data *Wojtek J. Krzanowski and David J. Hand* (2009)
112. Antedependence Models for Longitudinal Data *Dale L. Zimmerman and Vicente A. Núñez-Antón* (2009)
113. Mixed Effects Models for Complex Data *Lang Wu* (2010)
114. Intoduction to Time Series Modeling *Genshiro Kitagawa* (2010)
115. Expansions and Asymptotics for Statistics *Christopher G. Small* (2010)
116. Statistical Inference: An Integrated Bayesian/Likelihood Approach *Murray Aitkin* (2010)
117. Circular and Linear Regression: Fitting Circles and Lines by Least Squares *Nikolai Chernov* (2010)
118. Simultaneous Inference in Regression *Wei Liu* (2010)
119. Robust Nonparametric Statistical Methods, Second Edition
Thomas P. Hettmansperger and Joseph W. McKean (2011)
120. Statistical Inference: The Minimum Distance Approach
Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park (2011)
121. Smoothing Splines: Methods and Applications *Yuedong Wang* (2011)
122. Extreme Value Methods with Applications to Finance *Serguei Y. Novak* (2012)
123. Dynamic Prediction in Clinical Survival Analysis *Hans C. van Houwelingen and Hein Putter* (2012)
124. Statistical Methods for Stochastic Differential Equations
Mathieu Kessler, Alexander Lindner, and Michael Sørensen (2012)
125. Maximum Likelihood Estimation for Sample Surveys
R. L. Chambers, D. G. Steel, Suojin Wang, and A. H. Welsh (2012)
126. Mean Field Simulation for Monte Carlo Integration *Pierre Del Moral* (2013)
127. Analysis of Variance for Functional Data *Jin-Ting Zhang* (2013)
128. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition *Peter J. Diggle* (2013)
129. Constrained Principal Component Analysis and Related Techniques *Yoshio Takane* (2014)
130. Randomised Response-Adaptive Designs in Clinical Trials *Anthony C. Atkinson and Atanu Biswas* (2014)
131. Theory of Factorial Design: Single- and Multi-Stratum Experiments *Ching-Shui Cheng* (2014)

Monographs on Statistics and Applied Probability 131

Theory of Factorial Design

Single- and Multi-Stratum Experiments

Ching-Shui Cheng

University of California, Berkeley, USA

and

Academia Sinica, Taiwan

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20131017

International Standard Book Number-13: 978-1-4665-0558-2 (eBook - PDF)

Dedicated to the memory of

Leonard Chung-Wei Cheng

Contents

Preface

Factorial designs are widely used in many scientific and industrial investigations. The objective of this book is to provide a rigorous, systematic, and up-to-date treatment of the theoretical aspects of this subject. Despite its long history, research in factorial design has grown considerably in the past two decades. Several books covering these advances are available; nevertheless new discoveries continued to emerge. There are also old useful results that seem to have been overlooked in recent literature and, in my view, deserve to be better known.

Factorial experiments with multiple error terms (strata) that result from complicated structures of experimental units are common in agriculture. In recent years, the design of such experiments also received much attention in industrial applications. A theory of orthogonal block structures that goes back to John Nelder provides a unifying framework for the design and analysis of multi-stratum experiments. One feature of the present book is to present this elegant and general theory which, once understood, is simple to use, and can be applied to various structures of experimental units in a unified and systematic way. The mathematics required to understand this theory is perhaps what, in Rosemary Bailey's words, "obscured the essential simplicity" of the theory. In this book, I tried to minimize the mathematics needed, and did not present the theory in the most general form as developed by Bailey and her coauthors. To prepare readers for the general theory, a unified treatment of some simple designs such as completely randomized designs, block designs, and row-column designs is presented first. Therefore the book also covers these elementary non-factorial-design topics. It is suitable as a reference book for researchers and as a textbook for graduate students who have taken a first course in the design of experiments. Since the book is self-contained and includes many examples, it should also be accessible to readers with minimal previous exposure to experimental design as long as they have good mathematical and statistical backgrounds. Readers are required to be familiar with linear algebra. A review of linear model theory is given in Chapter 2, and a brief survey of some basic algebraic results on finite groups and fields can be found in the Appendix. Sections that can be skipped, at least on the first reading, without affecting the understanding of the material in later parts of the book are marked with stars.

In addition to a general theory of multi-stratum factorial design, the book covers many other topics and results that have not been reported in books. These include, among others, the useful method of design key for constructing multi-stratum factorial designs, the methods of partial foldover and doubling for constructing two-level resolution IV designs, some results on the structures of two-level resolution IV designs taken from the literature of projective geometry, the extension of minimum

aberration to nonregular designs, and the minimum moment aberration criterion, which is equivalent to minimum aberration.

The book does not devote much space to the analysis of factorial designs due to its theoretical nature, and also because excellent treatment of strategies for data analysis can be found in several more applied books. Another subject that does not receive a full treatment is the so-called nonregular designs. It is touched upon in Chapter 8 when orthogonal arrays are introduced, and some selected topics are surveyed in Chapter 15. The research on nonregular designs is still very active and expands rapidly. It deserves another volume.

The writing of this book originated from a ten-lecture workshop on "Recent Developments in Factorial Design" I gave in June 2002 at the Institute of Statistical Science, Academia Sinica, in Taiwan. I thank Chen-Hsin Chen, Director of the institute at the time, for his invitation. The book was written over a long period of time while I taught at the University of California, Berkeley, and also during visits to the National Center for Theoretical Sciences in Hsinchu, Taiwan, and the Issac Newton Institute for Mathematical Sciences in Cambridge, United Kingdom. The support of these institutions and the US National Science Foundation is acknowledged. The book could not have been completed without the help of many people. It contains results from joint works with Rosemary Bailey, Dursun Bulutoglu, Hegang Chen, Lih-Yuan Deng, Mike Jacroux, Bobby Mee, Rahull Mukerjee, Nam-Ky Nguyen, David Steinberg, Don Sun, Boxin Tang, Pi-Wen Tsai, Hongquan Xu, and Oksoun Yee. I had the privilege of working with them. I also had the fortune to know Rosemary Bailey early in my career. Her work has had a great impact on me, and this book uses the framework she had developed. Boxin Tang read the entire book, and both Rosemary Bailey and Don Ylvisaker read more than half of it. They provided numerous detailed and very helpful comments as well as pointing out many errors. Hegang Chen, Chen-Tuo Liao, and Hongquan Xu helped check the accuracy of some parts of the book. As a LaTex novice, I am very grateful to Pi-Wen Tsai for her help whenever I ran into problems with LaTex. She also read and commented on earlier versions of several chapters. Yu-Ting Chen and Chiun-How Kao helped fix some figures. I would also like to acknowledge our daughter Adelaide for her endearing love and support as well as her upbeat reminder to always see the bright side. Last but not least, I am most grateful to my wife Suzanne Pan for her thankless support and care over the years and for patiently reading this "Tian Shu" from cover to cover.

Additional material for the book will be maintained at http://www.crcpress.com/product/isbn/9781466505575/ and http://www.stat.sinica.edu.tw/factorial-design/.

Chapter 1

Introduction

Many of the fundamental ideas and principles of experimental design were developed by Sir R. A. Fisher at the Rothamsted Experimental Station (Fisher, 1926). This agricultural background is reflected in some terminology of experimental design that is still being used today. Agricultural experiments are conducted, e.g., to compare different varieties of a certain crop or different fertilizers. In general, those that are under comparison in an experiment are called *treatments*. Manufacturing processes in industrial experiments and drugs in pharmaceutical studies are examples of treatments. In an agricultural experiment, the varieties or fertilizers are assigned to plots, and the yields are compared after harvesting. Each plot is called an *experimental unit* (or unit). In general, an experimental unit can be defined as the smallest division of the experimental material such that different units may receive different treatments (Cox, 1958, p. 2). At the design stage, a treatment is chosen for each experimental unit.

One fundamental difficulty in such *comparative experiments* is inherent variability of the experimental units. No two plots have exactly the same soil quality, and there are other variations beyond the experimenter's control such as weather conditions. Consequently, effects of the treatments may be biased by uncontrolled variations. A solution is to assign the treatments randomly to the units. In addition to guarding against potential systematic biases, *randomization* also provides a basis for appropriate statistical analysis.

The simplest kind of randomized experiment is one in which treatments are assigned to units completely at random. In a *completely randomized experiment*, the precision of a treatment comparison depends on the overall variability of the experimental units. When the experimental units are highly variable, the treatment comparisons do not have good precision. In this case, the method of *blocking* is an effective way to reduce experimental error. The idea is to divide the experimental units into more homogeneous groups called blocks. When the treatments are compared on the units within each block, the precision is improved since it depends on the smaller within-block variability.

Suppose the experimental units are grouped into b blocks of size k. Even though efforts are made for the units in the same block to be as alike as possible, they are still not the same. Given an initial assignment of the treatments to the bk unit labels based on statistical, practical and/or other considerations, randomization is carried

out by randomly permuting the unit labels within each block (done independently from block to block), and also randomly permuting the block labels. The additional step of randomly permuting block labels is to assure that an observation intended on a given treatment is equally likely to occur at any of the experimental units.

Under a completely randomized experiment, the experimental units are considered to be unstructured. The structure of the experimental units under a block design is an example of *nesting*. Suppose there are b blocks each consisting of k units; then each experimental unit can be labeled by a pair (i, j), $i = 1, \ldots, b$, $j = 1, \ldots, k$. This involves two factors with b and k levels, respectively. Here if $i \neq i'$, unit (i, j) bears no relation to unit (i', j); indeed, within-block randomization renders positions of the units in each block immaterial. We say that the k-level factor is nested in the b-level factor, and denote this structure by b/k or block/unit if the two factors involved are named "block" and "unit," respectively.

Another commonly encountered structure of the experimental units involves two blocking factors. For example, in agricultural experiments the plots may be laid out in rows and columns, and we try to eliminate from the treatment comparisons the spatial variations due to row-to-row and column-to-column differences. In experiments that are carried out on several different days and in several different time slots on each day, the observed responses might be affected by day-to-day and time-to-time variations. In this case each experimental run can be represented by a cell of a rectangular grid with those corresponding to experimental runs taking place on the same day (respectively, in the same time slot) falling in the same row (respectively, the same column). In general, suppose rc experimental units can be arranged in r rows and c columns such that any two units in the same row have a definite relation, and so do those in the same column. Then we have an example of *crossing*. This structure of experimental units is denoted by $r \times c$ or row $\times$ column if the two factors involved are named "row" and "column," respectively. In such a *row-column experiment*, given an initial assignment of the treatments to the rc unit labels, randomization is carried out by randomly permuting the row labels and, independently, randomly permuting the column labels. This assures that the structure of the experimental units is preserved: two treatments originally assigned to the same row (respectively, column) remain in the same row (respectively, column) after randomization.

For example, suppose there are four different manufacturing processes compared in four time slots on each of four days. With the days represented by rows and times represented by columns, a possible design is

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

where the four numbers 1, 2, 3, and 4 are labels of the treatments assigned to the units represented by the 16 row-column combinations. We see that each of the four numbers appears once in each row and once in each column. Under such a design, called a *Latin square*, all the treatments can be compared on each of the four days

as well as in each of the four time slots. If the random permutation is such that the first, second, third, and fourth rows of the Latin square displayed above are mapped to the first, fourth, second, and third rows, respectively, and the first, second, third, and fourth columns are mapped to the fourth, third, first, and second columns, respectively, then it results in the following Latin square to be used in actual experimentation.

3	4	2	1
1	2	4	3
2	1	3	4
4	3	1	2

The structures of experimental units are called *block structures*. Block and row-column designs are based on the two simplest block structures involving nesting and crossing, respectively. Nelder (1965a) defined *simple block structures* to be those that can be obtained by iterations of nesting ($/$) and crossing ($\times$) operators. For example, $n_1/(n_2 \times n_3)$ represents the block structure under a *nested row-column design*, where $n_1 n_2 n_3$ experimental units are grouped into n_1 blocks of size $n_2 n_3$, and within each block the $n_2 n_3$ units are arranged in n_2 rows and n_3 columns. Randomization of such an experiment can be done by randomly permuting the block labels and carrying out the appropriate randomization for the block structure $n_2 \times n_3$ within each block, that is, randomly permuting the row labels and column labels separately. In an experiment with the block structure $n_1/(n_2/n_3)$, $n_1 n_2 n_3$ experimental units are grouped into n_1 blocks, and within each block the $n_2 n_3$ units are further grouped into n_2 smaller blocks (often called *whole-plots*) of n_3 units (often called *subplots*). To randomize such a *blocked split-plot experiment*, we randomly permute the block labels and carry out the appropriate randomization for the block structure n_2/n_3 within each block, that is, randomly permute the whole-plot labels within each block, and randomly permute the subplot labels within each whole-plot. Note that $n_1/(n_2/n_3)$ is the same as $(n_1/n_2)/n_3$.

In general, randomization of an experiment with a simple block structure is carried out according to the appropriate randomization for nesting or crossing at each stage of the block structure formula.

Like experimental units, the treatments may also have a structure. One can compare treatments by examining the pairwise differences of treatment effects. When the treatments do not have a structure (for example, when they are different varieties of a crop), one may be equally interested in all the pairwise comparisons of treatment effects. However, if they do have a certain structure, then some comparisons may be more important than others. For example, suppose one of the treatments is a control. Then one may be more interested in the comparisons between the control and new treatments.

In this book, treatments are to have a *factorial structure*: each treatment is a combination of multiple factors (variables) called *treatment factors*. Suppose there are n treatment factors and the ith factor has s_i values or settings to be studied. Each of these values or settings is called a *level*. The treatments, also called *treatment com-*

binations in this context, consist of all $s_1 \cdots s_n$ possible combinations of the factor levels. The experiment is called an $s_1 \times \cdots \times s_n$ factorial experiment, and is called an s^n experiment when $s_1 = \cdots = s_n = s$. For example, a fertilizer may be a combination of the levels of three factors N (nitrogen), P (phosphate), and K (potash), and a chemical process might involve temperature, pressure, concentration of a catalyst, etc. Fisher (1926) introduced factorial design to agricultural experiments, and Yates (1935, 1937) made significant contributions to its early development.

When the treatments have a factorial structure, typically we are interested in the effects of individual factors, as well as how the factors interact with one another. Special functions of the treatment effects, called *main effects* and *interactions*, can be defined to represent such effects of interest. We say that the treatment factors do not interact if, when the levels of a factor are changed while those of the other factors are kept constant, the changes in the treatment effects only depend on the levels of the varying factor. In this case, we can separate the effects of individual factors, and the effect of each treatment combination can be obtained by summing up these individual effects. Under such *additivity* of the treatment factors, for example, to determine the combination of N, P, and K with the highest average yield, one can simply find the best level of each of the three factors separately. Otherwise, the factors need to be considered simultaneously. Roughly speaking, the main effect of a treatment factor is its effects averaged over the levels of the other factors, and the interaction effects measure departures from additivity. Precise definitions of these effects, collectively called *factorial effects*, will be given in Chapter 6.

A factorial experiment with each treatment combination observed once is called a *complete* factorial experiment. We also refer to it as a single-replicate complete factorial experiment. The analysis of completely randomized experiments in which each treatment combination is observed the same number of times, to be presented in Chapter 6, is straightforward. It becomes more involved if the experimental units have a more complicated block structure and/or if not all the treatment combinations can be observed.

When a factorial experiment is blocked, with each block consisting of one replicate of all the treatment combinations, the analysis is still very simple. As will be discussed in Chapter 6, in this case all the treatment main effects and interactions can be estimated in the same way as if there were no blocking, except that the variances of these estimators depend on the within-block variability instead of the overall variability of the experimental units. Since the total number of treatment combinations increases rapidly as the number of factors becomes large, a design that accommodates all the treatment combinations in each block requires large blocks whose homogeneity is difficult to control. In order to achieve smaller within-block variability, we cannot accommodate all the treatment combinations in the same block and must use *incomplete blocks*. It may also be impractical to carry out experiments in large blocks. Then, since not all the treatment combinations appear in the same block, the estimates of some treatment factorial effects cannot be based on within-block comparisons alone. This may result in less precision for such estimates. For example, suppose an experiment on two two-level factors A_1 and A_2 is to be run on two different days with the two combinations $(0,0)$ and $(1,1)$ of the levels of A_1 and

A_2 observed on one day, and the other two combinations $(0,1)$ and $(1,0)$ observed on the other day, where 0 and 1 are the two factor levels. Then estimates of the main effect (comparison of the two levels) of factor A_1 and the main effect of A_2 are based on within-block comparisons, but as will be seen in Chapter 7, the interaction of the two factors would have to be estimated by comparing the observations on the first day with those on the second day, resulting in less precision. We say that this two-factor interaction is *confounded* with blocks.

When a factorial experiment must be run in incomplete blocks, we choose a design in such a way that only those factorial effects that are less important or are known to be negligible are confounded with blocks. Typically the main effects are deemed more important, and one would avoid confounding them with blocks. However, due to practical constraints, sometimes one must confound certain main effects with blocks. For instance, it may be difficult to change the levels of some factors. In the aforementioned example, if a factor must be kept at the same level on each day, then the main effect of that factor can only be estimated by a more variable between-day comparison.

Often the number of treatment combinations is so large that it is practically possible to observe only a small subset of the treatment combinations. This is called a *fractional factorial design*. Then, since not all the treatment combinations are observed, some factorial effects are mixed up and cannot be distinguished. We say that they are *aliased*. For example, when only 16 treatment combinations are to be observed in an experiment involving six two-level factors, there are 63 factorial effects (6 main effects, 15 two-factor interactions, 20 three-factor interactions, 15 four-factor interactions, 6 five-factor interactions, and 1 six-factor interaction), but only 15 degrees of freedom are available for estimating them. This is possible if many of the factorial effects are negligible. One design issue is which 16 of the 64 treatment combinations are to be selected.

An important property of a fractional factorial design, called *resolution*, pertains to the extent to which the lower-order effects are mixed up with higher-order effects. For example, under a design of resolution III, no main effect is aliased with other main effects, but some main effects are aliased with two-factor interactions; under a design of resolution IV, no main effect is aliased with other main effects or two-factor interactions, but some two-factor interactions are aliased with other two-factor interactions; under a design of resolution V, no main effects and two-factor interactions are aliased with one another. When the experimenter has little knowledge about the relative importance of the factorial effects, it is common to assume that the lower-order effects are more important than higher-order effects (the main effects are more important than interactions, and two-factor interactions are more important than three-factor interactions, etc.), and effects of the same order are equally important. Under such a *hierarchical assumption*, it is desirable to have a design with high resolution. A popular criterion of selecting fractional factorial designs and a refinement of maximum resolution, called *minimum aberration*, is based on the idea of minimizing the aliasing among the more important lower-order effects.

When the experimental units have a certain block structure, in addition to picking a fraction of the treatment combinations, we also have to decide how to assign

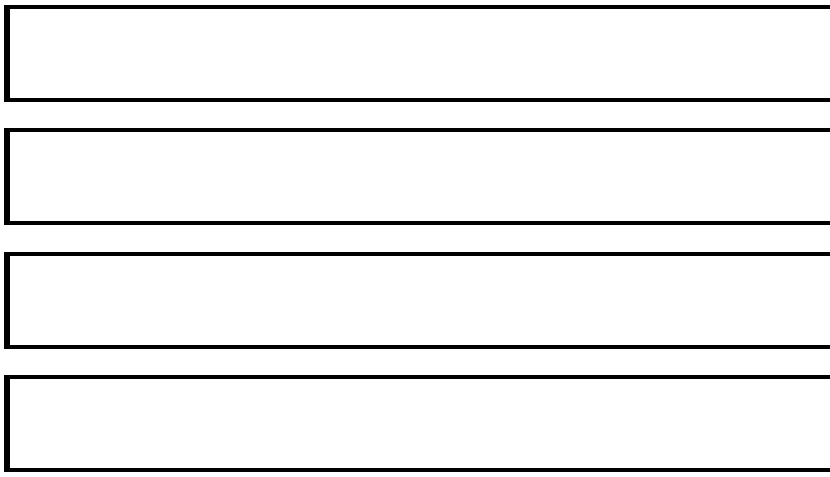
the selected treatment combinations to the units. In highly fractionated factorial experiments with complicated block structures, we have complex aliasing of treatment factorial effects as well as multiple levels of precision for their estimates. The bulk of this book is about the study of such designs, including their analysis, selection, and construction. The term "multi-stratum" in the subtitle of the book refers to multiple sources of errors that arise from complicated block structures, while "single-stratum" is synonymous with "complete randomization" where there is one single error term.

Treatment and block structures are two important components of a randomized experiment. Nelder (1965a,b) emphasized their distinction and developed a theory for the analysis of randomized experiments with simple block structures. Simple block structures cover most, albeit not all the block structures that are commonly encountered in practice. Speed and Bailey (1982) and Tjur (1984) further developed the theory to cover the more general *orthogonal block structures*. This theory, an account of which can be found in Bailey (2008), provides the basis for the approach adopted in this book.

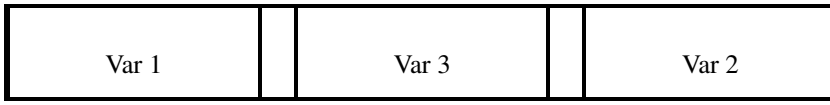
We turn to five examples of factorial experiments to motivate some of the topics to be discussed in the book. The first three examples involve simple block structures. The block structures in Examples 1.4 and 1.5 are not simple block structures, but the theory developed by Speed and Bailey (1982) and Tjur (1984) is applicable. We will return to these examples from time to time in later chapters to illustrate applications of the theory as it is developed.

Our first example is a replicated complete factorial experiment with a relatively complicated block structure.

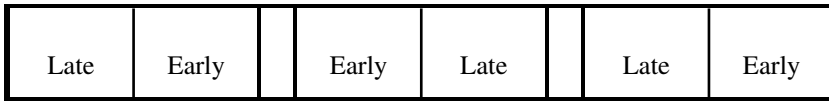
Example 1.1. Loughin (2005) studied the design of an experiment on weed control. Herbicides can kill the weeds that reduce soybean yields, but they can also kill soybeans. On the other hand, soybean varieties can be bred or engineered to be resistant to certain herbicides. An experiment is to be carried out to study what factors influence weed control and yield of genetically altered soybean varieties. Four factors studied in the experiment are *soybean variety/herbicide combinations* in which the herbicide is safe for the soybean variety, *dates* and *rates* of herbicide application, and *weed species*. There are three variety/herbicide combinations, two dates (early and late), three rates (1/4, 1/2, and 1), and seven weed species, giving a total of 126 treatments with a $3 \times 2 \times 3 \times 7$ factorial structure. Soybeans and weeds are planted together and a herbicide safe for the soybean variety is sprayed at the designated time and rate. Then weed properties (numbers, density, mass) and soybean yields are measured. However, there are some practical constraints on how the experiment can be run. Due to herbicide drift, different varieties cannot be planted too close together and buffer zones between varieties are needed, but the field size is not large enough to allow for 126 plots of adequate size with large buffers between each pair of adjacent plots. Therefore, for efficient use of space, one needs to plant all of a given soybean variety contiguously so that fewer buffers are needed. Additional drift concerns lead to a design described as follows. First the field is divided into four blocks to accommodate four replications:



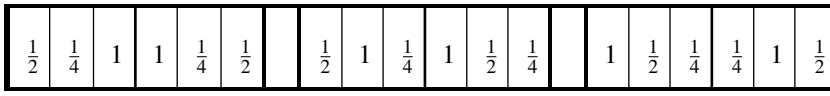
Each block is split into three plots with two buffer zones, and the variety/herbicide combinations are randomly assigned to the plots within blocks:



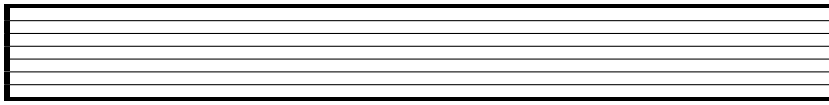
Each plot is then split into two subplots, with application times randomly assigned to subplots within plots:



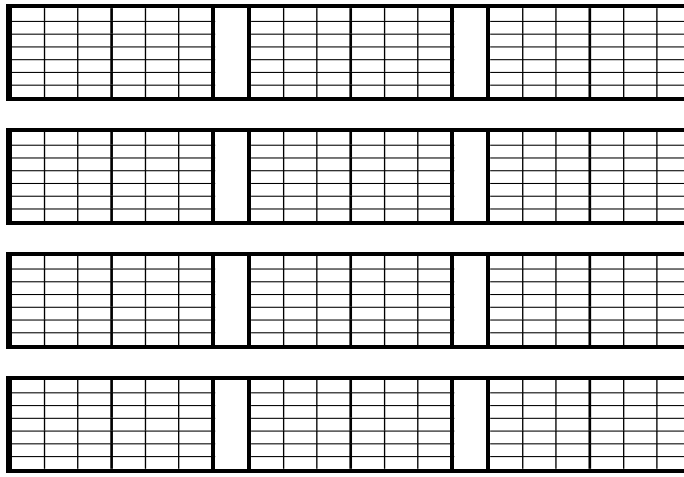
Furthermore each subplot is split into three sub-subplots, with application rates randomly assigned to sub-subplots within subplots:



Each block is divided into seven horizontal strips, with the weed species randomly assigned to the strips within blocks:



We end up with 504 combinations of sub-subplots and strips:



Each of the 126 treatment combinations appears once in each of the four blocks. To summarize, we have four replicates of a complete $3 \times 2 \times 3 \times 7$ factorial experiment with the block structure $4/[(3/2/3) \times 7]$. Both subplots in the same plot are assigned the same variety/herbicide combination, all the sub-subplots in the same subplot are assigned the same herbicide application time, and all the sub-subplot and strip intersections in the same strip are assigned the same weed species. Various aspects of the analysis of this design will be discussed in Sections 12.1, 12.9, 12.10, and 13.10.

In Example 1.1, there are 18 sub-subplots in each block. If different soybean varieties were to be assigned to neighboring sub-subplots, then 17 buffer zones would be needed in each block. With only two buffer zones per block under the proposed design, comparisons of soybean varieties are based on between-plot comparisons, which are expected to be more variable than those between subplots and sub-subplots. The precision of the estimates of such effects is sacrificed in order to satisfy the practical constraints.

Example 1.2. McLeod and Brewster (2004) discussed an experiment for identifying key factors that would affect the quality of a chrome-plating process. Suppose six two-level treatment factors are to be considered in the experiment: A, chrome concentration; B, chrome to sulfate ratio; C, bath temperature; S, etching current density; T, plating current density; and U, part geometry. The response variables include, e.g., the numbers of pits and cracks. The chrome plating is done in a bath (tank), which contains several rectifiers, but only two will be used. On any given day the levels of A, B, and C cannot be changed since they represent characteristics of the bath. On the other hand, the levels of factors S, T, and U can be changed at the rectifier level. The experiment is to be run on 16 days, with four days in each of four weeks. Therefore there are a total of 32 runs with the block structure (4 weeks)/(4 days)/(2 runs), and

one has to choose 32 out of the $2^6 = 64$ treatment combinations. Weeks, days, and runs can be considered as blocks, whole-plots, and subplots, respectively. The three factors A, B, and C must have constant levels on the two experimental runs on the same day, and are called whole-plot treatment factors. The other three factors S, T, and U are not subject to this constraint and are called subplot treatment factors. We will return to this example in Sections 12.9, 13.2, 13.4, 13.5, 13.7, 14.4, and 14.13.

Example 1.3. Miller (1997) described a laundry experiment for investigating methods of reducing the wrinkling of clothes. Suppose the experiment is to be run in two blocks, with four washers and four dryers to be used. After four cloth samples have been washed in each washer, the 16 samples are divided into four groups with each group containing one sample from each washer. Each of these groups is then assigned to one dryer. The extent of wrinkling on each sample is evaluated at the end of the experiment. This results in 32 experimental runs that can be thought to have the $2/(4 \times 4)$ block structure shown in Figure 1.1, where each cell represents a cloth sample, rows represent sets of samples that are washed together, and columns represent sets of samples that are dried together. There are ten two-level treatment factors,

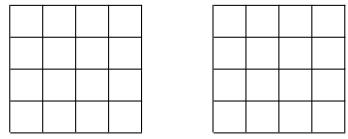


Figure 1.1 *A* $2/(4 \times 4)$ *block structure*

six of which (A, B, C, D, E, F) are configurations of washers and four (S, T, U, V) are configurations of dryers. One has to choose 32 out of the $2^{10} = 1024$ treatment combinations. Furthermore, since the experimental runs on the cloth samples in the same row are conducted in the same washing cycle, each of A, B, C, D, E, F must have a constant level in each row. Likewise, each of S, T, U, V must have a constant level in each column. Thus in each block, four combinations of the levels of A, B, C, D, E, F are chosen, one for each row, and four combinations of the levels of S, T, U, V are chosen, one for each column. The four combinations of washer settings are then coupled with the four combinations of dryer settings to form 16 treatment combinations of the ten treatment factors in the same block. An experiment run in this way requires only four washer loads and four dryer loads in each block. If one were to do complete randomization in each block, then four washer loads and four dryer loads could produce only four observations. The trade-off is that the main effect of each treatment factor is confounded with either rows or columns. Construction and analysis of designs for such *blocked strip-plot experiments* will be discussed in Sections 12.2, 12.9, 12.10, 13.3, 13.4, 13.5, 13.6, 13.7, 14.5, 14.6, and 14.14.

Federer and King (2006) gave a comprehensive treatment of split-plot and strip-

plot designs and their many variations. In this book, we present a unifying theory that can be systematically applied to a very general class of multi-stratum experiments.

Example 1.3 is an experiment with two processing stages: washing and drying. Many industrial experiments involve a sequence of processing stages, with the levels of various treatment factors assigned and processed at different stages. At each stage the experimental units are partitioned into disjoint classes. Those in the same class, which will be processed together, are assigned the same level of each of the treatment factors that are to be processed at that stage. We call the treatment factors processed at the ith stage the ith-stage treatment factors. In Example 1.3, levels of the six washer factors are set at the first stage and those of the four dryer factors are set at the second stage. So the washer configurations are first-stage treatment factors and the dryer configurations are second-stage treatment factors. Such an experiment with two processing stages can be thought to have experimental units with a row-column structure.

In Examples 1.1–1.3, the experimental units can be represented by *all* the level combinations of some *unit factors*. In the next two examples, we present experiments in which the experimental units are a *fraction* of unit-factor level combinations.

Example 1.4. Mee and Bates (1998) discussed designs of experiments with multiple processing stages in the fabrication of integrated circuits. Suppose that at the first stage 16 batches of material are divided into four groups of equal size, with the same level of each first-stage treatment factor assigned to all the batches in the same group. At the second stage they are rearranged into another four groups of equal size, again with the same level of each second-stage treatment factor assigned to all the batches in the same group. As in Example 1.3, the groupings at the two stages can be represented by rows and columns. Then each of the first-stage groups and each of the second-stage groups have exactly one batch in common. This is a desirable property whose advantage will be explained in Section 12.13. Now suppose there is a third stage. Then we need a third grouping of the batches. One possibility is to group according to the numbers in the Latin square shown earlier:

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

One can assign the same level of each third-stage treatment factor to all the units (batches) corresponding to the same number in the Latin square. One advantage is that each of the third-stage groups has exactly one unit in common with any group at the first two stages. If a fourth stage is needed, then one may group according to the following Latin square:

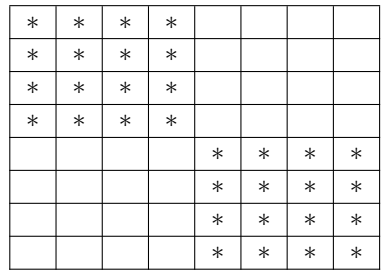
1	2	3	4
4	3	2	1
2	1	4	3
3	4	1	2

This and the previous Latin square have the property that when one is superimposed on the other, each of the 16 pairs of numbers (i, j), $1 \leq i,\ j \leq 4$, appears in exactly one cell. We say that these two Latin squares are *orthogonal* to each other. If the fourth-stage grouping is done according to the numbers in the second Latin square, then each of the fourth-stage groups also has exactly one unit in common with each group at any of the first three stages. This kind of block structure cannot be obtained by iterations of nesting and crossing operators. To be a simple block structure, with four groups at each of three or four stages, one would need $4^3 = 64$ or $4^4 = 256$ units, respectively. Thus the 16 units can be regarded as a quarter or one-sixteenth fraction of the combinations of three or four 4-level factors, respectively. The following is a complete 2^4 design which can be used for experiments in which the levels of the four treatment factors are set at four stages, one factor per stage: the first factor has a constant level in each row, the second factor has a constant level in each column, the third factor has a constant level in each cell occupied by the same number in the first Latin square, and the fourth factor has a constant level in each cell occupied by the same number in the second Latin square.

0000	0011	0101	0110
0010	0001	0111	0100
1001	1010	1100	1111
1011	1000	1110	1101

We will return to this example in Sections 12.5, 12.10, 12.13, and 13.11.

Example 1.5. Bingham, Sitter, Kelly, Moore, and Olivas (2008) discussed experiments with multiple processing stages where more groups are needed at each stage, which makes it impossible for all the groups at different stages to share common units. For example, in an experiment with two processing stages, suppose 32 experimental units are to be partitioned into 8 groups of size 4 at each of the two stages. One possibility is to partition the 32 units as in Figure 1.1. The eight rows of size 4, four of which from each of the two blocks, together constitute the eight first-stage groups, and the eight columns in the two blocks together constitute the eight second-stage groups. As shown in the following figure, the 32 starred experimental units are a *fraction* of the 64 units in a completely crossed 8×8 square.



As in Example 1.4, the 32 units do not have a simple block structure. An important difference, however, is that in the current setting not all the first-stage groups can meet with every second-stage group, causing some complications in the design and analysis (to be discussed in Section 13.12). The figure shows that the 32 units are divided into two groups of size 16, which we call *pseudo* blocks since they are not part of the originally intended block structure. We will revisit this example in Sections 12.13 and 13.12, and show that it can be treated as an experiment with the block structure $2/(4 \times 4)$. A similar problem was studied in Vivacqua and Bisgaard (2009), to be discussed in Section 14.7.

An overview

Some introductory material is presented in Chapters 2–5. Chapter 2 is a review of some results on linear models, with emphasis on one-way and two-way layout models and a geometric characterization of the condition of proportional frequencies between two factors. Under the assumption of treatment-unit additivity, randomization models are developed in Chapter 3 for some designs with simple block structures, including block designs and row-column designs. In Chapter 4, the condition of proportional frequencies is extended to a notion of orthogonal factors that plays an important role in the block structures studied in this book. Some mathematical results on factors that are needed throughout the book are also gathered there. A condition that entails simple analysis of a randomized design (Theorem 5.1) is established in Chapter 5. This result is used to present a unified treatment of the analyses of three classes of orthogonal designs (completely randomized designs, complete block designs, and Latin square designs) under the randomization models derived in Chapter 3. It is also a key result for developing, in later chapters, a general theory of orthogonal designs for experiments with more complicated block structures.

The treatment factorial structure is introduced in Chapter 6. It is shown how certain special functions of the treatment effects can be defined to represent main effects and interactions of the treatment factors. Unless all the treatment factors have two levels, the choices of such functions are not unique. Several methods of constructing them based on orthogonal polynomials, finite Euclidean geometry, and Abelian groups are presented. The discussion of complete factorial designs is continued in Chapter 7, for experiments that are conducted in incomplete blocks or row-column

layouts, including split-plot and strip-plot designs. In this case, there is more than one error term and some factorial effects are confounded with blocks, rows, or columns. A construction method based on design keys is presented in addition to a commonly used method, and is shown to enjoy several advantages.

Fractional factorial designs under complete randomization are treated in Chapters 8–11. In Chapter 8, the important combinatorial structure of orthogonal arrays is introduced. Some basic properties of orthogonal arrays as fractional factorial designs, including upper bounds on the number of factors that can be accommodated by orthogonal arrays of a given run size, are derived. We also present several methods of constructing orthogonal arrays, in particular the foldover method and the construction via difference matrices. The chapter is concluded with a brief discussion of applications of orthogonal arrays to computer experiments and three variants of orthogonal arrays recently introduced for this purpose. The emphasis of this book is mainly on the so-called regular fractional factorial designs, which are easy to construct and analyze, and have nice structures and a rich theory. In Chapter 9 we provide a treatment of their basics, including design construction, aliasing and estimability of factorial effects, resolution, a search algorithm for finding designs under which some required effects can be estimated, and the connection with the linear codes of coding theory. The criterion of minimum aberration and some related criteria for selecting regular fractional factorial designs are discussed in Chapter 10. The statistical meaning of minimum aberration is clarified via its implications on the aliasing pattern of factorial effects. It is shown that this criterion produces designs with good properties under model uncertainty and good lower-dimensional projections. The connection to coding theory provides two powerful tools for constructing minimum aberration designs: the MacWilliams identities can be used to establish a complementary design theory that is useful for determining minimum aberration designs when there are many factors; the Pless power moment identities lead to the criterion of minimum moment aberration, which is equivalent to minimum aberration. Besides the theoretical interest, this equivalence is useful for analytical characterization and algorithmic construction of minimum aberration designs. A Bayesian approach to the design and analysis of factorial experiments, also applicable to nonregular designs, is presented at the end of Chapter 10. Regular designs are also closely related to finite projective geometries. The connection is made in two optional sections in Chapter 9, and is used to characterize and construct minimum aberration designs in Chapter 10. The geometric connection culminates in an elegant theory of the construction and structures of resolution IV designs in Chapter 11. While foldover is a well-known method of constructing resolution IV designs, many resolution IV designs cannot be constructed by this method. We translate the geometric results into design language, and among other topics, present the methods of doubling and partial foldover for constructing them.

In Chapters 12–14, we turn to factorial designs with more complicated block structures called multi-stratum designs. Some basic results on Nelder's simple block structures and the more general orthogonal block structures are derived in Chapter 12. A general theory for the design and analysis of orthogonal multi-stratum complete factorial designs is developed in Chapter 13. This theory is applied to several

settings, including blocked split-plot designs, blocked strip-plot designs, and design of experiments with multiple processing stages. Chapter 14 is devoted to the construction of multi-stratum fractional factorial designs and criteria for their selection under model uncertainty in the spirit of minimum aberration. The five motivating examples presented above are revisited.

We survey a few nonregular design topics in Chapter 15. Under nonregular designs, the factorial effects are aliased in a complicated way, but their run sizes are more flexible than regular designs. At the initial stage of experimentation, often only a small number of the potential factors are important. Due to their run-size economy, nonregular designs are suitable for conducting factor screening experiments under the *factor sparsity* principle. In this context, it is useful to study the property of the design when it is projected onto small subsets of factors. We also discuss the relevant topics of search designs and supersaturated designs. The objective of a search design is to identify and discriminate nonnegligible effects under the assumption that the number of nonnegligible effects is small. Supersaturated designs have more unknown parameters than the degrees of freedom available for estimating them and are useful for screening active factors. In addition to these and other miscellaneous topics, we show how some of the results presented in earlier chapters can be extended to nonregular designs. For example, coding theory again proves useful for providing a way to extend minimum aberration to nonregular designs.

Throughout this book, the starred sections can be skipped, at least on the first reading. Relevant exercises are also marked with stars.

Chapter 2

Linear Model Basics

In this chapter we review some basic results from linear model theory, including least squares estimation, the Gauss–Markov Theorem, and tests of linear hypotheses, with applications to the analysis of fixed-effect one-way and additive two-way layout models. We show how the analysis of an additive two-way layout model is simplified when certain vector spaces associated with the two factors are orthogonal. This geometric condition is shown to be equivalent to that the two factors satisfy the condition of proportional frequencies.

2.1 Least squares

Consider the linear model

$$y_i = \sum_{j=1}^{p} x_{ij}\theta_j + \varepsilon_i,\ i = 1, \ldots, N, \tag{2.1}$$

where $x_{ij}, 1 \le i \le N, 1 \le j \le p$, are known constants, $\theta_1, \ldots, \theta_p$ are unknown parameters, and $\varepsilon_1, \ldots, \varepsilon_N$ are uncorrelated random variables with zero mean and common variance σ^2. Let $\mathbf{y} = (y_1, \ldots, y_N)^T$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^T$, and $\mathbf{X}$ be the $N \times p$ matrix with the (i, j)th entry equal to x_{ij}, where T stands for "transpose." Then (2.1) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{2.2}$$

with

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0},\ \text{and}\ \mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_N, \tag{2.3}$$

where $\mathrm{E}(\boldsymbol{\varepsilon}) = (\mathrm{E}(\varepsilon_1), \ldots, \mathrm{E}(\varepsilon_N))^T$, $\mathrm{cov}(\boldsymbol{\varepsilon})$ is the covariance matrix of $\boldsymbol{\varepsilon}$, $\mathbf{0}$ is the vector of zeros, and $\mathbf{I}_N$ is the identity matrix of order N. We call $\mathbf{X}$ the *model matrix*.

Least squares estimators $\widehat{\theta}_1, \ldots, \widehat{\theta}_p$ of $\theta_1, \ldots, \theta_p$ are obtained by minimizing $\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}\theta_j)^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$.

Let $\mathrm{E}(\mathbf{y}) = (\mathrm{E}(y_1), \ldots, \mathrm{E}(y_N))^T$. Then under (2.2) and (2.3), $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\theta}$. This implies that $\mathrm{E}(\mathbf{y})$ is a linear combination of the column vectors of $\mathbf{X}$. Let $\mathrm{R}(\mathbf{X})$, called the column space of $\mathbf{X}$, be the linear space generated by the column vectors of $\mathbf{X}$. Then $\mathrm{E}(\mathbf{y}) \in \mathrm{R}(\mathbf{X})$. The least squares method is to find a vector $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\theta}}$ in $\mathrm{R}(\mathbf{X})$ such that $\|\mathbf{y} - \widehat{\mathbf{y}}\|$ is minimized. This is achieved if $\widehat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto

$\mathrm{R}(\mathbf{X})$. Denoting the orthogonal projection matrix onto a space V by $\mathbf{P}_V$, we have

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{P}_{\mathrm{R}(\mathbf{X})}\mathbf{y}.$$

Here $\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}$, called the *residual*, is the orthogonal projection of $\mathbf{y}$ onto $\mathrm{R}(\mathbf{X})^{\perp}$, where $\mathrm{R}(\mathbf{X})^{\perp} = \{\mathbf{x} \in \mathbb{R}^N : \mathbf{x} \text{ is orthogonal to all the vectors in } \mathrm{R}(\mathbf{X})\}$ is the orthogonal complement of $\mathrm{R}(\mathbf{X})$ in $\mathbb{R}^N$. Therefore $\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}$ is orthogonal to all the column vectors of $\mathbf{X}$, and it follows that

$$\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}\right) = \mathbf{0}$$

or

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{X}^T\mathbf{y}. \tag{2.4}$$

Equation (2.4) is called the *normal equation*. If $\mathbf{X}^T\mathbf{X}$ is invertible, then $\widehat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, with

$$\mathrm{E}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta} \text{ and } \mathrm{cov}(\widehat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \tag{2.5}$$

Let the rank of $\mathbf{X}$ be r. Then $\mathbf{X}^T\mathbf{X}$ is invertible if and only if $r = p$. Unless $r = p$, not all the parameters in $\boldsymbol{\theta}$ are identifiable, and solutions to (2.4) are not unique. In this case, (2.4) can be solved by using generalized inverses.

A *generalized inverse* of a matrix $\mathbf{A}$ is defined as any matrix $\mathbf{A}^-$ such that $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$. For any generalized inverse $(\mathbf{X}^T\mathbf{X})^-$ of $\mathbf{X}^T\mathbf{X}$, $\widehat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$ is a solution to the normal equation. Even though $\widehat{\boldsymbol{\theta}}$ may not be unique, since $\mathbf{y}$ has a unique orthogonal projection onto $\mathrm{R}(\mathbf{X})$, $\mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{y}$ is unique and does not depend on the choice of $(\mathbf{X}^T\mathbf{X})^-$. A byproduct of this is an explicit expression for the orthogonal projection matrix onto the column space of any matrix.

Theorem 2.1. *The orthogonal projection matrix onto* $R(\mathbf{X})$ *is* $\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$, *where* $(\mathbf{X}^T\mathbf{X})^-$ *is any generalized inverse of* $\mathbf{X}^T\mathbf{X}$.

A linear function $\mathbf{c}^T\boldsymbol{\theta} = \sum_{i=1}^{p} c_i\theta_i$ of the unknown parameters is said to be *estimable* if there exists an $N \times 1$ vector $\mathbf{a}$ such that $\mathrm{E}(\mathbf{a}^T\mathbf{y}) = \mathbf{c}^T\boldsymbol{\theta}$ for all $\boldsymbol{\theta}$. Such an estimator of $\mathbf{c}^T\boldsymbol{\theta}$ is called a a linear unbiased estimator. Since $\mathrm{E}(\mathbf{a}^T\mathbf{y}) = \mathbf{a}^T\mathbf{X}\boldsymbol{\theta}$, it is equal to $\mathbf{c}^T\boldsymbol{\theta}$ for all $\boldsymbol{\theta}$ if and only if $\mathbf{c} = \mathbf{X}^T\mathbf{a}$. This shows that $\mathbf{c}^T\boldsymbol{\theta}$ is estimable if and only if $\mathbf{c} \in \mathrm{R}(\mathbf{X}^T)$. From matrix algebra, it is known that $\mathrm{R}(\mathbf{X}^T) = \mathrm{R}(\mathbf{X}^T\mathbf{X})$. Therefore,

$$\mathbf{c}^T\boldsymbol{\theta} \text{ is estimable if and only if } \mathbf{c} \in \mathrm{R}(\mathbf{X}^T\mathbf{X}). \tag{2.6}$$

If $\mathbf{c}^T\boldsymbol{\theta}$ is estimable, then $\mathbf{c}^T\widehat{\boldsymbol{\theta}}$ does not depend on the solution to the normal equation.

Theorem 2.2. *(Gauss–Markov Theorem) Under (2.2)–(2.3), if* $\mathbf{c}^T\boldsymbol{\theta}$ *is estimable, then for any solution* $\widehat{\boldsymbol{\theta}}$ *to (2.4),* $\mathbf{c}^T\widehat{\boldsymbol{\theta}}$ *has the smallest variance among all the linear unbiased estimators of* $\mathbf{c}^T\boldsymbol{\theta}$.

It is common to refer to $\mathbf{c}^T\widehat{\boldsymbol{\theta}}$ as the best linear unbiased estimator, or BLUE, of $\mathbf{c}^T\boldsymbol{\theta}$.

For any estimable function $\mathbf{c}^T\boldsymbol{\theta}$, we have

$$\text{var}\left(\mathbf{c}^T\widehat{\boldsymbol{\theta}}\right) = \sigma^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{c},$$

where $(\mathbf{X}^T\mathbf{X})^-$ is any generalized inverse of $\mathbf{X}^T\mathbf{X}$. We call $\mathbf{X}^T\mathbf{X}$ the *information matrix* for $\boldsymbol{\theta}$.

Suppose the model matrix $\mathbf{X}$ is design dependent. We say that a design is D-, A-, or E-optimal if it minimizes, respectively, the determinant, trace, or the largest eigenvalue of the covariance matrix of $\widehat{\boldsymbol{\theta}}$ among all the competing designs. By (2.5), these optimality criteria, which are referred to as the D-, A-, and E-criterion, respectively, are equivalent to maximizing $\det(\mathbf{X}^T\mathbf{X})$, minimizing $\text{tr}(\mathbf{X}^T\mathbf{X})^{-1}$, and maximizing the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}$, respectively.

2.2 Estimation of σ^2

We have

$$\mathbf{y} = \mathbf{X}\widehat{\boldsymbol{\theta}} + (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}), \tag{2.7}$$

where $\mathbf{X}\widehat{\boldsymbol{\theta}}$ and $\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}$ are orthogonal to each other. By the Pythagorean Theorem,

$$\|\mathbf{y}\|^2 = \left\|\mathbf{X}\widehat{\boldsymbol{\theta}}\right\|^2 + \left\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}\right\|^2.$$

The last term in this identity, $\left\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}\right\|^2$, is called the *residual sum of squares*. Since $\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{P}_{\text{R}(\mathbf{X})^\perp}\mathbf{y}$,

$$\begin{aligned}
\text{E}\left[\left\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}\right\|^2\right] &= \text{E}\left[\left(\mathbf{P}_{\text{R}(\mathbf{X})^\perp}\mathbf{y}\right)^T\left(\mathbf{P}_{\text{R}(\mathbf{X})^\perp}\mathbf{y}\right)\right] \\
&= \text{E}\left(\mathbf{y}^T\mathbf{P}_{\text{R}(\mathbf{X})^\perp}\mathbf{y}\right) \\
&= [\text{E}(\mathbf{y})]^T\mathbf{P}_{\text{R}(\mathbf{X})^\perp}[\text{E}(\mathbf{y})] + \sigma^2\text{tr}\left(\mathbf{P}_{\text{R}(\mathbf{X})^\perp}\right) \\
&= \sigma^2\text{tr}\left(\mathbf{P}_{\text{R}(\mathbf{X})^\perp}\right) \\
&= \sigma^2\dim\left(\text{R}(\mathbf{X})^\perp\right) \\
&= \sigma^2(N - r),
\end{aligned}$$

where the second equality follows from (A.2) and (A.3) in the Appendix, the third equality follows from (A.5), the fourth equality holds since $\text{E}(\mathbf{y}) \in \text{R}(\mathbf{X})$ and $\text{R}(\mathbf{X}) \perp \text{R}(\mathbf{X})^\perp$, and the fifth equality follows from (A.4).

Let $s^2 = \left\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}\right\|^2/(N - r)$. Then s^2, called the residual *mean square*, is an unbiased estimator of σ^2. The dimension of $\text{R}(\mathbf{X})^\perp$, $N - r$, is called the *degrees of freedom* associated with the residual sum of squares.

Under the assumption that $\mathbf{y}$ has a normal distribution, $\frac{1}{\sigma^2}\left\|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\theta}}\right\|^2$ has a χ^2-distribution with $N-r$ degrees of freedom. Thus, for any estimable function $\mathbf{c}^T\boldsymbol{\theta}$,

$$\frac{\mathbf{c}^T\widehat{\boldsymbol{\theta}}}{s\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{c}}} \text{ has a } t\text{-distribution with } N-r \text{ degrees of freedom.}$$

Therefore a $100(1-\alpha)\%$ confidence interval for $\mathbf{c}^T\boldsymbol{\theta}$ is $\mathbf{c}^T\widehat{\boldsymbol{\theta}} \pm t_{N-r;1-\alpha/2}s \cdot \sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{c}}$, where $t_{N-r;1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the t-distribution with $N-r$ degrees of freedom.

In the rest of the book, we often abbreviate degrees of freedom, sum of squares, and mean square as d.f., SS, and MS, respectively.

2.3 F-test

We have seen that under linear model (2.2)–(2.3), $\mathrm{E}(\mathbf{y}) \in \mathrm{R}(\mathbf{X})$. Suppose V is a subspace of $\mathrm{R}(\mathbf{X})$ with $\dim(V) = q$. Let $\mathrm{R}(\mathbf{X}) \ominus V = \{\mathbf{y} \in \mathrm{R}(\mathbf{X}) : \mathbf{y} \text{ is orthogonal to } V\}$ be the orthogonal complement of V relative to $\mathrm{R}(\mathbf{X})$. Then $\mathrm{R}(\mathbf{X}) \ominus V$ has dimension $r-q$, and $\mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{P}_{\mathrm{R}(\mathbf{X})}\mathbf{y}$ can be decomposed as

$$\mathbf{P}_{\mathrm{R}(\mathbf{X})}\mathbf{y} = \mathbf{P}_V\mathbf{y} + \mathbf{P}_{\mathrm{R}(\mathbf{X})\ominus V}\mathbf{y}. \tag{2.8}$$

By combining (2.7) and (2.8), we have

$$\mathbf{y} - \mathbf{P}_V\mathbf{y} = \mathbf{P}_{\mathrm{R}(\mathbf{X})\ominus V}\mathbf{y} + \mathbf{P}_{\mathrm{R}(\mathbf{X})^\perp}\mathbf{y}, \tag{2.9}$$

where the two components on the right side are orthogonal. Thus

$$\|\mathbf{y} - \mathbf{P}_V\mathbf{y}\|^2 = \|\mathbf{P}_{\mathrm{R}(\mathbf{X})\ominus V}\mathbf{y}\|^2 + \left\|\mathbf{P}_{\mathrm{R}(\mathbf{X})^\perp}\mathbf{y}\right\|^2. \tag{2.10}$$

Under the assumption that $\mathbf{y}$ has a normal distribution, a test of the null hypothesis H_0: $\mathrm{E}(\mathbf{y}) \in V$ against the alternative hypothesis that $\mathrm{E}(\mathbf{y}) \notin V$ is based on the ratio

$$F = \frac{\|\mathbf{P}_{\mathrm{R}(\mathbf{X})\ominus V}\mathbf{y}\|^2/(r-q)}{\left\|\mathbf{P}_{\mathrm{R}(\mathbf{X})^\perp}\mathbf{y}\right\|^2/(N-r)}. \tag{2.11}$$

It can be shown that the likelihood ratio test is to reject H_0 for large values of F. Under H_0, F has an F-distribution with $r-q$ and $N-r$ degrees of freedom. Therefore the null hypothesis is rejected at level α if $F > F_{r-q,N-r;1-\alpha}$, where $F_{r-q,N-r;1-\alpha}$ is the $(1-\alpha)$th quantile of the F-distribution with $r-q$ and $N-r$ degrees of freedom.

The left side of (2.10) is the residual sum of squares when H_0 is true, and the second term on the right side is the residual sum of squares under the full model (2.2). Thus the sum of squares $\|\mathbf{P}_{\mathrm{R}(\mathbf{X})\ominus V}\mathbf{y}\|^2$ that appears in the numerator of the test statistic in (2.11) is the difference of the residual sums of squares under the full model and the reduced model specified by H_0.

2.4 One-way layout

Let $\mathbf{y} = (y_{11}, \ldots, y_{1r_1}, \ldots, y_{t1}, \ldots, y_{tr_t})^T$, where $y_{11}, \ldots, y_{1r_1}, \ldots, y_{t1}, \ldots, y_{tr_t}$ are uncorrelated random variables with constant variance σ^2, and $\mathrm{E}(y_{lh}) = \alpha_l$, for all $1 \le h \le r_l$, $1 \le l \le t$. Then we can express $\mathbf{y}$ as in (2.2)–(2.3), with $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_t)^T$, and

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_{r_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{r_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{r_t} \end{bmatrix},$$

where $\mathbf{1}_{r_l}$ is the $r_l \times 1$ vector of ones. This model arises, for example, when $y_{l1}, \ldots, y_{lr_l}$ are a random sample from a population with mean α_l and variance σ^2. It is also commonly used as a model for analyzing a completely randomized experiment, where the lth treatment, $1 \le l \le t$, is replicated r_l times, $y_{l1}, \ldots, y_{lr_l}$ are the observations on the lth treatment, and $\alpha_1, \ldots, \alpha_t$ are effects of the treatments.

Clearly $\mathbf{X}^T\mathbf{X} = \mathrm{diag}(r_1, \ldots, r_t)$, the $t \times t$ diagonal matrix with $r_1, \ldots, r_t$ as the diagonal entries; therefore (2.4) has a unique solution: $\widehat{\alpha}_l = y_{l\cdot}$, where $y_{l\cdot} = \frac{1}{r_l}\sum_{h=1}^{r_l} y_{lh}$. For any function $\sum_{l=1}^t c_l\alpha_l$, we have $\sum_{l=1}^t c_l\widehat{\alpha}_l = \sum_{l=1}^t c_l y_{l\cdot}$, with

$$\mathrm{var}\left(\sum_{l=1}^t c_l\widehat{\alpha}_l\right) = \sigma^2 \sum_{l=1}^t \frac{c_l^2}{r_l}.$$

The projection $\mathbf{P}_{\mathrm{R}(\mathbf{X})}\mathbf{y} = \mathbf{X}\widehat{\boldsymbol{\theta}}$ has its first r_1 components equal to $y_{1\cdot}$, the next r_2 components equal to $y_{2\cdot}, \ldots$, etc. Therefore the residual sum of squares can be expressed as

$$\left\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}\right\|^2 = \sum_{l=1}^t \sum_{h=1}^{r_l} (y_{lh} - y_{l\cdot})^2.$$

Call this the *within-group* sum of squares and denote it by W. We have $N = \sum_{l=1}^t r_l$, and $\mathrm{rank}(\mathbf{X}) = t$. Therefore the residual sum of squares has $N - t$ degrees of freedom, and so if $s^2 = W/(N-t)$, then $\mathrm{E}(s^2) = \sigma^2$.

Now we impose the normality assumption and consider the test of H_0: $\alpha_1 = \cdots = \alpha_t$. Under H_0, $\mathrm{E}(\mathbf{y}) \in V$, where V is the one-dimensional space consisting of all the vectors with constant entries. Then $\mathbf{P}_V\mathbf{y}$ is the vector with all the entries equal to the overall average $y_{\cdot\cdot} = \frac{1}{N}\sum_{l=1}^t \sum_{h=1}^{r_l} y_{lh}$. Componentwise, (2.9) can be expressed as

$$y_{lh} - y_{\cdot\cdot} = (y_{l\cdot} - y_{\cdot\cdot}) + (y_{lh} - y_{l\cdot})$$

and, in the present context, (2.10) reduces to

$$\sum_{l=1}^t \sum_{h=1}^{r_l} (y_{lh} - y_{\cdot\cdot})^2 = \sum_{l=1}^t r_l(y_{l\cdot} - y_{\cdot\cdot})^2 + \sum_{l=1}^t \sum_{h=1}^{r_l} (y_{lh} - y_{l\cdot})^2.$$

Thus $\|\mathbf{P}_{\mathrm{R}(\mathbf{X})\ominus V}\mathbf{y}\|^2 = \sum_{l=1}^t r_l(y_{l\cdot} - y_{\cdot\cdot})^2$, which has $t-1$ degrees of freedom and is called the *between-group* sum of squares. Denote it by B. Then the F-test statistic

is $\frac{B/(t-1)}{W/(N-t)}$. In the application to completely randomized experiments, the between-group sum of squares is also called the treatment sum of squares.

These results can be summarized in Table 2.1, called an ANOVA (Analysis of Variance) table.

Table 2.1 *ANOVA table for one-way layout*

source	sum of squares	d.f.	mean square
Between groups	$\sum_{l=1}^{t} r_l(y_{l\cdot} - y_{\cdot\cdot})^2$	$t-1$	$\frac{1}{t-1}\sum_{l=1}^{t} r_l(y_{l\cdot} - y_{\cdot\cdot})^2$
Within groups	$\sum_{l=1}^{t}\sum_{h=1}^{r_l}(y_{lh} - y_{l\cdot})^2$	$N-t$	$\frac{1}{N-t}\sum_{l=1}^{t}\sum_{h=1}^{r_l}(y_{lh} - y_{l\cdot})^2$
Total	$\sum_{l=1}^{t}\sum_{h=1}^{r_l}(y_{lh} - y_{\cdot\cdot})^2$	$N-1$	

Remark 2.1. If we write the model as $\mathrm{E}(y_{lh}) = \mu + \alpha_l$, then the model matrix $\mathbf{X}$ has an extra column of 1's and $p = t+1$. Since $\mathrm{rank}(\mathbf{X}) = t < p$, the parameters themselves are not identifiable. By using (2.6), one can verify that $\sum_{l=1}^{t} c_l\alpha_l$ is estimable if and only if $\sum_{l=1}^{t} c_l = 0$. Such functions are called *contrasts*. The pairwise differences $\alpha_l - \alpha_{l'}$, $1 \le l \ne l' \le t$, are examples of contrasts and are referred to as elementary contrasts. Because of the constraint $\sum_{l=1}^{t} c_l = 0$, the treatment contrasts form a $(t-1)$-dimensional vector space that is generated by the elementary contrasts. Functions such as $\alpha_1 - \frac{1}{2}(\alpha_2+\alpha_3)$ and $\frac{1}{2}(\alpha_1+\alpha_2) - \frac{1}{3}(\alpha_3+\alpha_4+\alpha_5)$ are also contrasts. If $\sum_{l=1}^{t} c_l\alpha_l$ is a contrast, then $\sum_{l=1}^{t} c_l\alpha_l = \sum_{l=1}^{t} c_l(\mu+\alpha_l)$, and it can be shown that, as before, $\sum_{l=1}^{t} c_l\widehat{\alpha}_l = \sum_{l=1}^{t} c_l y_{l\cdot}$. Therefore, if the interest is in estimating the contrasts, then it does not matter whether $\mathrm{E}(y_{lh})$ is written as α_l or $\mu+\alpha_l$. A contrast $\sum_{l=1}^{t} c_l\alpha_l$ is said to be normalized if $\sum_{l=1}^{t} c_l^2 = 1$.

2.5 Estimation of a subset of parameters

Suppose $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T\ \boldsymbol{\theta}_2^T)^T$, where $\boldsymbol{\theta}_1$ is $q \times 1$ and $\boldsymbol{\theta}_2$ is $(p-q)\times 1$; e.g., the parameters in $\boldsymbol{\theta}_2$ are nuisance parameters, or the components of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are effects of the levels of two different factors. Partition $\mathbf{X}$ as $[\mathbf{X}_1\ \mathbf{X}_2]$ accordingly. Then (2.2) can be written as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\theta}_1 + \mathbf{X}_2\boldsymbol{\theta}_2 + \boldsymbol{\varepsilon}, \tag{2.12}$$

and (2.4) is the same as

$$\mathbf{X}_1^T\mathbf{X}_1\widehat{\boldsymbol{\theta}}_1 + \mathbf{X}_1^T\mathbf{X}_2\widehat{\boldsymbol{\theta}}_2 = \mathbf{X}_1^T\mathbf{y}, \tag{2.13}$$

and

$$\mathbf{X}_2^T\mathbf{X}_1\widehat{\boldsymbol{\theta}}_1 + \mathbf{X}_2^T\mathbf{X}_2\widehat{\boldsymbol{\theta}}_2 = \mathbf{X}_2^T\mathbf{y}. \tag{2.14}$$

If $\mathrm{R}(\mathbf{X}_1)$ and $\mathrm{R}(\mathbf{X}_2)$ are orthogonal, $(\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0})$, then $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$ can be computed by solving $\mathbf{X}_1^T\mathbf{X}_1\widehat{\boldsymbol{\theta}}_1 = \mathbf{X}_1^T\mathbf{y}$ and $\mathbf{X}_2^T\mathbf{X}_2\widehat{\boldsymbol{\theta}}_2 = \mathbf{X}_2^T\mathbf{y}$ separately. In this case, least squares estimators of estimable functions of $\boldsymbol{\theta}_1$ are the same regardless of whether $\boldsymbol{\theta}_2$ is

in the model. Likewise, dropping $\boldsymbol{\theta}_1$ from (2.12) does not change the least squares estimators of estimable functions of $\boldsymbol{\theta}_2$.

If $\mathrm{R}(\mathbf{X}_1)$ and $\mathrm{R}(\mathbf{X}_2)$ are not orthogonal, then for estimating $\boldsymbol{\theta}_1$ one needs to adjust for $\boldsymbol{\theta}_2$, and vice versa. By (2.14), $\widehat{\boldsymbol{\theta}}_2$ can be written as

$$\widehat{\boldsymbol{\theta}}_2 = \left(\mathbf{X}_2^T\mathbf{X}_2\right)^- \left[\mathbf{X}_2^T\mathbf{y} - \mathbf{X}_2^T\mathbf{X}_1\widehat{\boldsymbol{\theta}}_1\right]. \tag{2.15}$$

We eliminate $\widehat{\boldsymbol{\theta}}_2$ by substituting the right side of (2.15) for the $\widehat{\boldsymbol{\theta}}_2$ in (2.13). Then $\widehat{\boldsymbol{\theta}}_1$ can be obtained by solving

$$\left[\mathbf{X}_1^T\mathbf{X}_1 - \mathbf{X}_1^T\mathbf{X}_2\left(\mathbf{X}_2^T\mathbf{X}_2\right)^-\mathbf{X}_2^T\mathbf{X}_1\right]\widehat{\boldsymbol{\theta}}_1 = \mathbf{X}_1^T\mathbf{y} - \mathbf{X}_1^T\mathbf{X}_2\left(\mathbf{X}_2^T\mathbf{X}_2\right)^-\mathbf{X}_2^T\mathbf{y}$$

or

$$\left\{\mathbf{X}_1^T\left[\mathbf{I} - \mathbf{X}_2\left(\mathbf{X}_2^T\mathbf{X}_2\right)^-\mathbf{X}_2^T\right]\mathbf{X}_1\right\}\widehat{\boldsymbol{\theta}}_1 = \mathbf{X}_1^T\left[\mathbf{I} - \mathbf{X}_2\left(\mathbf{X}_2^T\mathbf{X}_2\right)^-\mathbf{X}_2^T\right]\mathbf{y}. \tag{2.16}$$

This is called the *reduced normal equation* for $\boldsymbol{\theta}_1$.

We write the reduced normal equation as

$$\mathbf{C}_1\widehat{\boldsymbol{\theta}}_1 = \mathbf{Q}_1, \tag{2.17}$$

where

$$\mathbf{C}_1 = \mathbf{X}_1^T\left[\mathbf{I} - \mathbf{X}_2\left(\mathbf{X}_2^T\mathbf{X}_2\right)^-\mathbf{X}_2^T\right]\mathbf{X}_1, \tag{2.18}$$

and

$$\mathbf{Q}_1 = \mathbf{X}_1^T\left[\mathbf{I} - \mathbf{X}_2\left(\mathbf{X}_2^T\mathbf{X}_2\right)^-\mathbf{X}_2^T\right]\mathbf{y}. \tag{2.19}$$

Since $\mathbf{X}_2(\mathbf{X}_2^T\mathbf{X}_2)^-\mathbf{X}_2^T$ is the orthogonal projection matrix onto $\mathrm{R}(\mathbf{X}_2)$, $\mathbf{I} - \mathbf{X}_2(\mathbf{X}_2^T\mathbf{X}_2)^-\mathbf{X}_2^T$ is the orthogonal projection matrix onto $\mathrm{R}(\mathbf{X}_2)^\perp$, and

$$\mathbf{C}_1 = \mathbf{X}_1^T\mathbf{P}_{\mathrm{R}(\mathbf{X}_2)^\perp}\mathbf{X}_1.$$

If we put

$$\tilde{\mathbf{X}}_1 = \mathbf{P}_{\mathrm{R}(\mathbf{X}_2)^\perp}\mathbf{X}_1,$$

then we can express $\mathbf{C}_1$ as

$$\mathbf{C}_1 = \tilde{\mathbf{X}}_1^T\tilde{\mathbf{X}}_1,$$

and the reduced normal equation (2.16) can be written as

$$\tilde{\mathbf{X}}_1^T\tilde{\mathbf{X}}_1\widehat{\boldsymbol{\theta}}_1 = \tilde{\mathbf{X}}_1^T\tilde{\mathbf{y}},$$

where $\tilde{\mathbf{y}} = \mathbf{P}_{\mathrm{R}(\mathbf{X}_2)^\perp}\mathbf{y}$. Therefore the least squares estimators of estimable functions of $\boldsymbol{\theta}_1$ are functions of $\tilde{\mathbf{y}} = \mathbf{P}_{\mathrm{R}(\mathbf{X}_2)^\perp}\mathbf{y}$: to eliminate $\boldsymbol{\theta}_2$, (2.12) is projected onto $\mathrm{R}(\mathbf{X}_2)^\perp$ to become

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}_1\boldsymbol{\theta}_1 + \tilde{\boldsymbol{\varepsilon}},$$

where $\tilde{\boldsymbol{\varepsilon}} = \mathbf{P}_{\mathrm{R}(\mathbf{X}_2)^\perp}\boldsymbol{\varepsilon}$.

Theorem 2.3. *Under (2.12) and (2.3), a linear function* $\mathbf{c}_1^T\boldsymbol{\theta}_1$ *of* $\boldsymbol{\theta}_1$ *is estimable if and only if* $\mathbf{c}_1$ *is a linear combination of the column vectors of* $\mathbf{C}_1$*. If* $\mathbf{c}_1^T\boldsymbol{\theta}_1$ *is estimable, then* $\mathbf{c}_1^T\widehat{\boldsymbol{\theta}}_1$ *is its best linear unbiased estimator, and*

$$var\left(\mathbf{c}_1^T\widehat{\boldsymbol{\theta}}_1\right) = \sigma^2\mathbf{c}_1^T\mathbf{C}_1^-\mathbf{c}_1,$$

where $\widehat{\boldsymbol{\theta}}_1$ *is any solution to (2.17), and* $\mathbf{C}_1^-$ *is any generalized inverse of* $\mathbf{C}_1$*.*

In particular, if $\mathbf{C}_1$ is invertible, then all the parameters in $\boldsymbol{\theta}_1$ are estimable, with

$$\text{cov}\left(\widehat{\boldsymbol{\theta}}_1\right) = \sigma^2\mathbf{C}_1^{-1}. \tag{2.20}$$

The matrix $\mathbf{C}_1$ is called the information matrix for $\boldsymbol{\theta}_1$.

Suppose the parameters in $\boldsymbol{\theta}_2$ are nuisance parameters, and we are only interested in estimating $\boldsymbol{\theta}_1$. Then a design is said to be D_s-, A_s-, or E_s-optimal if it minimizes, respectively, the determinant, trace, or the largest eigenvalue of the covariance matrix of $\widehat{\boldsymbol{\theta}}_1$ among all the competing designs. By (2.20), these optimality criteria are equivalent to maximizing $\det(\mathbf{C}_1)$, minimizing $\text{tr}(\mathbf{C}_1^{-1})$, and maximizing the smallest eigenvalue of $\mathbf{C}_1$, respectively. Here *s* refers to a *subset* of parameters.

2.6 Hypothesis testing for a subset of parameters

Under (2.12) and (2.3), suppose we further assume that $\boldsymbol{\varepsilon}$ has a normal distribution. Consider testing the null hypothesis

$$H_0\text{: } \text{E}(\mathbf{y}) = \mathbf{X}_2\boldsymbol{\theta}_2.$$

Under (2.12), $\text{E}(\mathbf{y}) \in \text{R}(\mathbf{X}_1)+\text{R}(\mathbf{X}_2)$, and under H_0, $\text{E}(\mathbf{y}) \in \text{R}(\mathbf{X}_2)$. By (2.11), the F-test statistic in this case is

$$F = \frac{\left\|\mathbf{P}_{[\text{R}(\mathbf{X}_1)+\text{R}(\mathbf{X}_2)]\ominus\text{R}(\mathbf{X}_2)}\mathbf{y}\right\|^2/\dim([\text{R}(\mathbf{X}_1)+\text{R}(\mathbf{X}_2)]\ominus\text{R}(\mathbf{X}_2))}{\left\|\mathbf{P}_{[\text{R}(\mathbf{X}_1)+\text{R}(\mathbf{X}_2)]^\perp}\mathbf{y}\right\|^2/(N-\dim[\text{R}(\mathbf{X}_1)+\text{R}(\mathbf{X}_2)])}.$$

This is based on the decomposition

$$\mathbb{R}^N = \text{R}(\mathbf{X}_2)\oplus([\text{R}(\mathbf{X}_1)+\text{R}(\mathbf{X}_2)]\ominus\text{R}(\mathbf{X}_2))\oplus[\text{R}(\mathbf{X}_1)+\text{R}(\mathbf{X}_2)]^\perp. \tag{2.21}$$

It can be shown that

$$\left\|\mathbf{P}_{[\text{R}(\mathbf{X}_1)+\text{R}(\mathbf{X}_2)]\ominus\text{R}(\mathbf{X}_2)}\mathbf{y}\right\|^2 = \widehat{\boldsymbol{\theta}}_1^T\mathbf{Q}_1, \tag{2.22}$$

and

$$\dim([\text{R}(\mathbf{X}_1)+\text{R}(\mathbf{X}_2)]\ominus\text{R}(\mathbf{X}_2)) = \text{rank}(\mathbf{C}_1). \tag{2.23}$$

We leave the proofs of these as an exercise.

A test of the null hypothesis H_0: $\text{E}(\mathbf{y}) = \mathbf{X}_1\boldsymbol{\theta}_1$ is based on the decomposition

$$\mathbb{R}^N = \text{R}(\mathbf{X}_1)\oplus([\text{R}(\mathbf{X}_1)+\text{R}(\mathbf{X}_2)]\ominus\text{R}(\mathbf{X}_1))\oplus[\text{R}(\mathbf{X}_1)+\text{R}(\mathbf{X}_2)]^\perp. \tag{2.24}$$

In this case,

$$\left\|\mathbf{P}_{[\mathrm{R}(\mathbf{X}_1)+\mathrm{R}(\mathbf{X}_2)]\ominus\mathrm{R}(\mathbf{X}_1)}\mathbf{y}\right\|^2 = \widehat{\boldsymbol{\theta}}_2^T\mathbf{Q}_2,$$

where

$$\mathbf{Q}_2 = \mathbf{X}_2^T\left[\mathbf{I}-\mathbf{X}_1\left(\mathbf{X}_1^T\mathbf{X}_1\right)^-\mathbf{X}_1^T\right]\mathbf{y}.$$

When $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$, both (2.21) and (2.24) reduce to

$$\mathbb{R}^N = \mathrm{R}(\mathbf{X}_1)\oplus\mathrm{R}(\mathbf{X}_2)\oplus[\mathrm{R}(\mathbf{X}_1)\oplus\mathrm{R}(\mathbf{X}_2)]^\perp.$$

2.7 Adjusted orthogonality

Consider the model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\theta}_1+\mathbf{X}_2\boldsymbol{\theta}_2+\mathbf{X}_3\boldsymbol{\theta}_3+\boldsymbol{\varepsilon}, \tag{2.25}$$

with

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \text{ and } \mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_N.$$

Suppose the parameters in $\boldsymbol{\theta}_2$ are nuisance parameters. To estimate $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_3$, we eliminate $\boldsymbol{\theta}_2$ by projecting $\mathbf{y}$ onto $\mathrm{R}(\mathbf{X}_2)^\perp$. This results in a reduced normal equation for $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_3$:

$$\begin{bmatrix}\tilde{\mathbf{X}}_1^T\tilde{\mathbf{X}}_1 & \tilde{\mathbf{X}}_1^T\tilde{\mathbf{X}}_3\\ \tilde{\mathbf{X}}_3^T\tilde{\mathbf{X}}_1 & \tilde{\mathbf{X}}_3^T\tilde{\mathbf{X}}_3\end{bmatrix}\begin{bmatrix}\widehat{\boldsymbol{\theta}}_1\\ \widehat{\boldsymbol{\theta}}_3\end{bmatrix} = \begin{bmatrix}\tilde{\mathbf{X}}_1^T\tilde{\mathbf{y}}\\ \tilde{\mathbf{X}}_3^T\tilde{\mathbf{y}}\end{bmatrix},$$

where

$$\tilde{\mathbf{X}}_i = \mathbf{P}_{\mathrm{R}(\mathbf{X}_2)^\perp}\mathbf{X}_i, i = 1,3, \text{ and } \tilde{\mathbf{y}} = \mathbf{P}_{\mathrm{R}(\mathbf{X}_2)^\perp}\mathbf{y}.$$

If $\tilde{\mathbf{X}}_1^T\tilde{\mathbf{X}}_3 = \mathbf{0}$, then $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_3$ can be computed by solving the equations $\tilde{\mathbf{X}}_1^T\tilde{\mathbf{X}}_1\widehat{\boldsymbol{\theta}}_1 = \tilde{\mathbf{X}}_1^T\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}_3^T\tilde{\mathbf{X}}_3\widehat{\boldsymbol{\theta}}_3 = \tilde{\mathbf{X}}_3^T\tilde{\mathbf{y}}$ separately. In this case, least squares estimators of estimable functions of $\boldsymbol{\theta}_1$ under (2.25) are the same regardless of whether $\boldsymbol{\theta}_3$ is in the model. Likewise, dropping $\boldsymbol{\theta}_1$ from (2.25) does not change the least squares estimators of estimable functions of $\boldsymbol{\theta}_3$. Loosely we say that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_3$ are orthogonal adjusted for $\boldsymbol{\theta}_2$.

Note that $\tilde{\mathbf{X}}_1^T\tilde{\mathbf{X}}_3 = \mathbf{0}$ is equivalent to

$$P_{\mathrm{R}(\mathbf{X}_2)^\perp}[\mathrm{R}(\mathbf{X}_1)] \text{ is orthogonal to } P_{\mathrm{R}(\mathbf{X}_2)^\perp}[\mathrm{R}(\mathbf{X}_3)]. \tag{2.26}$$

Under normality, the F-test statistic for the null hypothesis that $\mathrm{E}(\mathbf{y}) = \mathbf{X}_2\boldsymbol{\theta}_2+\mathbf{X}_3\boldsymbol{\theta}_3$ is

$$\frac{\left\|\mathbf{P}_{[\mathrm{R}(\mathbf{X}_1)+\mathrm{R}(\mathbf{X}_2)+\mathrm{R}(\mathbf{X}_3)]\ominus[\mathrm{R}(\mathbf{X}_2)+\mathrm{R}(\mathbf{X}_3)]}\mathbf{y}\right\|^2/\dim([\mathrm{R}(\mathbf{X}_1)+\mathrm{R}(\mathbf{X}_2)+\mathrm{R}(\mathbf{X}_3)]\ominus[\mathrm{R}(\mathbf{X}_2)+\mathrm{R}(\mathbf{X}_3)])}{\left\|\mathbf{P}_{[\mathrm{R}(\mathbf{X}_1)+\mathrm{R}(\mathbf{X}_2)+\mathrm{R}(\mathbf{X}_3)]^\perp}\mathbf{y}\right\|^2/(N-\dim[\mathrm{R}(\mathbf{X}_1)+\mathrm{R}(\mathbf{X}_2)+\mathrm{R}(\mathbf{X}_3)])}.$$

If (2.26) holds, then we have

$$\mathrm{R}(\mathbf{X}_1)+\mathrm{R}(\mathbf{X}_2)+\mathrm{R}(\mathbf{X}_3) = \mathrm{R}(\mathbf{X}_2)\oplus P_{\mathrm{R}(\mathbf{X}_2)^\perp}[\mathrm{R}(\mathbf{X}_1)]\oplus P_{\mathrm{R}(\mathbf{X}_2)^\perp}[\mathrm{R}(\mathbf{X}_3)].$$

In this case,

$$\begin{aligned}[\mathrm{R}(\mathbf{X}_1)+\mathrm{R}(\mathbf{X}_2)+\mathrm{R}(\mathbf{X}_3)]\ominus[\mathrm{R}(\mathbf{X}_2)+\mathrm{R}(\mathbf{X}_3)] &= P_{\mathrm{R}(\mathbf{X}_2)^\perp}[\mathrm{R}(\mathbf{X}_1)] \\ &= [\mathrm{R}(\mathbf{X}_1)+\mathrm{R}(\mathbf{X}_2)]\ominus\mathrm{R}(\mathbf{X}_2).\end{aligned}$$

So the sum of squares in the numerator of the F-test statistic is equal to the quantity $\widehat{\boldsymbol{\theta}}_1^T\mathbf{Q}_1$ given in (2.22), with its degrees of freedom equal to $\mathrm{rank}(\mathbf{C}_1)$, where $\mathbf{C}_1$ and $\mathbf{Q}_1$ are as in (2.18) and (2.19), respectively. A similar conclusion can be drawn for testing the hypothesis that $\mathrm{E}(\mathbf{y})=\mathbf{X}_1\boldsymbol{\theta}_1+\mathbf{X}_2\boldsymbol{\theta}_2$.

2.8 Additive two-way layout

In a two-way layout, the observations are classified according to the levels of two factors. Suppose the two factors have t and b levels, respectively. At each level combination (i,j), $1\le i\le t$, $1\le j\le b$, there are n_{ij} observations y_{ijh}, $0\le h\le n_{ij}$, such that

$$y_{ijh}=\alpha_i+\beta_j+\varepsilon_{ijh}, \tag{2.27}$$

where the ε_{ijh}'s are uncorrelated random variables with zero mean and constant variance σ^2. We require $\sum_{j=1}^b n_{ij}>0$ for all i, and $\sum_{i=1}^t n_{ij}>0$ for all j, so that there is at least one observation on each level of the two factors, but some n_{ij}'s may be zero. This is called an additive two-way layout model, which is commonly used for analyzing block designs, where each y_{ijh} is an observation on the ith treatment in the jth block. With this in mind, we call the two factors treatment and block factors, and denote them by $\mathcal{T}$ and $\mathcal{B}$, respectively; then $\alpha_1,\ldots,\alpha_t$ are the treatment effects and $\beta_1,\ldots,\beta_b$ are the block effects. Let $N=\sum_{i=1}^t\sum_{j=1}^b n_{ij}$, and think of the observations as taken on N units that are grouped into b blocks. Define an $N\times t$ matrix $\mathbf{X}_\mathcal{T}$ with 0 and 1 entries such that the (v,i)th entry of $\mathbf{X}_\mathcal{T}$, $1\le v\le N$, $1\le i\le t$, is 1 if and only if the ith treatment is assigned to the vth unit. Similarly, let $\mathbf{X}_\mathcal{B}$ be the $N\times b$ matrix with 0 and 1 entries such that the (v,j)th entry of $\mathbf{X}_\mathcal{B}$, $1\le v\le N$, $1\le j\le b$, is 1 if and only if the vth unit is in the jth block. The two matrices $\mathbf{X}_\mathcal{T}$ and $\mathbf{X}_\mathcal{B}$ are called unit-treatment and unit-block incidence matrices, respectively. Then we can write (2.27) as

$$\mathbf{y}=\mathbf{X}_\mathcal{T}\boldsymbol{\alpha}+\mathbf{X}_\mathcal{B}\boldsymbol{\beta}+\boldsymbol{\varepsilon},$$

where $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_t)^T$ and $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_b)^T$.

We use the results in Sections 2.5 and 2.6 to derive the analysis for such models. Then we apply the results in Section 2.7 to show in Section 2.9 that the analysis can be much simplified when certain conditions are satisfied.

Let $\mathbf{N}$ be the $t\times b$ matrix whose (i,j)th entry is n_{ij}, $n_{i+}=\sum_{j=1}^b n_{ij}$, and $n_{+j}=\sum_{i=1}^t n_{ij}$. For block designs, n_{i+} is the number of observations on the ith treatment, and n_{+j} is the size of the jth block. Also, let $y_{i++}=\sum_{j=1}^b\sum_h y_{ijh}$ and $y_{+j+}=\sum_{i=1}^t\sum_h y_{ijh}$, the ith treatment total and jth block total, respectively. Then $\mathbf{X}_\mathcal{T}^T\mathbf{X}_\mathcal{T}$ is the diagonal matrix with diagonal entries $n_{1+},\ldots,n_{t+}$, $\mathbf{X}_\mathcal{B}^T\mathbf{X}_\mathcal{B}$ is the diagonal matrix with diagonal entries $n_{+1},\ldots,n_{+b}$, $\mathbf{X}_\mathcal{T}^T\mathbf{y}=(y_{1++},\ldots,y_{t++})^T$, $\mathbf{X}_\mathcal{B}^T\mathbf{y}=(y_{+1+},\ldots,y_{+b+})^T$, and

$$\mathbf{X}_\mathcal{T}^T\mathbf{X}_\mathcal{B}=\mathbf{N}. \tag{2.28}$$

By (2.17), (2.18), and (2.19), the reduced normal equations for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are, respectively,

$$\mathbf{C}_{\mathcal{T}}\widehat{\boldsymbol{\alpha}} = \mathbf{Q}_{\mathcal{T}} \tag{2.29}$$

and

$$\mathbf{C}_{\mathcal{B}}\widehat{\boldsymbol{\beta}} = \mathbf{Q}_{\mathcal{B}},$$

where

$$\mathbf{C}_{\mathcal{T}} = \begin{bmatrix} n_{1+} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & n_{t+} \end{bmatrix} - \mathbf{N} \begin{bmatrix} \frac{1}{n_{+1}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{n_{+b}} \end{bmatrix} \mathbf{N}^T,$$

$$\mathbf{Q}_{\mathcal{T}} = \begin{bmatrix} y_{1++} \\ \vdots \\ y_{t++} \end{bmatrix} - \mathbf{N} \begin{bmatrix} \frac{1}{n_{+1}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{n_{+b}} \end{bmatrix} \begin{bmatrix} y_{+1+} \\ \vdots \\ y_{+b+} \end{bmatrix},$$

$$\mathbf{C}_{\mathcal{B}} = \begin{bmatrix} n_{+1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & n_{+b} \end{bmatrix} - \mathbf{N}^T \begin{bmatrix} \frac{1}{n_{1+}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{n_{t+}} \end{bmatrix} \mathbf{N},$$

and

$$\mathbf{Q}_{\mathcal{B}} = \begin{bmatrix} y_{+1+} \\ \vdots \\ y_{+b+} \end{bmatrix} - \mathbf{N}^T \begin{bmatrix} \frac{1}{n_{1+}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{n_{t+}} \end{bmatrix} \begin{bmatrix} y_{1++} \\ \vdots \\ y_{t++} \end{bmatrix}.$$

It can be verified that both $\mathbf{C}_{\mathcal{T}}$ and $\mathbf{C}_{\mathcal{B}}$ have zero column sums. Therefore $\operatorname{rank}(\mathbf{C}_{\mathcal{T}}) \le t-1$, $\operatorname{rank}(\mathbf{C}_{\mathcal{B}}) \le b-1$, and, by Theorem 2.3, if $\sum_{i=1}^{t} c_i\alpha_i$ is estimable, then $\sum_{i=1}^{t} c_i = 0$. All such contrasts are estimable if and only if $\operatorname{rank}(\mathbf{C}_{\mathcal{T}}) = t-1$. Similarly, if $\sum_{j=1}^{b} d_j\beta_j$ is estimable, then $\sum_{j=1}^{b} d_j = 0$, and all such contrasts are estimable if and only if $\operatorname{rank}(\mathbf{C}_{\mathcal{B}}) = b-1$.

Theorem 2.4. *All the contrasts of $\alpha_1, \ldots, \alpha_t$ are estimable if and only if for any $1 \le i \ne i' \le t$, there is a sequence*

$$i_1, j_1, i_2, j_2, \ldots, i_k, j_k, i_{k+1}$$

such that $i_1 = i$, $i_{k+1} = i'$, and for all $1 \le s \le k$, $n_{i_s j_s} > 0$ and $n_{i_{s+1} j_s} > 0$.

Proof. Suppose the condition in the theorem holds. We need to show that all the contrasts of $\alpha_1, \ldots, \alpha_t$ are estimable. Since the space of all the contrasts is generated by the pairwise differences, it suffices to show that all the pairwise differences $\alpha_i - \alpha_{i'}$ are estimable.

Let y_{ij} be any of the observations at the level combination (i, j). Then

$$\begin{aligned} & \mathrm{E}(y_{i_1 j_1} - y_{i_2 j_1} + y_{i_2 j_2} - \cdots + y_{i_k j_k} - y_{i_{k+1} j_k}) \\ = \; & (\alpha_{i_1} + \beta_{j_1}) - (\alpha_{i_2} + \beta_{j_1}) + (\alpha_{i_2} + \beta_{j_2}) - \cdots + (\alpha_{i_k} + \beta_{j_k}) - (\alpha_{i_{k+1}} + \beta_{j_k}) \\ = \; & \alpha_{i_1} - \alpha_{i_{k+1}} \\ = \; & \alpha_i - \alpha_{i'}. \end{aligned}$$

This shows that $\alpha_i - \alpha_{i'}$ is estimable.

Conversely, if the condition in the theorem does not hold, then the treatments can be partitioned into two disjoint sets such that any treatment from one set never appears in the same block with any treatment from the other set. Then it can be seen that $\mathbf{C}_{\mathcal{T}}$ is of the form

$$\begin{bmatrix} \mathbf{C}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 \end{bmatrix}.$$

Since $\mathbf{C}_{\mathcal{T}}$ has zero column sums, both $\mathbf{C}_1$ and $\mathbf{C}_2$ also have zero column sums. It follows that $\mathrm{rank}(\mathbf{C}_{\mathcal{T}}) = \mathrm{rank}(\mathbf{C}_1) + \mathrm{rank}(\mathbf{C}_2) \leq t - 2$; therefore not all the contrasts of $\alpha_1, \ldots, \alpha_t$ are estimable. □

If any two treatments can be connected by a chain of alternating treatments and blocks as in Theorem 2.4, then any two blocks can also be connected by such a sequence. Therefore the condition in Theorem 2.4 is also a necessary and sufficient condition for all the contrasts of $\beta_1, \ldots, \beta_b$ to be estimable. In particular, all the contrasts of $\alpha_1, \ldots, \alpha_t$ are estimable if and only if all the contrasts of $\beta_1, \ldots, \beta_b$ are estimable.

Throughout the rest of this section, we assume that the condition in Theorem 2.4 holds; therefore $\mathrm{rank}(C_{\mathcal{T}}) = t - 1$ and $\mathrm{rank}(\mathbf{C}_{\mathcal{B}}) = b - 1$. This is the case, for example, when there is at least one observation at each level combination of the two factors. In view of Theorem 2.4, designs with $\mathrm{rank}(C_{\mathcal{T}}) = t - 1$ are called *connected* designs.

Suppose we would like to test the hypothesis that $\alpha_1 = \cdots = \alpha_t$. Under the null hypothesis, let the common value of the α_i's be α; then $\mathrm{E}(y_{ijh}) = \alpha + \beta_j$. This reduces (2.27) to a one-way layout model. Absorb α into β_j (see Remark 2.1); then it is the same as to test that $\mathrm{E}(\mathbf{y}) = \mathbf{X}_{\mathcal{B}}\boldsymbol{\beta}$. Therefore the results in Section 2.6 can be applied. In particular, the sum of squares that appears in the numerator of the F-test statistic is equal to $\widehat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\mathcal{T}} = \widehat{\boldsymbol{\alpha}}^T \mathbf{C}_{\mathcal{T}} \widehat{\boldsymbol{\alpha}}$, with $t - 1$ degrees of freedom. The residual sum of squares can be computed by subtracting $\widehat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\mathcal{T}}$ from the residual (within-group) sum of squares under the one-way layout model containing block effects only. Let $y_{i\cdot\cdot} = \frac{1}{n_{i+}} \sum_{j=1}^{b} \sum_h y_{ijh}$, $y_{\cdot j\cdot} = \frac{1}{n_{+j}} \sum_{i=1}^{t} \sum_h y_{ijh}$, and $y_{\cdots} = \frac{1}{N} \sum_{i=1}^{t} \sum_{j=1}^{b} \sum_h y_{ijh}$ be the ith treatment mean, jth block mean, and overall mean, respectively. Then we have the ANOVA in Table 2.2.

A test of the hypothesis that $\mathrm{E}(\mathbf{y}) = \mathbf{X}_{\mathcal{T}}\boldsymbol{\alpha}$ can be based on an ANOVA similar to that in Table 2.2 with the roles of treatments and blocks reversed.